

https://dspace.org/

A retrospective analysis of laboratory data to identify clinical practice trends in prostate cancer screening and diagnosis by level of care in the Gauteng Province between 2006 and 2016

Cassim, Naseem

2021
https://hdl.handle.net/10539/33567

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

A retrospective analysis of laboratory data to identify clinical practice trends in prostate cancer screening and diagnosis by level of care in the Gauteng Province between 2006 and 2016

Submitted by
Naseem Cassim

A thesis submitted to the Department of Molecular Medicine and Haematology, Faculty of Health Sciences, University of the Witwatersrand, in fulfilment of the requirements for the degree of Doctor of Philosophy

September, 2021

Declaration

I, Naseem Cassim declare that this research report is my own unaided work. The report is being submitted of requirements of the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted previously for any degree or examination at any other university.



_____

Signed

16 September 2021

Date

DEDICATION

I dedicate this thesis to my late parents Fatima Jassat and Abdool Hay Cassim, my dear wife Zarina Khan and our sons Muhammed Rashaad and Rameez Cassim.

AUTHORS CONTRIBUTION TO THE WORK

I was involved in all the stages of this study from identifying the problem statement, developing a hypothesis, study design, ethics application, protocol development and executing the data analysis. I requested the data from the corporate data warehouse (CDW) after obtaining approval from the Academic Affairs, Research and Quality Assurance (AARQA) department of the National Health Laboratory Service (NHLS). I performed all the data preparation and conducted the data analysis including statistical analysis. I also carried out the data checks and coded data to facilitate the analysis.

PUBLICATIONS ARISING FROM THIS THESIS

The following publications are submitted

Cassim N, Mapundu M, Olago V, Celik T, George JA and Glencross DK. (2021) Using text mining techniques to extract prostate cancer predictive information (Gleason score) from semi-structured narrative laboratory reports in the Gauteng province, South Africa.
BMC Medical Informatics and Decision Making

Contributions to the paper by each author:

NC providing leadership, technical assistance to validate the study findings and prepared the initial draft. MM and VO developed the methodology and conducted the research. All authors contributed to reviewing initial draft. TC, JAG & DKG contributed to the drafting and revising of the work critically for important intellectual content as well as overall supervision. All authors read and approved the final manuscript.

The following publications are submitted and currently under review

Cassim M, Rebbeck TR, Glencross DK and George JA. (2021) Retrospective analysis of prostate-specific antigen testing trends and primary health care public-sector guidelines in the Gauteng Province, South Africa, between 2006 and 2016. BMJ Open

Contributions to the paper by each author:

NC made substantial contributions to the conception or design of the work, acquisition of laboratory data, data analysis, drafting the work and revising it critically for important intellectual content. TRR made substantial contributions to the conceptualisation of the study design, interpretation of data and revising the work critically for important intellectual content. DKG & JAG contributed to the drafting and revising of the work critically for important intellectual content as well as overall supervision. All authors read and approved the final manuscript.

The following publications have been published

Cassim N, Ahmad A, Wadee R, Rebbeck TR, Glencross DK and George JA (2020) Prostate Cancer age-standardised incidence increase between 2006 and 2016 in the Gauteng Province, South Africa: a laboratory data-based analysis.

S Afr Med J. 111 (1), 26-32 https://doi.org/10.7196/SAMJ.2020.v111i4850 (Appendix D)

Contributions to the paper by each author:

NC made substantial contributions to the conception or design of the work, acquisition of laboratory data, data analysis, preparing the first draft and final draft submitted. AA & RW assisted with the SNOMED lookup table development, data review and reviewing the work. TRR contributed to the conception or design of the work, data review, reviewing the work and editorial input. DKG & JAG contributed to the drafting and revising of the work critically for important intellectual content as well as overall supervision. All authors read and approved the final manuscript.

Cassim N, Ahmad A, Wadee R, George JA and Glencross DK. (2020) Using Systematized Nomenclature of Medicine (SNOMED) clinical terms (CT) codes to assign histological findings for prostate biopsies in the Gauteng Province, South Africa: Lessons learnt.

Afr J Lab. 2020; 9 (1):1-9 https://doi.org/10.4102/ajlm.v9i1.909 (Appendix A)

Contributions to the paper by each author:

NC developed the methodology, conducted the research and prepared the initial draft. All authors contributed to reviewing the initial draft. RW and AA assisted with the SNOMED lookup table development. JAG & DKG contributed to the drafting and revising of the work critically for important intellectual content as well as overall supervision.

Peer-reviewed conference proceeding

Authors: Cassim N, Mapundu M, Olago V George JA Glencross DK and George JA. (2019)

Title: Using big data techniques to improve prostate cancer reporting in the Gauteng Province, South Africa.

Date: 23-30 August 2019 (17th World Congress of Medical and Health Informatics).

Place: Lyon, France

Citation: Cassim, N., Mapundu, M., Olago, V., George, J. A., & Glencross, D. K. (2019). Using Big Data Techniques to Improve Prostate Cancer Reporting in the Gauteng Province, South Africa. Studies in health technology and informatics, 264, 1437–1438. https://doi.org/10.3233/SHTI190472

Conference proceedings

Authors: Cassim N, Ahmad A, Wadee R, Rebbeck TR, Glencross DK and George JA. (2019)

Title: Prostate Cancer age-standardised incidence increase between 2006 and 2016 in the Gauteng Province, South Africa: a laboratory data-based analyses.

Date: 18-21 July 2018 (Path Red conference).

Place: Bredell, South Africa

Authors: Cassim N, Mapundu M, Olago V, Wadee R, George JA and Glencross DK. (2019)

Title: Big data approaches to extract clinical information from prostate biopsy results in the Gauteng Province.

Date: 18-21 July 2018 (Path Red conference).

Place: Bredell, South Africa

Authors: Mapundu M, Olago V, Cassim N, Wadee R, George JA and Glencross DK. (2019)

Title: Using supervised machine learning algorithms to assign prostate biopsy histological findings.

Date: 22 August 2018 (School of Public Health Research Day).

Place: Johannesburg, South Africa

Grants

National Health Laboratory Service (NHLS) development grant – R79 000

Awards

School of Public Health Research Day best oral presentation for the paper “Using supervised machine learning algorithms to assign prostate biopsy histological findings”.

Abstract

Background:

Prostate cancer (PCa) is the leading male neoplasm in South Africa, with an age-standardised incidence rate (ASIR) of 68.0 per 100,000 population reported in 2018. PCa is also the most diagnosed neoplasm among men globally. Local studies have reported that Black African men present with higher grade and stage disease, higher serum PSA and less often receive potentially curative treatment than men of other race groups. Updated local guidelines recommend informed patient-based screening for males with a life expectancy ≥10 years, commencing at 40 years for Black Africans. PCa risk categories are defined using the total prostate specific antigen (PSA), Gleason score (GS) and clinical stage. The GS is the strongest predictive factor for treatment. The updated grade group (GG) was developed based on the GS as follows: (i) GS ≥6: GG1, (ii) GS 3 + 4 = 7: GG2, (iii) GS 4 + 3 = 7: GG3, (iv) GS = 8: GG4 and (v) GS ≥9: GG5. One of the challenges is the tedious manual extraction of an adenocarcinoma histological finding and the GS as this information is embedded within the semi-structured narrative prostate biopsy report. There is a paucity of local data, with most studies reporting data in a urological setting with small sample sizes, based at one or more academic hospitals. Data mining has the potential to turn narrative reports into information by applying various computational techniques.

Aims:

The aim of this thesis was to describe PSA testing and histological diagnosis of PCa using laboratory data for men attending public-health facilities in the Gauteng Province. A further aim was to use text mining to extract and describe GS, specifically for Black Africans. A further aim was to assess trends in PSA testing for primary health care facilities. Automated methods to extract PCa information to reduce the burden of manual coding were a further aim.

Methods:

A retrospective descriptive study design was used to analyse prostate biopsy and PSA laboratory data between 2006 and 2016. The inclusion criteria were men ≥30 years. The Systematized Nomenclature of Medicine (SNOMED) clinical terms (CT) morphology (M) and topography (T) codes were used to develop lookup tables to assign the following histological findings: (i) diagnosis (benign/malignant) (ii) sub-diagnosis (adenocarcinoma) and (iii) sub-result (inflammation type). Two experts manually coded the diagnosis after reading the narrative prostate biopsy report to assess the positive predictive value (PPV) of the pathologist assigned SNOMED CT codes. For 1000 randomly selected prostate biopsies with PCa, predictive analytics and text mining was used to automate the extraction of the GS, reporting precision, recall and the F score. For the prostate biopsy data, data was reported for only prostatic biopsies. M or T codes were done manually. The GS was manually coded for an

adenocarcinoma sub-diagnosis. PCa was defined as an adenocarcinoma histological finding with the GS reported. We reported associations of PCa with a GG ≥4 for race group. Both the biopsy and PSA were de-duplicated using the CDW unique patient identifier to report: (i) ASIR and (ii) develop a presentation cohort (first-ever PSA). The PSA data was analysed to report patient numbers by calendar year, age category and race group as well as descriptive statistics. We used logistic regression to assess any association for race group and age with a PSA ≥4µg/L.

Results:

SNOMED CT lookup tables were able to report the histological findings for 88% of biopsies. The manual coding revealed a PPV of 0.96 for the pathologist assigned SNOMED CT codes. The predictive analytics and text mining accurately extracted the GS for all 1000 biopsies (F score of 1). There were 22 937 prostatic biopsies referred to NHLS between 2006 and 2016. Of these, a PCa finding was reported for 39% of Black Africans. A high-risk GS was reported for 46% Black Africans. Multiple logistic regression revealed that Black Africans were more likely to have PCa with a GG ≥4 with an odds ratio of 1.45. The ASIR increased from 44.9 in 2006 to 57.3 per 100 000 population by 2016. There were239 506 (86.2%) patients with a first-ever PSA, compared to 277 983 tests. Between 2006 and 2016, numbers of men tested increased from 1 782 to 67 025 respectively, with 186 984/239 506 (78.1%) of tests from clinics. The majority of testing was for men in the 50 - 59 age category and Black Africans. The logistic regression reported that the odds of having a PSA ≥4µg/L was significantly lower for Indian/Asians, Coloureds and Whites than for Black Africans.

Conclusions:

The feasibility of using SNOMED CT codes to automate PCa data was confirmed. In addition, the reliable extraction of the GS to assess late presentation and prognosis was confirmed. These approaches can be applied to national PCa data. Our findings reveal that Black African men are significantly more likely to present with a PCa with a GS ≥8 compared with other racial groups. Our data suggest that predominantly healthy patients were tested which in turn is indicative of some population-based screening. Local public-sector guidelines need to be aligned to the 2017 urological evidence-based recommendations. There is a need for additional research to understand why Black African men present with higher grade disease.

Acknowledgements

I wish to express my thanks and gratitude to the following people for their assistance, guidance and support:

- My supervisors Professors George and Glencross for their contributions to the study design, manuscript review and comments on the various chapters.
- Professor Mohammed Haffejee and Martin Hale for providing insight into the local context for prostate cancer work.
- Dr Innocent Maposa for help with statistical analysis.
- Dr Anushka Ajith and Marietha Nel for proof reading.
- The National Cancer registry for assisting with the hot deck race imputation for the study.
- The corporate data warehouse for implementing the probabilistic matching algorithm.
- I am indebted to the staff of the National Health Laboratory Service (NHLS) that generated the prostate specific antigen (PSA) and prostate biopsy data.

CONTENTS

Dedication iii
Authors contribution to the work iv
Publications arising from this thesis v
Contents xi
List of TABLES xv
List of Figures xviii
List of Abbreviations xxi
Preface xxiv
Chapter 1 – literature review 1
1.1 Introduction 1
1.2 Purpose of the literature review 1
1.3 Summary and research gaps 1
1.4 Aims 2
1.5 Scope of the literature review 2
1.6 Non-Communicable Diseases (NCD) 3
1.7 Cancer 3
1.8 Prostate cancer 4
1.9 Prostate cancer epidemiology 6
1.10 How prostate cancer biopsies are processed and reported 7
1.11 Urological guidelines for prostate cancer screening and diagnosis in South Africa 10
1.12 Local public sector and urological prostate cancer screening and diagnosis guidelines in South Africa between 2006 and 2016 13
1.13 Other prostate cancer guidelines 15
1.14 Controversy about prostate-specific antigen testing 16
1.15 Global prostate cancer studies 19
1.16 African prostate cancer studies 21
1.17 Local prostate cancer studies 26
1.18 National prostate cancer registry data 30
1.19 Challenges with the manual extraction of prostate biopsy data and the way forward 31
Chapter 2 - Using Systematized Nomenclature of Medicine (SNOMED) clinical terms (CT) codes to assign histological findings for prostate biopsies in the Gauteng PROVINCE, South Africa: Lessons learnt 37
2.1 List of definitions for terms used in this chapter 38
2.2 Introduction 39
2.3 Aim and Objective 41
2.4 Methods 41
2.4.1 Study design 41

2.4.2 Inclusion /Exclusion Criteria 41
2.4.3 Sample Population 41
2.4.4 Ethical considerations 41
2.4.5. Data extraction and preparation 41
2.4.6 Combining SNOMED descriptions for the lookup table 43
2.4.7 Coding the SNOMED M and T lookup tables 45
2.4.8 Combining the lookup tables with the prostate biopsy data 45
2.4.9 SNOMED M and T code descriptive analysis 48
2.4.10 Descriptive analysis of diagnosis and sub-diagnosis 48
2.4.11 Evaluating the accuracy of the SNOMED CT codes 48
2.5 Results 48
2.5.1 SNOMED M and T code descriptive analysis 48
2.5.2 Descriptive analysis of diagnosis and sub-diagnosis 49
2.5.3 Evaluating the accuracy of the SNOMED CT codes 51
2.6 Discussion 52
2.7 Conclusion 55
Chapter 3 - Using text mining techniques to extract prostate cancer prognostic information (Gleason score) from semi-structured narrative laboratory reports in the Gauteng Province, South Africa 57
3.1 Introduction 61
3.2 Materials and methods 63
3.2.1 Study design 63
3.2.2 Ethical considerations 63
3.2.3 Text mining algorithm development 63
3.2.4 Data acquisition 65
3.2.5 Pre-processing 66
3.2.6 Feature extraction 68
3.2.7 Feature value representation 68
3.2.9 Information extraction 71
3.2.10 Classification 71
3.2.11 Discovered knowledge 71

3.2.12 Text Mining Algorithm Evaluation 71
3.2.13 Statistical analysis 72
3.3 Results 72
3.3.1 Text Mining Algorithm performance 72
3.3.2 Text Mining precision and recall 72
3.3.3 Gleason score formats reported 73
3.3.4 Gleason score frequency analysis 74
3.3.4 Gleason risk category analysis 74
3.3.5 Validation analysis 75
3.4 Discussion 76
3.5 Conclusion 78
Chapter 4 - Retrospective analysis of prostate-specific antigen testing trends and primary health care public-sector guidelines in the Gauteng Province, South Africa, between 2006 and 2016 79
4.1 Introduction 82
4.2 Methods 83
4.2.1 Study design 83
4.2.3 Ethical considerations 83
4.2.4 Data extract 83
4.2.4 Data preparation 84
4.2.5 Statistical Methods 85
4.3 Results 85
4.3.1 Number of patients receiving first-ever PSA test 87
4.3.2 Number of patients receiving a first-ever PSA test by age category and race group 88
4.3.3 Descriptive statistics by age and race group 91
4.3.4 Association between an elevated PSA with age and race group 92
4.4 Discussion 93
Chapter 5 - Prostate Cancer age-standardised incidence increase between 2006 and 2016 in the Gauteng Province, South Africa: a laboratory data-based analysis 97
5.1 Introduction 99
5.2 Methods 100
5.2.1 Study design 100

5.2.2 Inclusion criteria 100
5.2.3 Biopsy data preparation 100
5.3 Results 102
5.3.1 Prostate biopsy outcomes by age and race group 102
5.3.2 Analysis of low, intermediate and high-risk Gleason scores 106
5.5 Discussion 110
Chapter 6 – Discussion 113
6.1 Summary of findings 113
6.1.1 SNOMED CT Lookup Tables 115
6.1.2 Using text mining and predictive analytics to extract the GS 115
6.1.3 Primary health care first-ever PSA trends 116
6.1.4 Prostate biopsy 117
6.1.5 Prostate cancer age-standardised incidence 118
6.1.6 Implications of our findings 119
6.2 Conceptual framework and practical aspects for local and national implementation 126
6.3 Areas of future research 127
6.4 Limitations and challenges 128
6.5 Recommendations 129
6.6 Conclusions 131
7. References 133
Appendix A – PUBLICATION 151
Appendix B – PUBLICATION 161
Appendix C – PUBLICATION 163
Appendix D – PUBLICATION 166
Appendix E – Ethics and other approval letters 174
Appendix F - Declaration and co-author signature 176
Appendix G - Similarity report 178
Appendix H - Detailed information on the HIV M&E dashboard 182
Appendix I – Other HIV M&E Dashboards 184
Appendix J – Additional visualisations for the text mining algorithm 186

LIST OF TABLES

No	Table Description	Page
1.1	Grade Group (GG) reporting indicating the corresponding Gleason score	9
1.2	Recommended terminology to be used to report prostate biopsy histological findings	10
1.3	The tumour, node, metastasis (TNM) classification system used for cancers	12
1.4	Risk stratification for planning treatment options	12
1.5	Risk of prostate cancer (PCa) in relation to the total prostate specific antigen (PSA) result and Gleason score (GS)	13
1.6	Local and other urological prostate cancer (PCa) guidelines	15
1.7	Prostate cancer recommendation from other countries	16
1.8	Studies providing conflicting prostate screening data	18
1.9	The GLOBOCON prostate cancer age-standardised incidence rate per 100 000 population in 2012 and 2018. The number of new cases are also reported	21
1.10	Number of prostate cancer cases and the age-standardised incidence rate per 100 000 population is reported for 46 African countries in the 2018 GLOBOCON study. Data for is study is sorted by age-standardised incidence rate in descending order. The matching highest the age-standardised incidence rate per 100 000 population reported by the Adeloye *et al* is reported in the last column	23
1.11	Pooled age-standardised incidence rate by age categories for African countries	24
1.12	Findings from African studies	25
1.13	Percentage of Black Africans for each study compared to the Census 2011 data for each respective Province	28
1.14	Findings from local prostate cancer (PCa) studies	29
1.15	Table describing how the semi-structured narrative prostate biopsy data meet the six criteria for big data	34
2.1	Example of four prostate biopsies where the SNOMED code descriptions were assigned a diagnosis and sub-diagnosis including the allocation of ICD-0-3 codes. The biopsy result text are also provided. Episode numbers are anonymised	47
2.2	Contingency table to assess the percentage of SNOMED M and T codes populated using the mapping table for prostate biopsies between 2006 and 2016 in the Gauteng Province, South Africa	49
2.3	Top ten most commonly requested SNOMED M code combinations from the prostate biopsy data between 2006 and 2016 in the Gauteng Province, South Africa	49
2.4	Descriptive analysis of diagnosis and sub-diagnosis where both a SNOMED M and T code are populated of prostatic origin between 2006 and 2016 in the Gauteng Province, South Africa. Where appropriate, the ICD-0-3 codes are provided in brackets	51
2.5	Contingency table used to determine the accuracy of the pathologist assigned SNOMED CT codes for 1000 random prostate biopsies with a 'Benign/negative for malignancy' or 'Neoplasm, malignant' diagnosis. These biopsies were randomly selected from prostate biopsies that were performed between 2006 and 2016 in the Gauteng Province, South Africa.	52

No	Table Description	Page
3.1	Example of the semi-structured narrative prostate biopsy report. The narrative biopsy report included the headings clinical history, macroscopy and pathological diagnosis	62
3.2	Description of the various python libraries used for the text mining algorithm to extract the Gleason score from narrative prostate biopsy reports.	64
3.3	N-grams feature extraction output for a sample of biopsies	68
3.4	Performance of the text mining algorithm to automate the extraction of the Gleason score from narrative prostate biopsy narrative reports. A contingency table was used to compare the manually coded and algorithm predicted values. We reported the precision, recall and F-score reported for the first and updated text mining algorithm output as well as for the validation dataset.	73
3.5	Different Gleason score formats reported. The clean extracted score reported, and the original value reported in the prostate biopsy report is indicated.	74
3.6	Biopsy volumes for the five most commonly reported Gleason scores. The table reported the Gleason score and frequency for the top five reported scores with the remaining scores grouped and reported as “Others”.	74
3.7	Comparison of low, intermediate and high-risk Gleason scores for the predicted and manually coded values. The macro-average F-score is reported.	75
3.8	Biopsy volumes for the five most commonly reported Gleason scores for the validation dataset. The table reported the Gleason score and frequency for the top five reported scores with the remaining scores grouped and reported as “Others”.	76
4.1	Descriptive statistics for a first-ever total prostate-specific antigen (PSA) test reported for age category and race group between 2006 and 2016 in the Gauteng Province, South Africa. The overall median and interquartile range (IQR) as well as the proportion of men with a with a first-ever total PSA test ≥4, <10 (low-risk), 10 - 19.9 (intermediate-risk) and ≥20µg/L (high-risk) are reported	92
4.2	Logistic regression to assess the association between samples with a first-ever total prostate-specific antigen (PSA) ≥4µg/L and race group as well as age category between 2006 and 2016 in the Gauteng Province, South Africa. We controlled for race group and age category.	93
5.1	Prostate biopsy descriptive statistics by race group. Prostate biopsy descriptive data for each race group reported as a table for the Gauteng Province between 2006 and 2016. Mean age (range), biopsy numbers and percentages reported for age category, race group, biopsy results and prostate cancer results. One-way analysis of variance (ANOVA) reported for age as a continuous variable and Chi-square test used to identify whether there was a statistical difference for age category and biopsy findings for the four race groups	103
5.2	Multiple logistic regression to assess the association between prostate cancer outcomes and Gleason score and race group and age category. For the dependent variable, a grade group	105

No	Table Description	Page
	(GG) of 4 or 5 was coded as 1 (GS ≥8) and GG 1-3 as 0 for a prostate cancer histological finding with a Gleason score (GS) reported	
6.1	Consolidated finding	113
6.2	Table describing some of the studies that could be undertaken once the Corporate Data Warehouse target area for Non-Communicable Diseases has been developed. Details are provided for the disease, laboratory test/s, study objectives and outcomes assessed	125

LIST OF FIGURES

No	Figure Description	Page
1.1	Anatomy of the prostate gland	5
1.2	Risk factors for prostate cancer (PCa)	7
1.3	Typical anatomical pathology workflow	8
1.4	Prostate cancer age-standardised incidence rate per 100 000 population per country in 2018 for the most commonly diagnosed neoplasm reference number. Data for South Africa is reported as a red bar	20
1.5	Bar chart reporting National Cancer Registry national age-standardised incidence rate (ASIR) estimates for South Africa between 1996 and 2014	31
1.6	The six criteria for big data	33
1.7	Illustrating the use of big data and the application of artificial intelligence	35
2.1	High-level overview of the steps taken to code each chained SNOMED code description combination. There were two separate SNOMED data extracts from the laboratory information system; morphology (M) and topography (T). The colour coding indicates the various processes: (i) green: data extracts, (ii) yellow: SNOMED code manipulation in preparation for lookup table development and (iii) orange: lookup tables with coded variables. The data extract was prepared using the two lookup tables to generate the following new coded	43
2.2	Six step procedure used to transform the chained comma separated SNOMED 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 SNOMED 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-00100 code description is 'Normal tissue (finding)' and (6) Microsoft Excel CONCATENATE function was used to combine code descriptions in a new column	44
2.3	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 SNOMED 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 (SQL) could be used to combine the required data for analysis	46
3.1	Diagram describing the logical processes used to analyse the raw narrative prostate biopsy report to generate the discovered knowledge. The steps were as follows: (i) data acquisition (ii) pre-processing and (iii) feature extraction, (iv) feature value representation, (v) feature selection, (vi) information extraction (vii) classification and (viii) discovered knowledge.	65

No	Figure Description	Page
3.2	Diagram depicting the visualisation of the corpus before (A) and after cleaning (B). The larger the text the more important and frequent the term is in the narrative biopsy reports	67
3.3	Horizontal bar graph depicting the top twenty occurring unigrams (A), bigrams (B), trigrams (C) and quadgrams (D). The number of occurrences is displayed on the x-axis.	70
4.1	Flow chart depicting all the data steps to generate the data used to create the first-ever PSA cohort for primary health services between 2006 and 2016 in the Gauteng Province, South Africa	86
4.2	Percentage year on year change reported as a bar chart for patients with a first-ever total PSA test at PHC services in the Gauteng Province, South Africa. The annual test volumes were reported as a line chart. As lower test volumes were reported in 2012 and 2013 (indicated by the dotted line), the 2014 and 2015 percentage year on year change calculations were based on the 2011 numbers.	87
4.3	Annual PSA test volumes by unit type for primary health care (PHC) services. The annual PSA volumes are reported for clinics (blue bars) and community health centres (CHC) (dark grey bars) between 2006 and 2016 in the Gauteng Province, South Africa	88
4.4	Line chart reporting the number of patients with a first-ever total prostate-specific antigen (PSA) test by year and age category (A) and race group (B) for primary health care (PHC) facilities between 2006 and 2016 in the Gauteng Province, South Africa	90
4.5	Bar chart reporting the number of patients with a first-ever total prostate-specific antigen (PSA) test by race group and age category for PHC facilities between 2006 and 2016 in the Gauteng Province, South Africa.	91
5.1	Age distribution for prostate cancer findings. Biopsy numbers reported by age category for both prostate cancer and non-prostate cancer histological findings in the Gauteng Province between 2006 and 2016 as a population pyramid.	104
5.2	Line chart reporting prostate biopsy numbers for the Gauteng Province between 2006 and 2016. Stacked bar charts indicating annual prostate biopsy numbers. The calendar year reported on the X-axis and the number of prostate biopsies on the Y-axis	106
5.3	Percentage of prostate biopsies with a low (GS ≤6), intermediate (GS =7: 3 + 4 =7 and 4 + 3 =7), and high (GS ≥8) risk Gleason score by race group. Bar chart used to report on the percentage of prostate biopsies with prostate cancer finding by GS risk categories. The grade group (GG) is indicated in brackets, with 4 and 5 combined to report a GS ≥8. The percentages reported as data labels per race group	107
5.4	Percentage of prostate biopsies with a low (GS ≤6), intermediate (GS =7: 3 + 4 =7 and 4 + 3 =7), and high (GS ≥8) risk Gleason score by calendar year. Bar chart used to report on the percentage of prostate biopsies with prostate cancer finding by GS risk categories. The grade group (GG) is indicated in brackets, with 4 and 5 combined to report a GS ≥8. The percentages reported as data labels per calendar year	108
5.5	Age-standardised prostate cancer incidence rates between 2006 and 2016 reported as orange dots with the average annual percentage change (AAPC) reported as a blue line. The year of diagnosis reported on the X-axis and the age-adjusted rate on the Y-axis	109
6.1	Prostate cancer conceptual framework	120
6.2	Flow chart depicting the relationship between guidelines, PSA and prostate biopsy testing. The information that PSA and prostate biopsy data informs is also indicated.	122

No	Figure Description	Page
6.3	Database diagram showing how the SNOMED CT lookup tables and text mining application programming interface (API) could automate the generation of a PCa diagnosis. The private sector prostate cancer data extract could also be accommodated through the CDW process. The Expert Committee would code all new SNOMED CT M and T code combinations to update the CDW lookup tables.	123
6.4	Prostate cancer conceptual framework	127
6.5	HIV Monitoring and Evaluation Facility Indicators and Trends (All Ages) dashboard screenshot (taken from the National Institute for Communicable Diseases (NICD) monitoring and evaluation online HIV programme dashboards). The dashboard displayed is titled 'HIV M&E Facility Indicators and Trends (All Ages)'.	130

LIST OF ABBREVIATIONS

Abbreviation	Description
AAPC	Annual Average Percentage Change
AARQA	Academic Affairs, Research and Quality Assurance
AFR	World Health Organisation African region
AI	Artificial Intelligence
AMR	Antimicrobial Resistance
ANOVA	Analysis of Variance
AP	Anatomical Pathologist
API	Application Programming Interface
ART	Antiretroviral Therapy
ASIR	Age-standardised incidence rate
AUA	American Urological Association
BPH	Benign Prostatic Hyperplasia
CDW	Corporate Data Warehouse
CHB	Chris Hani Baragwanath Hospital
CHC	Community Health Centre
CI	Confidence Interval
CMJAH	Charlotte Maxeke Johannesburg Academic Hospital
CSV	Comma separated values
CT	Clinical terms
DGM	Dr George Mukhari
DHIS	District Health Information System
DRE	Digital Rectal Examination
EAU	European Association of Urology
EML	Essential medicines list
ERSPC	European Random Study of Screening for Prostate Cancer
ESMO	European Society for Medical Oncology
ETL	Extract, Transform and Load
FN	False Negative
FP	False Positive
FPSA	Free PSA
FT	Free to total PSA ratio
GG	Gleason grade group
GS	Gleason score

Abbreviation	Description
GWAS	Genome wide association studies
HbA1c	Glycated Haemoglobin
HCW	Health care worker
HIE	Health Information Exchange
IARC	International Agency for Research on Cancer
ICD	International Classification of Diseases
ICD-0-3	International Classification of Diseases for Oncology – 3rd Edition
ICD-10	International Classification of Diseases, Tenth Revision
ICPCG	International Consortium for Prostate Cancer Genetics
IDE	Integrated Development Environment
IQR	Interquartile range
ISUP	International Society of Urological Pathology
LIS	Laboratory Information System
LMIC	Low-middle income country
LTFU	Loss-To-Follow Up
LUTS	Lower urinary tract symptoms
M	Morphology
M&E	Monitoring and Evaluation
ML	Machine learning
NCCN	National Comprehensive Cancer Network
NCCP	National Cancer Control Programme
NCD	Non-Communicable Diseases
NCR	National Cancer Registry
NDoH	National Department of Health
NHLS	National Health Laboratory Service
NICD	National Institute for Communicable Diseases
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
ODS	Operational Data Store
OR	Odds ratio
PBCR	Population-based cancer registry
PCa	Prostate Cancer
PCA3	Prostate cancer antigen 3
PCF	Prostate Cancer Foundation of South Africa
PHC	Primary Health Care

Abbreviation	Description
PMTCT	Prevention of Mother-to-Child Transmission
PSA	Prostate specific antigen
PSAV	PSA velocity
RCT	Randomised controlled trail
SAPCS	South African Prostate Cancer study
SBA	Steve Biko Academic Hospital
SCM	Supply chain management
SEER	Surveillance, Epidemiology, and End Results
SNOMED	Systematized Nomenclature of Medicine
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
SSA	Sub-Saharan Africa
STATS SA	Statistics South Africa
STG	Standard treatment guidelines
T	Topography
TB	Tuberculosis
TN	True Negative
TNM	The tumour, node and metastasis classification system
TP	True Positive
TRUS	Transrectal ultrasound
USA	United States of America
USPSTF	The United States Preventative Services Task Force
UTI	Urinary Tract Infection
WHO	World Health Organisation
WSP	World Standard Population

PREFACE

This PhD thesis is presented as a number of chapters as follows:

- Chapter 1 is the literature review.
- Chapters 2 to 5 contain the results of the studies undertaken for this PhD. Each chapter contains the results and discussion for each research article (published articles are provided as appendices).
- Chapter 6 is the conclusions for the study as well recommendations for further work.

CHAPTER 1 – LITERATURE REVIEW

1.1 INTRODUCTION

This chapter will discuss the related literature reviewed for this study which outlines the key concepts related to this work.

1.2 Purpose of the literature review

The purpose of the literature review is to identify and critically evaluate the available information, in the form of journal articles, policies and guidelines.

1.3 Summary and research gaps

There is limited local PCa data, with most studies using small sample sizes to predominantly evaluate patients presenting at urology clinics at a regional or tertiary hospital [1-6]. Data from a urology setting for patients referred from lower levels of care with a suspicion of PCa is then not representative of the broader male population. These studies also used small samples sizes affecting generalisability [1-6]. These studies reported late presentation with advanced PCa (PSA ≥20µg/L and a GS ≥8) disproportionately affecting Black African men, resulting in a poorer prognosis [1-3].

The NCR reports annual PCa incidence data at the national level from data from both the public and private sector laboratories [7]. The latest NCR data is reported for the 2016 calendar year in 2020, with 8 332 incident cases of PCa and an ASIR of 47.48 [8]. This data is unfortunately not available at the provincial level. To obtain findings, all local data was generated by manually reading the prostate biopsy semi-structured reports to code the histological findings. This is a time-consuming process with the potential for transcription errors as well as delays in PCa surveillance data. One of the challenges with NCR data is that the absence of GS findings makes it difficult to assess late presentation. The absence of the GS is significant given the local data indicating that late presentation disproportionally affects Black African men [1-3].

There is a need for comprehensive PCa data that includes the PSA results, prostate biopsy histological findings and GS for an entire province that will improve both generalisability and provide detailed data at a level below national reporting [7]. Furthermore, there is a need to move away from manual coding to speed up the reporting of local PCa data. This has never been attempted locally. Automating the reporting of histological findings as well as the GS makes it possible to report data on a regular basis to assess late presentation for the entire

population. This will provide important insights that have not been reported locally for a large sample size.

1.4 Aims

The aim of this study was to conduct a retrospective study to describe prostate cancer screening and histological diagnosis using prostate and PSA routinely collected laboratory data. A further aim was to describe presentation specifically for Black Africans, describe first-ever PSA testing trends for PHC services and to developed automated methods to reduce the burden of manual coding.

The objectives of the study were to use laboratory data to:

- Develop lookup tables to automate the reporting of prostate biopsy data.
- Develop a text mining algorithm to extract the GS.
- Describe the number of prostate biopsies performed annually.
- Describe the number of prostate biopsies by race group and age category.
- Describe the number of prostate biopsies with a PCa histological finding (Adenocarcinoma and GS reported).
- De-duplicate the prostate biopsy data to identify incident patients that were diagnosed with PCa to report ASIR.
- Describe the number of patients presenting with advanced disease (GS ≥8).
- Describe first-ever PSA testing trends for PHC services in the Gauteng Province.

1.5 Scope of the literature review

The literature review is presented according to the following headings:

- Non-communicable diseases (NCD).
- Cancer
- Prostate cancer
- Prostate cancer guidelines
- How prostate cancer biopsies are processed and reported
- Global prostate cancer data
- African prostate cancer data
- Local prostate cancer data
- Local cancer registry data
- Summary and research gaps

1.6 Non-Communicable Diseases (NCD)

Non-communicable diseases (NCD) are due to a combination of genetic-, physiological-, environmental- and behavioural factors [9]. Globally, the most prevalent NCD include cardiovascular, cancer, chronic respiratory disease and diabetes [10]. In South Africa, cardiovascular, diabetes mellitus, respiratory diseases and cancers contributed to 12% of the overall disease burden [11]. Unfortunately, the human-, social- and economic consequences of NCD are felt mainly in poor and vulnerable populations [12]. In South Africa, the burden of NCD disproportionately affects urban poor people [11]. Globally, modifiable and metabolic risk factors are the most common causes of NCD [13]. These modifiable risk factors include physical inactivity, tobacco use, unhealthy diet and the harmful use of alcohol [13]. The metabolic risk factors include raised blood pressure, obesity, hyperglycaemia and hyperlipidaemia [13]. Both the modifiable and non-modifiable risk factors would be different for each NCD, e.g. cancer. For modifiable risk factors, the cancers of interest are lung, liver, esophageal, liver, skin, cervical, stomach, colorectal, etc [14].

Insufficient physical activity is the leading modifiable risk factor for NCD related deaths, causing approximately 3.2 million global NCD deaths per annum [12]. This has been linked to colon and breast cancer [14]. Furthermore, tobacco usage increases the risk of cardiovascular, cancer, chronic respiratory disease and diabetes [12]. Smoking has been linked to a number of cancers in the lung, oral cavity, pharynx, oesophagus, stomach, colorectum, liver, pancreas, etc [14]. Excessive dietary sodium (salt) consumption is associated with an increased risk of hypertension and cardiovascular disease [12]. The harmful use of alcohol also increases the risk of developing NCD [12]. Alcohol intake has been associated with lip, oral cavity, pharynx, oesophageal, colorectal, larynx and breast cancer [14].

In 2008, there were 56 million deaths globally, of which 38 million (68%) were due to NCD. Most of these deaths were in low- and middle-income countries (LMIC) [15]. In South Africa, the probability of dying from the four main NCD is ≥25%, which are cardiovascular, cancer, respiratory disease and diabetes [16, 17]. By 2010, NCD accounted for 39% of all deaths in South Africa [17].

1.7 Cancer

Cancer is a generic term that describes a group of diseases that are characterised by the growth of abnormal cells beyond their usual boundaries [18]. The abnormal cancerous cells can invade adjoining organs [18]. Other terms for cancer include malignant tumours and neoplasms [18]. Cancer is the leading cause of global mortality for countries of all income

levels [19]. The number of cancer cases are projected to increase rapidly, fuelled by population growth, increasing life expectancy, and the adoption of lifestyle behaviours that increase cancer risk [19, 20]. Cancer can affect most parts of the body and has many anatomic and molecular subtypes that each require specific management strategies [21]. Globally, the most common male cancers are lung, prostate, colorectal, stomach and liver [21].

While NCD are responsible for the majority of global deaths in 2018, cancer is expected to rank as the leading cause of death and an important barrier to increasing life expectancy across the globe in the 21st century. [22]. In 2018, it was reported that cancer is expected to be the primary or secondary global cause of death before the age of 70 years for 91 of 172 countries (52.9%) [22]. In developed area such as the Americas and Australia, cancer ranks as the number one cause of premature death compared to 3rd to 4th in South Africa [22]. The lower ranking in South Africa is most likely due to the contribution of infectious diseases such as HIV and TB as well as the population distribution [23].

The GLOBOCON 2018 incidence estimates were produced by the International Agency for Research on Cancer (IARC) [22]. This study estimated that 18.1 million new cases and 9.6 million cancer deaths were reported globally in 2018 [22]. The most common male neoplasm was prostate cancer (PCa) for the North and South American, Australian, European and for sub-Saharan Africa (SSA) regions [22]. For two SSA countries, Kaposi sarcoma was the most reported male cancer [22]. The northern African countries (not SSA) reported multiple cancers including lung, liver and non-Hodgkin lymphoma [22]. Lung cancer was the most common male neoplasm for the Asian continent [22, 24].

Globally, lung cancer is the leading cause of mortality for 93/186 countries [22]. PCa mortality was highest in 46/186 countries followed by liver cancer (20/186) [22]. The PCa mortality rates do not follow the patterns reported for incidence, with elevated mortality rates reported in Sub-Saharan Africa and the Caribbean [22]. The highest PCa mortality rates were reported Benin, South Africa, Zambia, Barbados, Jamaica, and Haiti [22]. In South Africa, lung cancer is the leading cause of cancer-related mortality [22].

1.8 Prostate cancer

The prostate gland is similar in size and shape to a walnut and forms part of the male reproductive system [25]. Anatomically, the prostate gland is situated low in the male pelvis, below the bladder and in front of the rectum [25].

The prostate gland is divided into several anatomical zones or regions: (i) peripheral, (ii) transition, (iii) central and (iv) anterior fibromuscular stroma [25]. The peripheral zone is the largest area of the prostate that can be easily examined with a digital rectal exam (DRE) [26]. The transition zone is situated in an area where the urethra passes through the prostate [25]. The central zone is situated behind the transition zone through which the ejaculatory ducts pass [25]. The anterior fibromuscular stroma is the thickened area of tissue that surrounds the apex of the prostate [25].

The peripheral, the central and the transition zones comprise approximately 70%, 25% and 5% of the prostatic glandular tissue (Figure 1.1) [27]. The majority of PCa are located in the peripheral zone (70-80%), compared to between 20% and 30% arise in the transition zone [26, 27].

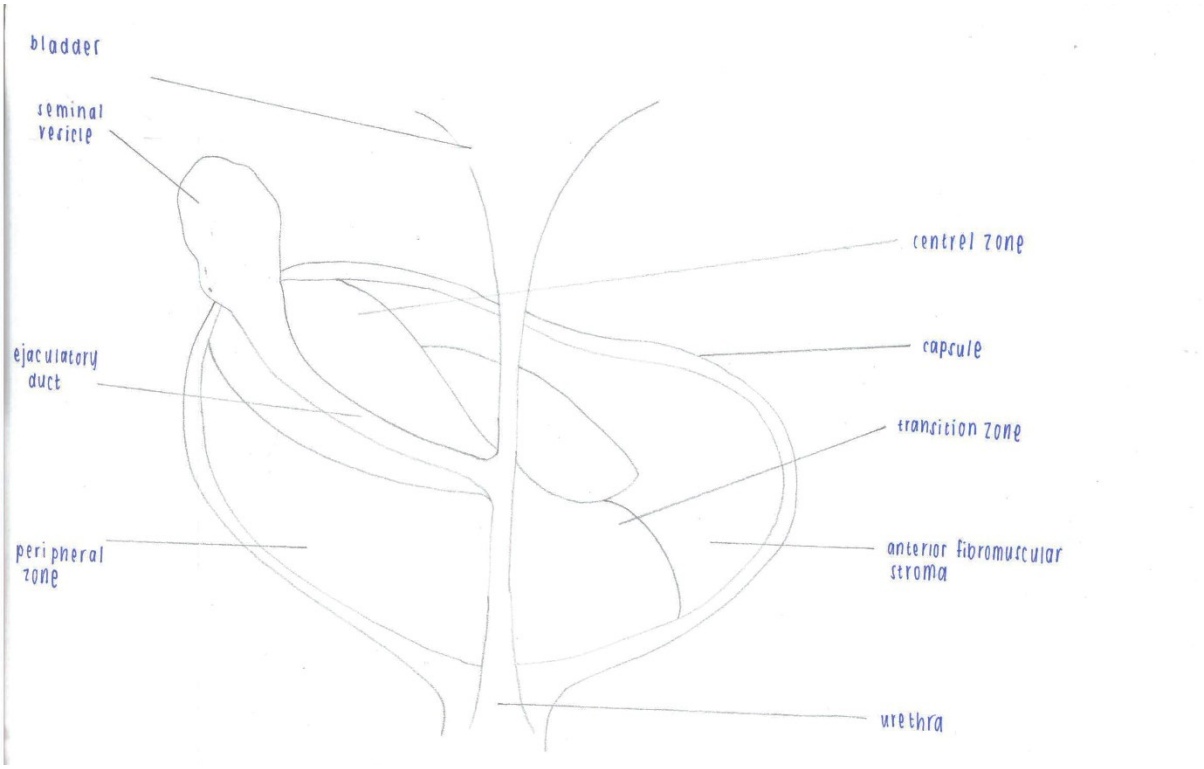


Figure 1.1 Anatomy of the prostate gland

Source: Hand drawing

The function of the prostate is to secrete one of the components of semen, which is a milky fluid that carries sperm from the testicles [18]. The prostate gland tends to grow larger as men age [18]. When the growth of the prostate gland squeezes on the urethra, it may cause lower urinary tract symptoms (LUTS) [28]. The LUTS symptoms include voiding, hesitancy, poor and/or intermittent stream, straining, prolonged micturition, feeling of incomplete bladder emptying, dribbling, urge incontinence, and nocturia [29].

Growing older is associated with an increased risk of prostate disease [15]. The most common prostate diseases include: (i) prostatitis, (ii) benign prostatic hyperplasia (BPH), and (iii) PCa [30]. Two studies have reported that prostatitis can significantly increase the odds for PCa [31, 32]. Similarly, a systematic review reported a significant association between BPH and PCa [33]. Prostatitis is an inflammation of the prostate gland that may be caused by acute or chronic infections [34]. BPH is a benign (non-malignant) enlargement of the prostate gland and refers to the stromal and glandular epithelial hyperplasia (abnormal cell growth) [25]. PCa occurs when the cells in the prostate start to grow uncontrollably [35]. The majority of PCa's are histologically confirmed as adenocarcinoma [35]. Other types of PCa include sarcomas and carcinomas which are rare [30]. Several risk factors have been identified for PCa including age (>50 years), race group (African men), family history and diet [30].

Due to increasing life expectancy, a growing world population and improvements in public health, PCa is a growing concern for global epidemiology [36, 37]. The global burden of PCa is high with one million cases histologically diagnosed annually with an annual mortality rate of 300 000 [36]. PCa is the second most frequently diagnosed cancer among men globally, but in South Africa it is the leading neoplasm in men [38].

1.9 Prostate cancer epidemiology

The most consistently reported risk factors for PCa are age, race group and family history [15]. PCa occurs predominantly in men over the age of 40 years [39]. Data from the histological examination of prostate biopsies and autopsy studies worldwide have indicated that the PCa age-standardised-incidence rate (ASIR) per 100 000 population increases with age [39]. Age is, therefore, the most important risk factor for PCa [15]. Several studies have indicated that PCa is higher among men of Black African descent who also have a higher stage of disease at presentation and poorer prognosis [36, 39-42]. There is significant epidemiological evidence for an increased PCa risk for African Americans compared to Asian Americans suggestive of both genetic and racial differences [39]. The risk of developing PCa doubles in men with a father or brother affected by PCa (family history) [41]. These findings suggest that racial and genetic factors are risk factors for PCa.

A number of genetic studies have reported several PCa risk loci identified through linkage studies [39]. The strongest linkage was noted on chromosome one, with candidate genes including HPC1 on chromosome 1q23-35, PCAP on chromosome 1q42-43, and CAPB on chromosome 1p36 [39]. An additional 12 PCa risk regions on different chromosomes including 1q23, 5q11, 5q35, 6p21, 8q12, 11q13, and 20p11-q11 were identified by the International

Consortium for Prostate Cancer Genetics (ICPCG) [39, 43]. Occupational exposure to herbicides and pesticides are also risk factors for PCa [39]. Obesity is associated with high-grade PCa, which is possibly from lower androgen levels [44]. These men also have a higher risk of dying with PCa [44]. Furthermore, diet may play a role in PCa risk as countries with a higher dietary fat intake also present with higher PCa mortality rates [44]. Saturated fat intake may increase the risk of PCa-related death, while vegetable fat may have a protective effect after a diagnosis of non-metastatic PCa [44]. Overall, no specific environmental or occupational exposure has been definitively shown to increase PCa risk [45]. However, a relationship between agent orange, a herbicide used by the Americans in Vietnam, with PCa risk as well as aggressive disease exist [45]. There is also epidemiological evidence suggestive of an inverse relationship between sun exposure and PCa risk [45]. While farming has been associated with increased PCa risk, the specific underlying mechanisms have not been established [45]. The PCa risk factors are described in Figure 1.2.

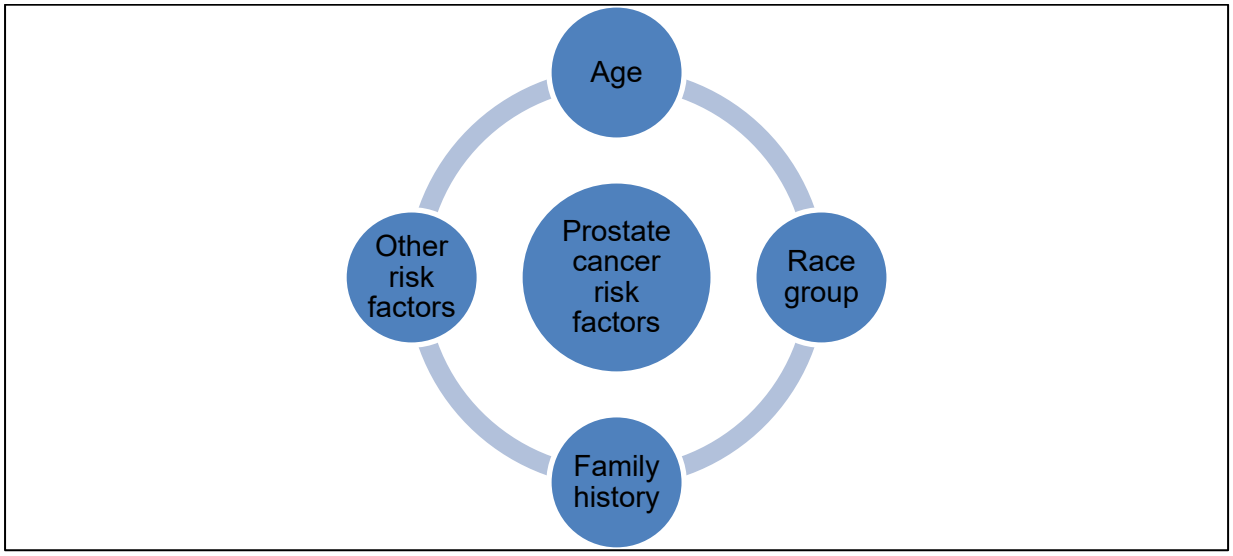


Figure 1.2: Risk factors for prostate cancer (PCa) (adapted from reference [26])

1.10 How prostate cancer biopsies are processed and reported

Within the Gauteng Province, prostate biopsies are processed in the anatomical pathology (AP) laboratories based at the Charlotte Maxeke Johannesburg Academic- (CMJAH), Chris Hani Baragwanath Academic- (CHB), Dr George Mukhari Academic- (DGM) and Steve Biko Academic (SBA) hospitals. In these laboratories, the prostate biopsies go through a number of steps to reach the endpoint of a histological diagnosis. The first step is the macroscopic description (also called gross examination) of the biopsy where the number of cores submitted is recorded, with the length of each core reported [26]. The length of the biopsy tissue significantly correlates with the PCa detection rate [26]. The next step involves cutting the

cores into blocks. Each block is placed in a cassette marked with the laboratory number. The cassettes are transferred to an embedding instrument to form wax tissue blocks. Once cooled, the wax tissue blocks are cut using a microtome and placed on glass slides. Next the slides are stained and screened by the pathologist, who will prepare the biopsy report that includes: (i) macroscopic description of the prostate cores, (ii) microscopic description and (iii) the conclusion/discussion section that summarises the histological findings. Here is an example of microscopic findings: 'Section confirm fibromuscular chips of tissue containing prostate glands. There is no significant epithelial hyperplasia. There is no prostate intraepithelial neoplasia or invasive malignancy in the sections examined'. For a PCa finding, the conclusion includes the tumour type, e.g., adenocarcinoma, and the Gleason score (GS), e.g., 3 + 4 =7 [26]. The typical AP workflow is described in Figure 1.3 below.

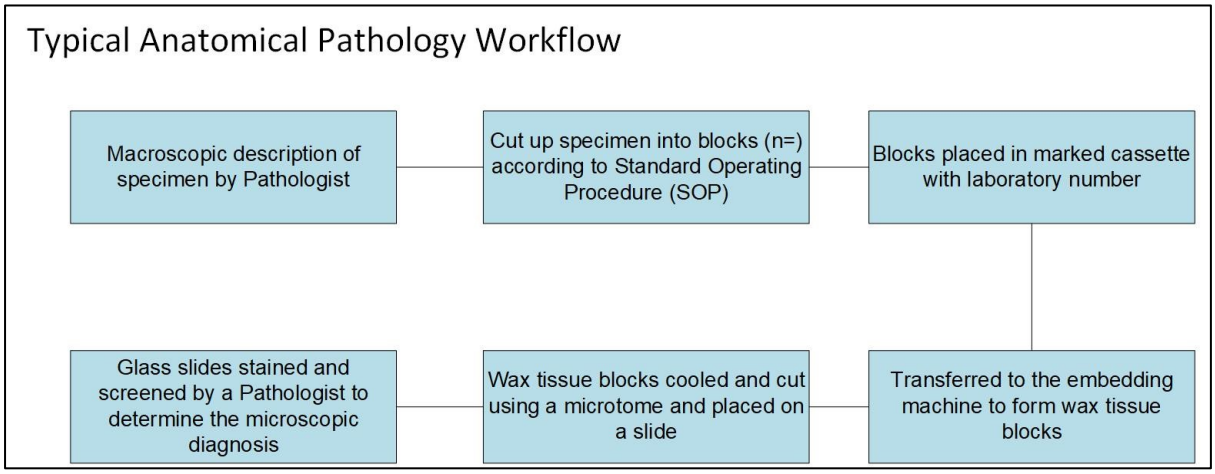


Figure 1.3: Typical anatomical pathology workflow (modified from the National Health Laboratory Service reflex testing document)

The Systematized Nomenclature of Medicine (SNOMED) clinical terms (CT) codes is a comprehensive clinical terminology system that makes it possible to both input and retrieve coded clinical information [46]. The SNOMED CT codes are an ontology supporting multiple types of relationships between clinical concepts.[46] This enables SNOMED CT to be used to efficiently classify a diagnosis [46]. For prostate biopsies, the anatomical pathologist will add the SNOMED Topography (T) and Morphology (M) codes based on the histological findings. The T code identifies anatomic terms for each organ, e.g., T-28000 for a lung biopsy. The M code describes the microscopic changes identified by the AP, e.g., M-4000 for inflammation. The SNOMED CT T and M codes for each biopsy are entered in the laboratory information system (LIS).

The prostate biopsy is used both to diagnose PCa and guide management with the AP an integral part of the decision-making process [47]. The management of patients with PCa is directly linked to the biopsy reported GS, percentage and numbers of cores as well as ancillary information such as seminal vesicle involvement and peri-neural invasion [47].

The GS is reported as five different patterns with increasing abnormalities of glandular structure [48]. To generate the GS, the most extensive (primary) pattern, plus the second most common (secondary) pattern is reported out of 5, e.g. 3/5 and 4/5 [26]. This equates to a GS of 3 + 4 =7.

The International Society of Urological Pathology (ISUP) developed grade groups (GG) in 2014 to provide a more accurate stratification of neoplasms than the GS [49, 50]. This GG simplifies the classification by reducing the number of grading categories to five [49]. This grading was accepted by the World Health Organisation (WHO) 2016 edition of 'Pathology and Genetics: Tumours of the Urinary System and Male Genital Organs' [51]. The GG's are described with the corresponding GS's in Table 1.1 below [26].

Table 1.1: Grade Group (GG) reporting indicating the corresponding Gleason score

Grade Group	Gleason Score
1	2 to 6
2	3 + 4 =7
3	4 + 3 =7
4	4 + 4 =8 or 3 + 5 =8 or 5 + 3 =8
5	9 to 10

Modified from reference [26]

The European Association of Urology (EAU) guidelines recommends using the following terminology for reporting prostate biopsies [26]:

Table 1.2: Recommended terminology to be used to report prostate biopsy histological findings

Histological findings
Benign/negative for malignancy
Active inflammation
Granulomatous inflammation
High-grade prostatic intraepithelial neoplasia (PIN)
High-grade PIN with atypical glands, suspicious for adenocarcinoma (PINATYP)
Focus of atypical glands/lesion suspicious for adenocarcinoma/atypical small acinar proliferation, suspicious for cancer
Adenocarcinoma

Modified from reference [26]

Furthermore, the following information should be reported for a PCa histological finding [26]:

- Type of carcinoma, e.g., adenocarcinoma
- Primary and secondary GS
- Extent of carcinoma (in mm or percentage of cores submitted)
- Seminal vesicle, lymphovascular or peri-neural invasion
- Gleason grade group (GG)

The prostate biopsy report contains essential information on the prognostic characteristics relevant for clinical decision making [26].

1.11 Urological guidelines for prostate cancer screening and diagnosis in South Africa

In 2013, the first local urological diagnostic and treatment PCa guidelines were introduced in South Africa [52]. These guidelines indicate that PCa diagnosis is based on taking a focused urological history and clinical examination [52]. A DRE is recommended for all patients presenting with a clinical suspicion of PCa (symptoms or an elevated PSA) [52]. The indications for a prostate biopsy include an abnormal DRE and/or a total prostate specific antigen (PSA) above the age-related reference range defined as follows: (i) 40 to 50 years: 0 - 2.5ng/ml, (ii) 50 to 60 years: 0 - 3.5ng/ml and (iii) >60 years: 0 - 4.0ng/ml [52]. At first presentation, for men with a normal DRE and a PSA below 10µg/L a repeat PSA is advised in six weeks [52]. PSA testing is recommended for males with a life expectancy of more than 10 years as follows: (i) Black African men from the age of 40 years with a positive family history (PCa and/or breast cancer in a first degree relative), (ii) from the age of 45 years for all males and (iii) patients with a history of lower urinary tract symptoms (LUTS) and/or clinical suspicion of PCa regardless of age group [52].

In addition to a DRE and PSA, additional diagnostic tests are advised [52]. The free PSA (FPSA) is advised for men with a negative first biopsy and an age-related elevated PSA <10µg/L [52]. The free to total PSA ratio (FT) is able to improve decision making for a repeat biopsy, i.e. for a FT >20% a repeat biopsy is not recommended [52]. The PSA velocity (PSAV) is also an additional diagnostic measure, reported as the absolute annual increase in serum PSA (µg/L/year) [52]. A PSAV greater than 0.75µg/L/year is also an indication for a prostate biopsy [52].

Other diagnostic testing include the PCA3 urine test [52]. Although the PCA3 score increases with PCa volume, conflicting data indicate that it may not be able to independently predict the GS [26]. The primary indication for the PCA3 test is to determine whether repeat biopsy is needed after an initial negative biopsy [26]. However, the clinical effectiveness of the PCA3 test for this purpose is uncertain [26]. Therefore, local guidelines do not recommend PCA3 to be used in place of PSA testing [52].

The tumour, node and metastasis (TNM) classification system is used to record the local growth and regional and distant spread of cancer [53, 54]. This classification system is used to stage PCa for patients with a similar outcome [26]. For a primary tumour, the following TMN clinical stages are defined in a urological setting (Table 1.3) [53].

Table 1.3: The tumour, node, metastasis (TNM) classification system used for cancers

Tx	**Primary tumour cannot be assessed**
T0	**No evidence of primary tumour**
T1	**Clinically inapparent tumour not palpable or visible by imaging**
T1a	Tumour incidental histological finding in 5% or less of tissue resected
T1b	Tumour incidental histological finding in more than 5% of tissue resected
T1c	Tumour identified by needle biopsy (e.g. because of elevated prostate-specific antigen (PSA) level)
T2	**Tumour confined within the prostate**
T2a	Tumour involves one half of one lobe or less
T2b	Tumour involves more than half of one lobe, but not both lobes
T2c	Tumour involves both lobes
T3	**Tumour extends through the prostatic capsule**
T3a	Extracapsular extension (unilateral or bilateral) including microscopic bladder neck involvement
T3b	Tumour invades seminal vesicle(s)
T4	**Tumour is fixed or invades adjacent structures other than seminal vesicles: external sphincter, rectum, levator muscles, and/or pelvic wall**

Modified from reference [26]

In addition to the TNM clinical staging, the D'Amico risk classification is used in a urological setting [55]. This is one of the most widely used approaches for assessing PCa risk, using the PSA, GS and TNM clinical stage to assign patients as either low, intermediate and high-risk [55]. Patient risk stratification is important for planning treatment options, as described in Table 1.4 below [26].

Table 1.4: Risk stratification for planning treatment options

Category	Low-Risk	Intermediate-Risk	High-Risk
TNM Clinical Stage	T1 to T2a	T2b to T2c	T3a or T3b
Gleason score	2 to 6	3 + 4 =7 or 4 +3=7	8 to10
PSA	<10µg/L	10 - 20µg/L	>20µg/L
PSA: prostate specific antigen			

Modified from reference [26]

In 2016, the EAU PCa guidelines recommended early PSA testing as follows: (i) all men 50 years and older, (ii) >50 years with a family history of PCa, (iii) >45 years for African

Americans, (iv) PSA >1µg/L at 40 years and (v) PSA >2µg/L at 60 years [26]. These guidelines were updated in 2021 and early PSA testing is recommended for well-informed men at elevated risk of having PCa as follows: (i) from 50 years of age, (ii) from 45 years of age for those with a family history of PCa, (iii) from 45 years of age for men of African descent and (iv) from 40 years for men carrying BRCA2 mutations [56].

These guidelines indicate the risk of PCa in relation to PSA value, as described in Table 1.5 below [26]. The risk of PCa is 26.9% for a PSA between 3.1 and 4.0µg/L [26].

Table 1.5: Risk of prostate cancer (PCa) in relation to the total prostate specific antigen (PSA) result and Gleason score (GS)

PSA level (µg/L)	Risk of PCa (%)	Risk of Gleason score >7 (%)
0.0 - 0.5	6.6	0.8
0.6 - 1.0	10.1	1.0
1.1 - 2.0	17.0	2.0
2.1 - 3.0	23.9	4.6
3.1 - 4.0	26.9	6.7
PSA: prostate specific antigen PCa: Prostate cancer		

Modified from reference [26]

1.12 Local public sector and urological prostate cancer screening and diagnosis guidelines in South Africa between 2006 and 2016

Local public sector PCa recommendations were first introduced in the primary health care (PHC) standard treatment guidelines (STG) and essential medicines list (EML) for South Africa in 2014 [57]. In March 1996, local public-sector STG were first published by the National Department of Health (NDoH) for use by health facilities [58]. They are used across South Africa to diagnose and treat patients presenting at public health facilities.

The STG with the associated EML, are defined as a systematically developed statement designed to assist health care workers (HCW) to make decisions for the provision of appropriate health care for specific clinical circumstances [59]. There is some evidence that indicates that when pharmaceutical supply is based on an approved EML, there is an opportunity for ineffective, unsafe, or wasteful prescribing [59]. The use of STG can benefit HCW, supply chain management (SCM) and patients by: (i) standardising guidelines for HCW, (ii) encouraging high quality care by directing HCW to use the most appropriate medicines for specific conditions, (iii) enabling HCW to concentrate on making the correct diagnosis using the provided treatment options (algorithms), (iv) utilising only formulary (essential) medicines,

(v) ensuring consistent treatment at all levels of care within the health care system and (vi) improving the availability of medicines due to consistent use and ordering [59, 60].

The 2014 public-sector PHC guidelines provided PCa recommendations for the very first time. These guidelines are used by all community health centres (CHC) and clinics across South Africa [61]. This is the fifth edition of these guidelines, with earlier editions produced in 1996, 1998, 2003 and 2008 [61, 62].

Local public-sector guidelines for hospitals were introduced in 2012 and 2015 [63, 64]. These guidelines include a urology section, with no specific PCa recommendations provided [63, 64]. The urology sections in these guidelines provide recommendation for haematuria, urinary tract infection (UTI), BPH, overactive bladder and erectile dysfunction [63, 64].

Local PCa diagnostic and treatment urological guidelines were released in 2013 by the Prostate Cancer Foundation (PCF) of South Africa [52]. Furthermore, these guidelines were reviewed by the South African Urological Association (SAUA), South African Society of Medical Oncology (SASMO) and the South African Society of Clinical and Radiation Oncologists (SASCRO) [52]. The table below summarises PCa recommendations by local public-sector and urological guidelines between 2006 and 2016.

Table 1.6: Local and other urological prostate cancer (PCa) guidelines

Year (Reference)	Guidelines Title and author	Recommendations
2008 [62]	*Public sector* Standard Treatment Guidelines and Essential Medicine List for South Africa: Primary Health Care Level **National Department of Health**	Urology section includes no guidelines for PCa
2012 [63]	*Public sector* Standard Treatment Guidelines and Essential Medicine List for South Africa: Hospital Level: Adults **National Department of Health**	No PCa specific guidelines in the Urology section. Information provided for benign prostatic hyperplasia (BPH) amongst others in the Urology sections.
2013 [52]	***Urological*** Prostate Cancer Diagnostic and Treatment Guidelines **Segone *et al***	PSA testing is recommended in males with a life expectancy of more than 10 years as follows: (i) from the age of 40 in black African patients with a positive family history (PCa/Breast cancer in a first degree relative) and (ii) from the age of 45 years in all other males. Indications for prostate biopsy include an abnormal DRE and/or PSA above the age-related norm as follows: (i) 40 - 50 ≤2.5, (ii) 50 - 69 ≤3.5 and >60 ≤4. If DRE is normal and PSA ≤10 at presentation, repeat PSA in 6 weeks advised.
2014 [57]	***Public sector*** Standard Treatment Guidelines and Essential Medicine List for South Africa: Primary Health Care Level **National Department of Health**	Usually occurs in men >50 years of age and is most often asymptomatic. Systemic symptoms, i.e. weight loss, bone pain, etc. occurs in 20% of patients. Referral to higher level of care for patients with suspected cancer
2015 [64]	***Public sector*** Standard Treatment Guidelines and Essential Medicine List for South Africa: Hospital Level: Adults **National Department of Health**	No PCa specific guidelines in the Urology section. Information provided for benign prostatic hyperplasia (BPH) amongst others in the Urology sections.

1.13 Other prostate cancer guidelines

Various other PCa guidelines have been released in other countries. The evidence-based guidelines from the EAU, American Association of Urology (AUA), National Comprehensive Cancer Network (NCCN) and European Society for Medical Oncology (ESMO) are described in Table 1.7 below [26, 65-67].

Table 1.7: Prostate cancer recommendation from other countries

Year (Reference)	Guidelines Title and author	Recommendations
2013 [66]	Early detection of Prostate Cancer **American Association of Urology**	Early detection using the PSA followed by prostate biopsy for diagnostic confirmation. In addition to PSA, DRE, PSA derivatives and isoforms such as free PSA, -2proPSA, prostate health index, hK2, PSA velocity or PSA doubling time and PCa# are recommended. Shared decision making between clinicians and men is a strategy for making health care decisions when there is more than one medically reasonable option. Test offered as follows: using a PSA threshold of 4.0 μg/L No testing for men <40 years Offer testing to men aged to 59 years at average risk. For men ages 55 to 69 years, PSA should be offered by weighing the benefits of preventing prostate cancer To reduce the harms of screening, a routine screening interval of two years is preferred over annual screening No testing for men >70+ years, based on a 10 to 15 year life expectancy.
2016 [26]	Guidelines on Prostate Cancer **European Association of Urology**	Offer early PSA testing in men at elevated risk of having PCa as follows: (i) > 50 years of age, (ii) > 45 years of age and a family history of PCa, (iii) African-Americans > 45 years of age, (iv) men with a PSA level of > 1ng/mL at 40 years of age and (v) men with a PSA level of >2 ng/mL at 60 years of age. PCa is usually suspected on the basis of DRE and/or PSA levels.
2019 [67]	Prostate Cancer **National Comprehensive Cancer Network**	Perform DRE to confirm clinical stage Perform PSA, calculate PSA density and doubling time Estimate life expectancy Obtain family history Perform prostate biopsy
2020 [65]	Prostate cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up **European Society for Medical Oncology**	Offer testing as follows: (i) men >50 years, (ii) men >45 years with a family history of prostate cancer, (iii) African-Americans >45 years and (iv) BRCA1/2 carriers >40 years. Testing in asymptomatic men should not be performed for men with a life expectancy <10 years

1.14 Controversy about prostate-specific antigen testing

There is controversy about the role of PSA testing for PCa diagnosis [26]. Some studies suggest that testing may result in over-diagnosis and over-treatment [68, 69]. Hayes *et al*

reported that the available evidence favours a discussion with the men on the pros and cons of PSA testing with average-risk, aged 55 to 69 years [69]. Furthermore, PSA testing should only be offered to men expressing a definite preference therefore [69]. Currently, our local urological guidelines in South Africa still advocate for PSA testing from 40 for Black Africans and 45 years for all other men [52]. The draft 2017 local urological guidelines recommend informed patient-based screening for males with a life expectancy of more than 10 years [47].

Since 2012, there have been several conflicting PCa guidelines [70]. The United States Preventative Services Task Force (USPSTF) guidelines recommended against screening men ≥75 years in 2008 [70]. In 2012, the updated USPSTF guidelines added recommendations to offer no PSA testing for men of all ages [70]. In contrast, the American Urological Association (AUA) guidelines introduced in 2013 recommended against screening for men <40 years [70]. The AUA guidelines did also not recommend routine screening for men aged 40 - 54, but recommended shared decision making about screening for men aged 55 to 69 years [70, 71].

The development of conflicting guidelines are a result of three large prospective randomised controlled trials (RCTs) that published data on screening between 2009 and 2010 [72-74]. Andriole *et al* compared annual PCa screening *versus* standard of care as the control at ten American study centres [72]. Following 7 to 10 years of follow-up, the PCa mortality rate was very low and did not differ significantly between the two study groups [72]. Schröder *et al* analysed data from cancer registries in seven European countries [74]. This study randomly assigned men to a group that offered PSA screening once every four years compared to a control group with very limited screening [74]. The intervention arm of this study reported that PCa mortality decreased by 20%, associated with a high risk of over-diagnosis [74]. Hugossun *et al* randomised men in a 1:1 ratio to either a screening group invited for PSA testing every two years or to a control group that were not screened [73]. This study showed that PCa mortality was reduced by almost half over 14 years in the intervention arm with the risk of over diagnosis [73].

A Cochrane review reported that PCa screening is not likely to result in a significant reduction in cancer-specific and overall mortality [75]. One of the challenges with PCa screening is over-diagnosis and over-treatment [75]. Therefore, mass screening of PCa is not indicated using a public health approach [26]. In contrast, early diagnosis is advised for well-informed men with 10 to 15 years of life expectancy, based on DRE and PSA testing [26]. Patient screening requires informed consent after a comprehensive discussion of the pros and cons of the complete procedure, considering the patient's risk factors, age and life expectancy [26].

Table 1.8: Studies providing conflicting prostate screening data

Year (Reference)	Source (Type)	PCa screening recommendation
2013 [75]	Ilic *et al* (Cochrane peer review)	All RCTs of screening versus no screening for prostate cancer were eligible for inclusion in this review (n=5). There were 341 342 participants. Patients that received a PSA with or without a DRE. Patients aged 45 to 80 years and duration of follow-up from 7 to 20 years. Men randomised to the screening and control group. There was no statistically significant reduction in PCa mortality between and screening group.
2009 [72]	Andriole *et al* (Prostate, lung, colorectal and ovarian (PLCO) cancer screening trial on prostate cancer mortality)	From 1993 to 2001, men aged 55-74 were enrolled at 10 study centres in the United States and assigned to either the annual screening or usual standard of care as the control arm (n=76 693) Cases: Offered annual PSA testing for 6 years and DRE for 4 years (n=38 343) Controls: Offered the standard of care (n=38 350) At 7 years, screening was associated with a relative increase of 22% in the rate of PCa diagnosis, with no reduction in PCa mortality.
2009 [74]	Schöder *et al* (Randomised prostate cancer screening study to assess mortality)	The European randomized study of screening for prostate cancer (ERSPC) is a randomised multi-centre trial reporting incidence and mortality for 9, 11 and 13 years of follow-up in the intervention and control arm. Men aged 50-74 were recruited from population registers. Cases: Offered PSA testing Controls: Not offered PSA testing A substantial mortality reduction due to PSA testing.
2010 [73]	Hugosson *et al* (Prostate cancer screening study – Göteborg study)	Men born between 1930 and 1944 randomly sampled (n=20 000) from the population register and randomised in a 1:1 ratio to the screening or control group. For both arms, incidence was checked by obtaining cancer registry data. The main outcome was the absolute and relative risk reduction in PCa mortality. Cases: Invited every second year for PSA testing until they reached the upper age limit (Median 69, range 67 – 71 years). Controls: Not offered PSA testing The benefits of PCa screening compare favourably to other studies with mortality reduced by half over 14 years. The risk of over-diagnosis is substantial.

1.15 Global prostate cancer studies

The 2012 GLOBOCAN data produced by the International Agency for Research on Cancer (IARC) reported global cancer incidence and mortality estimates from 184 countries [76]. This study reported that PCa was the second most common neoplasm in men with an estimated 1.1 million cases diagnosed in 2012 [76]. The global age-standardised incidence rate (ASIR) was 31.1 per 100 000 population in 2012 [76]. PCa incidence varied 25-fold globally across the 184 countries, with higher incidence reported in developed areas such as Australia/New Zealand (111.6), Northern America (97.2), Western- (94.9) and Northern Europe (85) per 100 000 population [76]. In less developed areas such as the Caribbean (79.8), Southern Africa (61.7) and South America (60.1) high incidence rates were also reported per 100 000 population [76]. Incidence remains low in Asian populations [76].

Ferlay *et al* reported the 2012 GLOBOCON cancer incidence data [38]. The estimated number of PCa cases was 758.7 and 353 thousand for more and less developed regions respectively [38]. It was postulated that the higher incidence rates in developed regions was due to widespread PSA testing and subsequent diagnosis [38]. In many developing countries, PSA testing is not routinely offered [38]. Torre *et al* reported that developed countries that implemented PSA testing in the 1980's have noted both a rapid increase in new case detection as well as an increasing incidence [19]. After a while, the incidence decreased rapidly as the pool of prevalent cases to be diagnosed reduced [19].

This data indicate that PCa was the predominant male cancer across most of SSA, North America, South America and Australia [19]. This highlights the burden of PCa in countries like South Africa as the leading male neoplasm. Bray *et al* reported the 2018 GLOBOCON incidence and mortality estimates [22]. PCa was the most commonly diagnosed male neoplasm in 105 countries including South Africa [22]. This study reported that PCa ASIR ranged from 1.0 per 100 000 population in Bhutan to 189.1 for France, Guadeloupe (Figure 1.4) [22] .

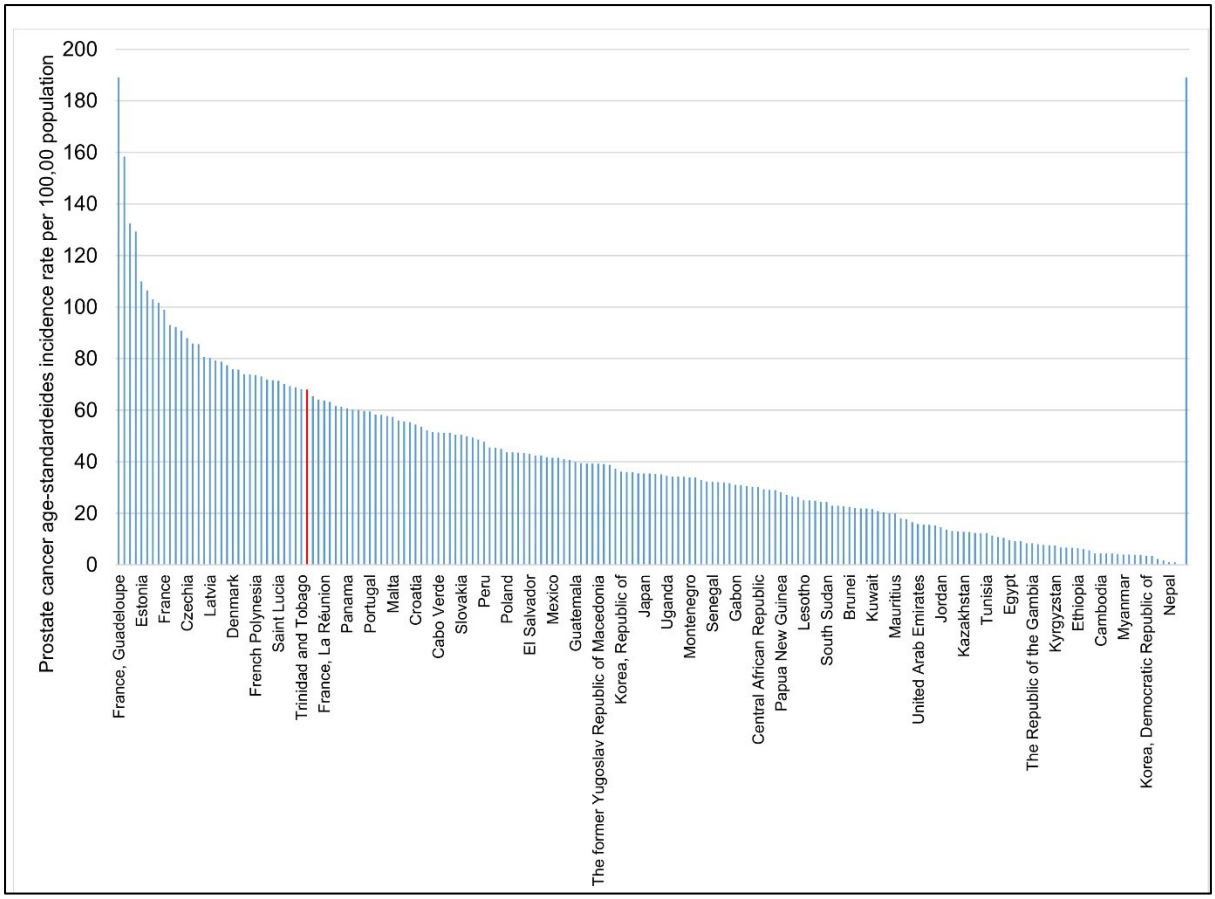


Figure 1.4: Prostate cancer age-standardised incidence rate per 100 000 population per country in 2018 for the most commonly diagnosed neoplasm reference number. Data for South Africa is reported as a red bar (modified from reference (10))

This data indicate that PCa ASIR ranged from 5.0 in South Central Asia to 86.4 per 100 000 population in Australia/New Zealand [22]. For Southern Africa, an ASIR of 64.1 per 100 000 population was reported with a relatively high mortality rate of 26.8 [22].

For South Africa, there were 3 639 new cases reported in 2012 that equates to an ASIR of 67.9 per 100 000 population (Table 1.9) [77]. For the same period, 26.4 PCa deaths were reported per 100 000 persons per year [77]. In 2018, a PCa ASIR of 68.0 was reported for South Africa with 12 452 new cases [78]. The number of deaths per 100 000 person per year increased to 27.9 [78].

Table 1.9: The GLOBOCON prostate cancer age-standardised incidence rate per 100 000 population in 2012 and 2018. The number of new cases are also reported

South African	Number of new cases	Age-standardised incidence per 100 000 population	Number of deaths per 100 000 persons per year
GLOBOCON 2012	3 639	67.9	26.4
GLOBOCON 2018	12 452	68.0	27.9
ASIR: Age standardised incidence rate			

Adapted from references [77] and [22].

1.16 African prostate cancer studies

Pilleron *et al* analysed cancer incidence data for calendar years 2008 to 2012 in four sub-Saharan populations in Kenya, South Africa, Uganda and Zimbabwe [79]. This study reported incidence data from high-quality population-based cancer registry (PBCR) stored in the IARC database [79]. The ASIR per 100 000 population was reported for two age categories, i.e., 0 - 59 and ≥60 years [79]. This study reported data for 8 944 cancer cases for adults aged 60 years and older [79]. The percentage of cases by country ranged from 23% in Uganda to 52% for South Africa [79]. Among males aged 60 years and older, PCa was the leading neoplasm in all regions except in South Africa, where oesophageal cancer exceeded all other cancer sites [79]. This study further reported that prostate and oesophageal cancers comprised between 40 and 60% of all male neoplasms for men 60 years and older [79]. The finding that oesophageal cancer was the most common male neoplasm in a South African population registry in the Eastern Cape conflicts with both National Cancer Registry (NCR) and GLOBOCON data [22, 38, 76, 77, 79-81]. This registry was specifically established to monitor oesophageal cancer in a known high-risk area, which may explain why these findings differ [79].

There are 55 countries in Africa, as defined by the African Union (AU) [82]. The 2018 GLOBOCON data provide ASIR estimates for 46/55 African countries (83.6%) [22]. The number of new PCa cases range from 22 in Djibouti to 13 078 for Nigeria [22]. New PCa cases in Nigeria, South Africa (n=12 452) and the Democratic Republic of Congo (n=5 718) represent 42.2% of all cases for this study [22]. The PCa ASIR ranged from 4.4 per 100 000 population in Benin to 68.0 for South Africa [22]. An ASIR of 55.7, 45.6, 43.7, 43.5 and 41.6 per 100 000 population was reported for Benin, Zambia, Ivory Coast, Zimbabwe and Cameroon respectively.

Adeloye *et al* conducted a systematic literature review of PCa studies conducted in Africa between 1980 and 2015 [83]. This study reported PCa incidence data for 40 studies spread across 16 countries [83]. The pooled PCa incidence increased with age, rising from 12.9 for

the 40 - 49 age category to 39.0 per 100 000 population for men 70 years and older [83]. Three studies reported PCa incidence data for South Africa ranging from 4.4 to 30.8 per 100 000 population, all from the Eastern Cape Province [83].

This study reported the highest PCa incidence per 100 000 population in Nigeria (182.5), followed by Cameroon (93.8), Zambia (37.7), Uganda (35.5), Ivory Coast (31.4) and South Africa (30.8) [83]. The lowest incidence per 100 000 population was reported for Egypt (0.41), Rwanda (1.02) and Gambia (3.46) [83].

The table below reports the GLOBOCON 2018 data for African countries as well as the highest ASIR as reported by Adeloye *et al* [22, 83].

Table 1.10: Number of prostate cancer cases and the age-standardised incidence rate per 100 000 population is reported for 46 African countries in the 2018 GLOBOCON study. Data for is study is sorted by age-standardised incidence rate in descending order. The matching highest the age-standardised incidence rate per 100 000 population reported by the Adeloye *et al* is reported in the last column.

Country	Number of cases[&]	% of Cases[&]	Age-standardised incidence rate[&]	Highest Age-standardised incidence rate[$]
South Africa	12 452	16.8%	68.0	30.8
Benin	1 315	1.8%	55.7	
Zambia	1 230	1.7%	45.6	37.7
Côte d'Ivoire	2 485	3.4%	43.7	31.4
Zimbabwe	1 299	1.8%	43.5	29.2
Cameroon	2 213	3.0%	41.6	93.8
Angola	2 016	2.7%	41.0	
Congo, Republic of	505	0.7%	40.7	
Liberia	376	0.5%	39.1	
Namibia	206	0.3%	37.3	
Equatorial Guinea	97	0.1%	35.9	
Burundi	754	1.0%	35.5	
Guinea	906	1.2%	35.3	8.1
Congo, Democratic Republic of	5 718	7.7%	35.1	
Uganda	2 086	2.8%	34.5	35.5
Eswatini	82	0.1%	34.2	21.5
Nigeria	13 078	17.7%	32.8	182.5
Ghana	2132	2.9%	32.3	
Senegal	959	1.3%	32.2	
Gabon	192	0.3%	31.0	
Kenya	2 864	3.9%	30.9	
Central African Republic	330	0.4%	30.2	
Rwanda	707	1.0%	29.1	1.02
Sierra Leone	390	0.5%	29.0	
Mozambique	1 651	2.2%	27.1	
Burkina Faso	689	0.9%	26.5	
Lesotho	138	0.2%	25.0	
South Sudan	716	1.0%	24.3	
Morocco	3 990	5.4%	22.7	
Chad	580	0.8%	22.0	
Mauritania	202	0.3%	21.9	

Country	Number of cases[&]	% of Cases[&]	Age-standardised incidence rate[&]	Highest Age-standardised incidence rate[$]
Togo	311	0.4%	20.4	
Mali	539	0.7%	17.7	4.7
Libya	317	0.4%	15.6	11.4
Malawi	525	0.7%	15.3	5.5
Botswana	72	0.1%	13.7	
Somalia	382	0.5%	13.1	
Algeria	2 578	3.5%	13.0	
Tunisia	819	1.1%	12.3	11.9
Egypt	3 109	4.2%	9.5	0.41
Sudan	938	1.3%	9.2	
The Republic of the Gambia	33	0.0%	8.4	
Eritrea	94	0.1%	7.7	
Djibouti	22	0.0%	7.6	
Ethiopia	1 701	2.3%	6.5	
Niger	172	0.2%	4.4	
[&]Globocon 2018 [$]Adeloye *et al*				

Modified from reference [22, 83]

Adeloye *et al* also reported the pooled ASIR estimates by age category (Table 1.11) [83]. The overall ASIR in Africa was 22.0 per 100 000 population [83]. ASIR increased with age from 12.9 for the 40 - 49 age category to 39.0 for men 70 years and older [83]. For the 60 - 69 and 50 - 59 age categories, an ASIR of 25.0 and 16.3 per 100 000 population was reported respectively.

Table 1.11: Pooled age-standardised incidence rate by age categories for African countries

Age Category	PCa ASIR per 100 000 population
40 - 49	12.9
50 - 59	16.3
60 - 69	25.0
70+	39.0
Total	**22.0**
ASIR: Age-standardised incidence rate	

Adapted from reference [83]

The findings from African studies are summarised in the table below:

Table 1.12: Findings from African studies

Author (Year) (Reference)	Study design	Population	Results
Adeloye *et al* (2016) [83]	Systematic literature review from 1980 to 2015 for studies that estimated incidence rate of PCa in any African location. This study included population, hospital, and/or registry based PCa studies. PCa identifier by three mechanisms included: (i) histologically confirmed, (ii) reporting diagnosis confirmed by medical practitioner and (iii) reported ICD-0-3 codes. Data extracted included the study location, study period, mean age/age range, population size, and number of incident cases.	Data were then sorted into five main African regions (central, east, north, south, and west) and countries represented within each region.	The overall pooled PCa incidence for Africa was 21.95 per 100 000 population (CI: 19.93–23.97). Pooled incidence increased as follows: 1980-1989: 15.7 per 100 000 1990-1999: 21.9 per 100 000 2000-2009: 26.1 per 100 000
Pilleron *et al* (2018) [79]	Cancer incidence data for the years 2008 – 2012 were obtained from high-quality population-based cancer registry (PBCR) data stored in the International Agency for Research on Cancer (IARC) database. Incidence data were available for 61 cancer sites as well for all sites combined by gender and 16 age groups. The ASIR per 100 000 population was reported for two age categories, i.e. 0 - 59 and ≥60 years	Five sub-Saharan population in Kenya, South Africa, Uganda and Zimbabwe	N=8 944 cancer cases (male and female) 52% of cases from South Africa (Eastern Cape population-based registry) Among males ≥60 years, PCa was the leading male cancer (except in South Africa - oesophageal cancer) The largest increase in the estimated average annual percentage change (EAPC) between 2008 and 2012 for males was noted from PCa.
Bray *et al* (2018) [22]	The 2018 GLOBOCON data provides prostate cancer age-standardised incidence rate estimates for 185 countries and 36 cancers by age and sex. This data is based on the best available sources of cancer incidence and mortality data in each country.	185 countries across the world	PCa ASIR range: Top Six countries with the highest ASIR: South Africa: 68.0 Benin: 55.7 Zambia: 45.6 Côte d'Ivoire: 437 Zimbabwe: 43.5 Cameroon: 41.6

1.17 Local prostate cancer studies

Several local studies have offered insights into local PCa disease burden. The findings from these studies are described below.

A study by Heyns *et al* conducted between January 1995 and December 2005, analysed prostate biopsy data for men presenting with histologically confirmed PCa at a urology oncology clinic at an urban tertiary hospital in the Western Cape Province (n=901) [1]. Most patients in this study were Coloured (59.8%), with data reported for only 7.9% of Black Africans [1]. The mean age at presentation was higher in the White than in the Coloured and Black African men at 69.7, 67.9 and 68.9 years respectively [1]. The mean and median PSA at presentation were significantly higher for Black Africans (766.1µg/L) and Coloureds (673.3µg/L) than Whites (196.1µg/L) [1]. This study reported that a TNM grade T1 PCa (well differentiated) was most commonly reported for Whites and Coloureds [1]. In contrast, T2 and T3 were more common for Black Africans indicating aggressive PCa [1]. This study reported that Black Africans presented with a higher grade and stage of PCa and received potentially curative treatment less often [1].

Le Roux *et al* conducted a study at a regional hospital in the KwaZulu-Natal Province [84]. This study reviewed National Health Laboratory Service (NHLS) prostate biopsy laboratory reports between November 2012 and April 2014 for only Black African patients with histologically confirmed PCa (n=81) [84]. Mean age at presentation was 72 ± 9 years with a median PSA of 154µg/L (IQR 39 - 448) [84]. Forty three percent of patients presented with a GS ≥8 (high-risk) [84]. Organ confined disease was reported for 27% of patients, with the majority of men presenting with radiographic evidence of metastatic PCa or a PSA level above 10ng/ml [84]. Delayed diagnosis was proposed as a possible reason for late presentation with advanced disease [84].

Similarly, the Southern African Prostate Cancer Study (SAPCS) enrolled only Black African cases (with PCa) and controls (patients with BPH) (n=909) [3]. This study commenced in 2008 at urban urology clinics at hospitals in the Limpopo and Gauteng Provinces [3]. Most of the patients were enrolled in the Limpopo- (n=643) compared to the Gauteng Province (n=266), with missing data for eight patients [3]. The majority of patients were PCa cases (n=576: 62.8%) *versus* controls (n=341: 37.2%). A urological complaint was cited by 86% of cases and 87.8% of controls [3]. Mean age at presentation was 71.0 years for cases and 70.6 years for controls [3]. PSA data was only available for 487/576 (84.5%) cases and 302/341 (88.6%) controls [3]. A PSA <10µg/L was reported for 6.4% of cases compared to 53.6% of controls [3]. For cases, 83.0% (404/487) presented with a PSA ≥20.0µg/L that was significantly

different to controls [3]. A GS >7 was reported for 39.5% and 24.7% of cases in the Limpopo and Gauteng Provinces respectively [3]. This study reported significant differences in clinical presentation between patients in rural Limpopo *versus* urban Gauteng [3].

Singh *et al* conducted a retrospective folder review for patients with a clinical suspicion of PCa that received a trans-rectal prostate biopsy between January and June 2013 at an urban regional hospital in the KwaZulu-Natal Province (n=106) [4]. These patients were referred from surrounding health facilities [4]. There were 74 Black African patients (69.8%), compared to 14, 13 and 5 Indian/Asians, Coloureds and Whites recruited for this study respectively [4]. An adenocarcinoma histological finding was reported for 49.1% of patients (n=52) [4]. Using the D'Amico risk stratification, 51/105 Black African patients were classified as high risk (68.9%), compared to 13.2%, 12.3% and 4.7% for Indian/Asians, Coloureds and Whites respectively [4]. The median PSA level for patients with PCa was 42.2, 14.2, 12.8 and 10.3µg/L for Black Africans, Coloureds, Whites and Indian/Asians respectively [4].

A local study by Dewar *et al* conducted a retrospective analysis of all patients that underwent a prostate biopsy at an urban academic hospital in the Western Cape Province between July 2008 and July 2014 (n=1 027) [5]. Eleven patients were excluded as PSA, race group, or histology results were missing [5]. There were 162/1016 Black African patients for this study (25.9%) [4]. The mean age of this cohort was 65.8 years [4]. The median PSA was 16.4µg/L for Black Africans compared to 10.9µg/L for the other race groups [4]. The likelihood of a positive biopsy among asymptomatic men was similar between the Black African and other race groups (41.4% *vs* 42.8%, respectively) [4]. Black Africans were statistically more likely to have a GS ≥8 (51.1%) compared to the other race groups (36.5%) [4].

Sheriff *et al* conducted a cross sectional analysis of patients with PCa who received treatment at an urban academic hospital in the Free State Province between 2008 and 2010 (n=497) [6]. This study employed a coded data sheet to collect the patient age, race, geographic location, Gleason score, PSA level, TNM staging, risk group and initial treatment administered [6]. The majority of patients were Black Africans (72.8%) compared to 21.3%, 5.4% and 0.4% for Whites, Coloureds and Indian/Asians respectively [6]. Most patients were between 65 and 75 years, with only five patients aged <50 years [6]. Most patients in this study reported a PSA >20µg/L (67.9%) [6]. Using the D'Amico classification system, 82.5% of patients were allocated to the high-risk group [6]. A GS ≥8 was reported for 43.7% of patients [6].

A local study by Maphayi *et al* extracted laboratory data for the 2013 calendar year for all PSA tests performed at a large academic laboratory in the Gauteng Province [85]. This laboratory serves surrounding hospitals and PHC facilities [85]. Sixty eight percent of PSA testing was requested by PHC facilities compared to 32% for hospitals [85]. For this study, 68% of samples were for Black Africans compared to 21% for other race groups [85]. The median PSA was 0.97µg/L (IQR: 0.54 - 2.39), increasing from 0.66µg/L for men ≤40 years to 1.68µg/L for the ≥60 years age category [85]. A PSA ≥4µg/L was reported for 17% of men [85]. The proportion of Black African men with a PSA ≥4µg/L was significantly different to other race groups [85]. A prostate biopsy was requested for 8% of Black African men with a PSA ≥4µg/L [85].

Table 1.13 below compares the proportion of Black African men recruited in each study with the provincial proportion reported in the Census 2011 data indicated [86]. These findings indicate that Black Africans were under-represented specifically in the Western Cape studies [86]. In comparison, other local studies that included men from all race groups ranged from 69.8% to 79.0% [4, 87]. Only two studies exclusively recruited Black African men [2, 3].

Table 1.13: Percentage of Black Africans for each study compared to the Census 2011 data for each respective province

Study (Published date)	Reference	Province	% Black Africans	Census % Black Africans*
Heyns *et al* (2011)	[1]	Western Cape	7.9%	50.6%
Le Roux *et al* (2015)	[2]	KwaZulu-Natal	**100%**	**47.4%**
Tindall *et al* (2014)	[3]	Limpopo	**100%**	**46.5%**
		Gauteng	**100%**	**50.8%**
Singh *et al* (2015)	[4]	KwaZulu-Natal	69.8%	47.4%
Dewar *et al* (2018)	[5]	Western Cape	25.9%	50.6%
Sheriff *et al* (2015)	[6]	Free State	72.8%	48.3%
Maphayi *et al* (2020)	[85]	Gauteng	79.0%	50.8%
*Data reported for the respective Provinces from Census 2011 data for males [88]. Studies in bold recruited only Black African men.				

The findings for the local PCa studies are presented in the table below.

Table 1.14: Findings from local prostate cancer (PCa) studies

Author (Year) (Reference)	Study design	Population	Results
Heyns *et al* (2011) [1]	Patients diagnosed with PCa in a urology setting (01/1995 to 12/2005). Western Cape Province.	n=901 7.9% Black African Mean age: 69.7 (Whites)	Median PSA: 33.5µg/L Grade 2 (3 + 4 =7) was the most common GS in Black Africans (39%) Median follow-up was 32 months.
Le Roux *et al* (2015) [2]	Patients diagnosed with PCa in a urology setting (11/2012 to 04/2014). KwaZulu-Natal Province.	n=81 100% Black African Mean age: 72	Median PSA: 154 43% of patients reported a GS ≥8
Tindall *et al* (2014) [3]	Case control study in a hospital setting (2009-2012) Limpopo (70%) and Gauteng (30%) Provinces	n=797 Controls: n=576 (patients with BPH) Cases: n=341 Mean age: Controls: 70.6 Cases: 71 100% Black Africa	PSA ≥98: (Median not reported) Controls: 3.3% Cases: 50.1% GS >7: Cases: 35.5%
Singh *et al* (2015) [4]	Retrospective folder review for patients that received a biopsy (PCa suspect) as a hospital (01-2016 to 06-2013). KwaZulu-Natal Province	n=106 Mean age 67.6 69.8% Black African	Median PSA (Men with PCa): Black African: 42.2µg/L Whites: 12.8µg/L Indian/Asians: 10.3µg/L Coloureds: 14.2µg/L 49.1% of biopsies with Adenocarcinoma Median referral time of 55 days
Dewar *et al* (2018) [5]	Retrospective analysis of patients that received a biopsy, referred to urology. (07-2008 to 07-2014) Western Cape Province	n=1016 Mean age 65.8 25.9% Black African	Median PSA: 16.4 (Black African) 10.9 (Other race groups) Adenocarcinoma: 57.4% (Black African) 44.1% (Other race groups) High risk PCa (GS ≥8)

Author (Year) (Reference)	Study design	Population	Results
			25.9% (Black African) 13.7% (Other race groups)
Sheriff *et al* (2015) [6]	Retrospective analysis of patients diagnosed with PCa who received treatment (2008-2010) Free State Province	n=497 72.8% Black African	High-risk group defined as GS 4 + 7 =7 and 8 - 10 For this group: 82.5% identified as high-risk 67.9% of patients reported a PSA ≥20µg/L. 4 + 3 =7 GS reported for 15% of patients GS ≥8 reported for 43.7% of patients
Maphayi *et al* (2020) [85]	Gauteng Province laboratory data for the Charlotte Maxeke Johannesburg Academic hospital laboratory (Gauteng Province) (2013).	n= 21 032 samples n= 18 022 patients 79% Black African	Median PSA: 0.97 (IQR 0.54 – 2.39 µg/L) Age Range: ≤60: 0.66µg/L (IQR 0.44 - 1.00µg/L) >60: 1.68µg/L (IQR 0.72 - 5.38µg/L) PSA Results: 1.7% ≥4.00µg/L 89% of patients received one PSA test. Black Africans >60 years reported an elevated PSA using age specific reference range (p <0.001) Only 8% of Black African men with a PSA ≥4.00µg/L received a biopsy (n=423)

1.18 National prostate cancer registry data

The NCR reports PCa ASIR for histologically confirmed incident cases [89]. The NCR is a passive pathology-based surveillance system, with copies of pathology reports confirming a cancer diagnosis submitted voluntarily by both public and private laboratories throughout South Africa [89, 90]. From the narrative pathology reports, data items including patient demographics, topography, morphology and date of diagnosis are abstracted [89]. The NCR collates laboratory confirmed cancer cases to report national annual cancer ASIR stratified by gender, age and race group [89]. A dramatic increase in national PCa ASIR was reported by Babb *et al* from 27.1 in 1996 to 30.8 per 100 000 population by 2006 [91]. Between 2007 and 2016, the national ASIR increased from 28.96 to 47.48 per 100 000 population.

The figure below, depicts the increase in national PCa ASIR reported by the NCR [7, 8, 80, 91-97]. A steep increase is noted between 2010 and 2011, rising from 29.90 to 44.30 per 100

000 population [80]. Between 2012 and 2016, the prostate cancer ASIR reached a plateau ranging from 46.53 to 47.48 per 100 000 population respectively [7, 8].

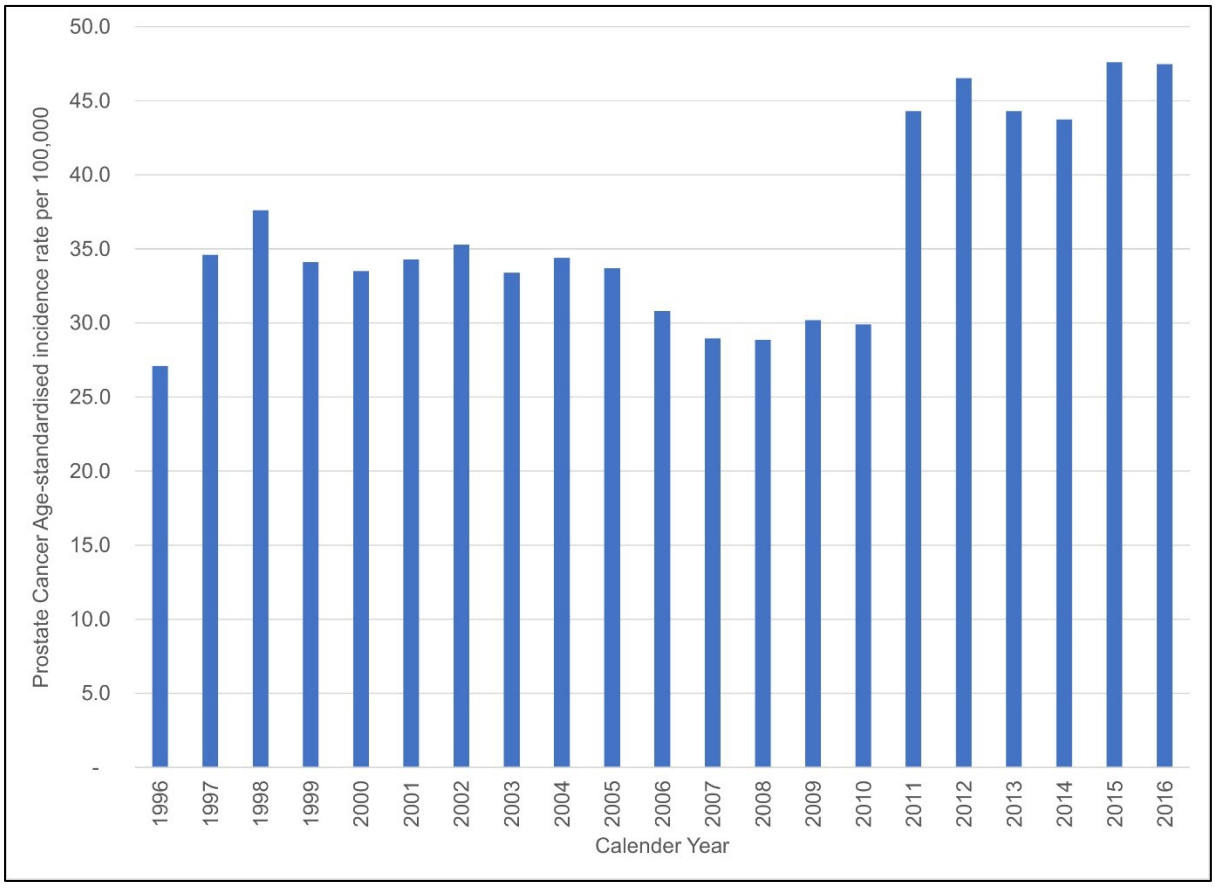


Figure 1.5: Bar chart reporting National Cancer Registry national age-standardised incidence rate (ASIR) estimates for South Africa between 1996 and 2014

Adapted from references [7, 8, 80, 91-97]

1.19 Challenges with the manual extraction of prostate biopsy data and the way forward

The PSA quantitative data were easily extracted as it was well-constructed, captured in a standardised format and easily analysed [98]. Structured data conform to a tabular format with a relationship between the different rows and columns in a relational database, where each variable holds important information [99]. As a result, these data are stored in a structured format in a database and require minimal expertise to start the analysis.

In contrast, unstructured and semi-structured data is information that either does not have a predefined data model or is not organised in a pre-defined manner [99]. Unstructured information is typically text-heavy and contains important information such as dates, numbers and facts [99]. Semi-structured data do not conform with the formal structure of data models

associated with relational databases, but contain markers to separate semantic elements within the data [99].

The text-intensive prostate biopsy narrative report is classified as semi-structured data, provided as a comma separated values (CSV) extract. However, while the narrative report are captured as free text with heading that define separate sections of the report, we still require trained coders who read the entire narrative report and enter key clinical information using a data system. This process is time consuming, tedious, subject to transcription errors and delays PCa surveillance activities. For example, the NCR uses a team of coders that abstract data items from the prostate biopsy narrative reports [100]. This manual coding takes a lot of time to code each biopsy report, as a result we only received the NCR data for 2016 in 2020 [7].

Due to these challenges, most local studies have small sample sizes ranging from 81 to 1016 to describe PCa [1-3, 5, 6]. This also explains the paucity of local data. There is a need for studies with much larger sample sizes that can describe PCa presentation across an entire province. We, therefore, need to develop automated mechanisms for extracting data from the semi-structured prostate biopsy reports.

This could be achieved through the application of big data, which offers the ability to handle large volumes of structured and unstructured data from multiple data sources, with analytical tools having the ability to study outcomes for a population [101]. Big data is used to handle datasets that are too big, too fast, and too complex to process and interpret with existing tools [101]. Big data is defined by a set of criteria, defined as the six v's: (i) value, (ii) volume, (iii) velocity, (iv) variety, (v) veracity and (vi) variability [101]. The prostate biopsy narrative report meet all these criteria [101, 102].

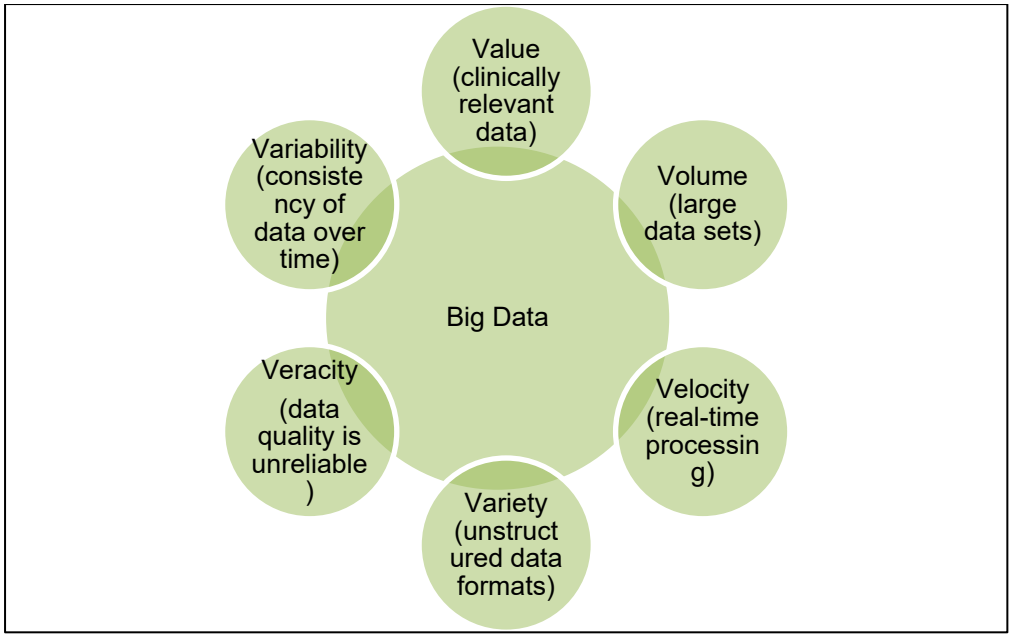


Figure 1.6: The six criteria for big data

Adapted from reference [101]

The table below describes each of the v's as well as how the prostate biopsy data meet these criteria:

Table 1.15: Table describing how the semi-structured narrative prostate biopsy data meet the six criteria for big data

Criteria	How prostate biopsy meet the criteria
Volume	Volume refers to enormous quantities of data [102]. This refers to very large datasets. There are large numbers of semi-structured narrative reports generated routinely across the NHLS. The National Cancer Registry reported data for 8332 cases of PCa [8]. The CDW has PCa narrative reports from 1996 that equates to a very large dataset. In 2016, data was reported for 39060 male and 42288 female cancer cases [8]. When combined with all the other cancers, the biopsy narrative reports stored is the CDW since 1996 meet the criteria for volume.
Velocity	Velocity refers to data created in real-time [102]. Across the NHLS, all laboratories are generating data in real-time using a single LIS. The prostate biopsy reports are routinely generated by anatomical pathologists across South Africa, with data captured in real-time electronically in the LIS, meeting the velocity criteria.
Variety	Variety relates to structured, semi-structured and unstructured data [102]. Within the LIS, laboratory results are captured in all three formats. The prostate biopsy narrative reports are captured with some structure and therefore referred to as semi-structured. However, although there are general headings used, each anatomical pathology laboratory may use different formats and wording in the narrative biopsy report. As a result, the pathologist has some freedom in the structure and content of the prostate biopsy report, meeting the variety criteria.
Veracity	Veracity relates to messy and noisy data [102]. This is definitely the case for prostate biopsies. The quality, structure, content and formatting of the narrative report is determined by each anatomical pathologist. Unfortunately, the semi-structured narrative biopsy report varies in terms of content.
Variability	Variability indicates that the meaning of data can change [102]. Over the last decade, there have been various attempts to standardise what is reported in the prostate biopsy report. In addition, laboratories switched to a new laboratory information system. As a result, the data reported by each LIS

Criteria	How prostate biopsy meet the criteria
	differs in terms of how data is captured and stored. In addition, we now have to report both the GS and GG, which meets the variability criteria.
Value	Value infers that multiple insights can be extracted from data [102]. The prostate biopsy holds immense clinical value to describe PCa presentation, diagnosis and follow-up. With a unique identifier it would be possible to undertake longitudinal analysis. A good example of value is the reporting of ASIR for a given population.

Adapted from references [101, 102].

In addition to big data, we require the computational capacity to extract clinical information from unstructured data without manual coding. This includes the use of artificial intelligence (AI) techniques that aim to mimic human cognitive function [32]. Examples of AI approaches include text mining, predictive analytics, pattern recognition, expert systems, fuzzy logic and machine learning (ML) [103]. Many AI models have statistical underpinnings [103]. The figure below describes how unstructured data is acquired using big data and then analysed using AI [103].

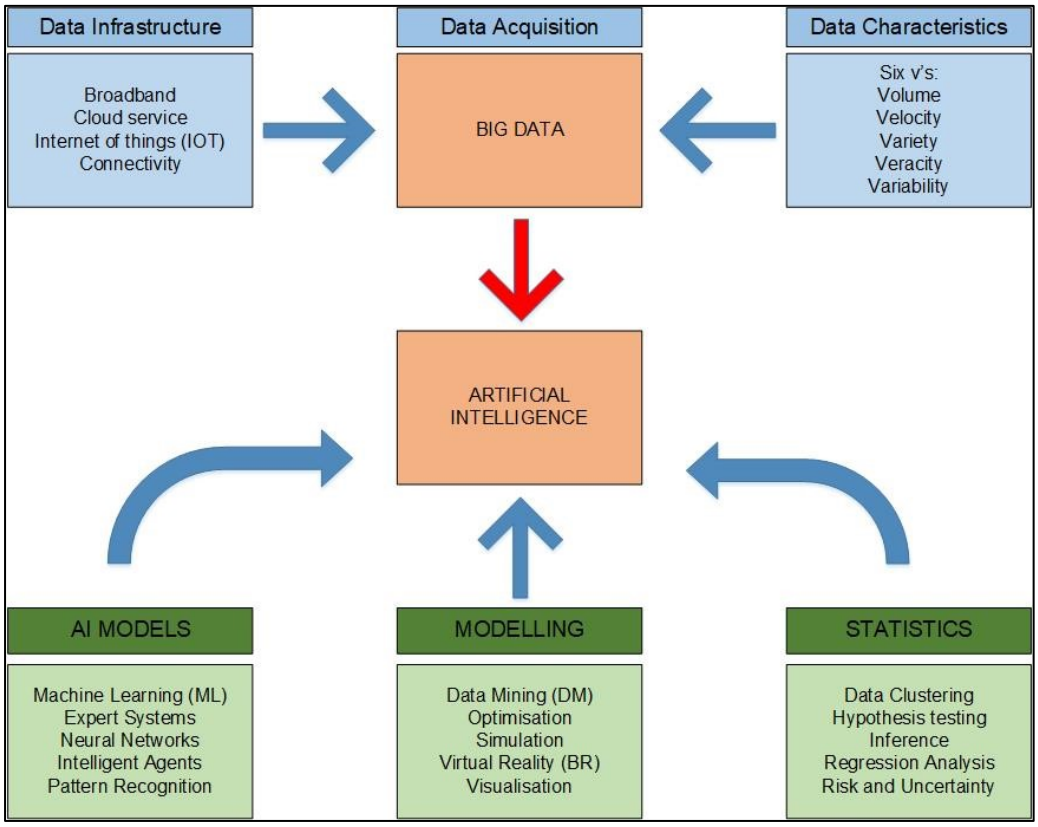


Figure 1.7: Illustrating the use of big data and the application of artificial intelligence (taken from reference (77))

For unstructured data, text mining and natural language processing (NLP) can be used to extract clinical information. Specifically for cancer, unstructured biomedical text has great value for determining the diagnosis and planning treatment [104]. The goal of text mining is to derive implicit knowledge from unstructured text and present it in an explicit form [104]. Text mining generally has four phases with multiple approaches applied in each, i.e. information retrieval, information extraction, knowledge discovery, and hypothesis generation [104]. Information retrieval entails converting the data file to a document format to start the extraction of textual information [104]. Entity recognition is the most important step in the extraction of knowledge with an overall aim of identifying specific terms of interest, e.g. gene, protein, disease, and drug [104, 105]. The final step of knowledge discovery relates to the extraction of the required information from large volumes of structured or unstructured data [104].

Text mining employs computational technologies such as natural language processing (NLP) and pattern recognition [104], NLP is a computational technique that can be used to understand and manipulate natural language text or speech [106, 107]. There are various NLP tasks such as coreference resolution, discourse analysis, named entity recognition (NER), sentiment analysis and word sense disambiguation (WSD) [108]. Many of the NLP approaches are used for the pre-processing of document text [108].

CHAPTER 2 - USING SYSTEMATIZED NOMENCLATURE OF MEDICINE (SNOMED) CLINICAL TERMS (CT) CODES TO ASSIGN HISTOLOGICAL FINDINGS FOR PROSTATE BIOPSIES IN THE GAUTENG PROVINCE, SOUTH AFRICA: LESSONS LEARNT

For uniformity of layout of this thesis, this chapter is laid out in the style of the thesis but is identical to the published paper in the African Journal of Laboratory Medicine. The paper is included at the end of this thesis as Appendix A.

Abstract

Introduction:

PCa is a leading male neoplasm in South Africa. The aim of our study was to describe PCa using SNOMED CT 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 CT morphology M and T codes where the term 'prostate' was captured in the narrative report. Microsoft Excel was used to prepare the code descriptions. Using M code descriptions, the diagnosis, sub-diagnosis, sub-result and ICD-0-3 codes were assigned using a lookup table. The T code descriptions identified biopsies of prostatic origin. Microsoft Access was used to combine the data extracts and lookup tables. Contingency tables reported M and T codes, diagnosis and sub-diagnosis frequencies. We manually coded 1000 random biopsies to assess the positive predictive value (PPV) of the SNOMED CT codes.

Results:

An M and T code was reported for 88% (n=22 009) of biopsies. Of these, 20 551 (93%) were of prostatic origin. A benign diagnosis (ICD-0-3:8000/0) was reported for 10 441 biopsies (51%) and 45% had a malignant diagnosis (n=9 302). An adenocarcinoma (8140/3) sub-diagnosis was reported for 88% of biopsies (n=8 201). An Atypia diagnosis was reported for 760 biopsies (4%). Inflammation (49%) and hyperplasia (29%) were the predominant benign sub-diagnoses. The manual coding revealed a PPV of 0.96 for the pathologist assigned SNOMED CT codes.

Conclusion:

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

2.1 List of definitions for terms used in this chapter

Term	Definition
Atypia	The presence of atypical cells within tissue.
Biopsy	The removal of a part of tissue for examination in a laboratory.
Chained	To link together a set of data values in a single row or variable.
Comma separated	Mechanism used to store multiple values in a single row or variable. The multiple values are separated by a comma, e.g., M-72450, M-00100.
Diagnosis	The histological diagnosis for a prostate biopsy, i.e., malignant or benign.
Dysplasia	The presence of abnormal cells within a tissue that are not indicative of cancer.
Hyperplasia	The enlargement of the prostate organ or tissue caused by an increase in the reproduction rate of its cells.
Inflammation	The presence of neutrophils, lymphocytes and plasma cells in tissue observed microscopically.
Interoperability	The ability of computer systems to exchange data in a coordinated manner without the need for manual coding
Join	A mechanism to join data from multiple tables in a relational database. There are multiple join types that determines the data populated, i.e. inner (common value) and outer (one master table and only matching values from the child tables). Outer joins are directional, i.e. left and right.
Lookup table	A lookup table is used to display information from one database table based on the value stored in a separate table that describes the original variable value.
Manipulation	The process of changing or amending data such that it would be easier to read, interpret and analyse.
Morphology (M)	Describes changes found in cells, tissues and organs.
Oncology	A branch of health sciences that is responsible for the prevention, diagnosis, treatment and study of cancers.
Query	A database query refers to the action of retrieving data from one or more tables. Queries are developed using the structure query language (SQL). Queries have the power to report data from multiple tables, including lookup tables, in a single query to facilitate the analysis of data.
Relational database	Database model that stores data in tables. Tables can be joined using a primary key. The primary key is a unique identifier for each row within the table that can be used to combine data from multiple tables in a query.
Sub-Diagnosis	The histological details for a diagnosis, e.g., malignant and adenocarcinoma (cell type).
Sub-Result	For an inflammation sub-diagnosis, indicate the specific type of inflammation, i.e., acute, chronic and granulomatous.
Table	A systematic representation of data in rows and columns in a database or Microsoft Excel worksheet.
Topography (T)	Describes the specific location in the body for the biopsy.

2.2 Introduction

PCa is a leading male cancer in South Africa in 2012, while globally it is the second most frequently diagnosed neoplasm [38]. A 2012 global cancer study reported an estimated ASIR of 67.9 per 100 000 for South Africa with an associated mortality rate of 26.4 per 100 000 [38, 76]. Presentation with highly aggressive PCa in African men in South Africa has been described [1, 2].

A national cancer control programme (NCCP) was recommended by the WHO with the aim of identifying priorities and assigning resources to ensure sustained progress to meet the stated objectives to reduce cancer incidence and mortality as well as to improve the quality of life for patients with cancer [109]. This can 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 health care resources [110]. The 2013 - 2020 Global Action Plan for the Prevention and Control of NCD aims to achieve a 25% reduction in the relative risk of premature mortality from cancers and other NCD [111]. This initiative emphasizes the importance of both surveillance and disease registries that should be integrated into existing health information systems to improve the availability of high-quality data [111]. 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 to generate information to assist planning of cancer control programs [112]. Cancer registries should typically publish annual data within 28 months after the close of the year in which incident case was diagnosed [113]. A delay in cancer registry reporting is a major limitation for understanding PCa trends [113].

In South Africa, the NCR is a passive reported registry that follows the internationally recommended methodology to manually code pathology reports using the ICD-O-3 coding system [100, 114, 115]. The ICD-O-3 describes both the anatomical site and cell type and behaviour (malignant or benign biopsy) [116]. The NCR had only reported data for 2016 in 2020 [117]. Cases are manually coded resulting in a reporting delay. Based on the Surveillance, Epidemiology, and End Results (SEER) program standard, the 2015 report should have been published by 2018 [113]. The SEER Program itself, on which the South African reporting standard is based, comprises 11 registries in 5 states and 6 metropolitan areas in the United States of America (USA) which generates annual cancer data approximately 28 months after diagnosis [112].

Both internationally and nationally, surveillance is defined as an ongoing, timely, and systematic collection and analysis of cancer data to assess risk factors, screening, diagnosis,

incident cancer cases and death [113]. The aim of surveillance is to analyse and disseminate cancer data to identify challenges and opportunities in the delivery of cancer control programmes in a timely fashion [113].

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-0-3 codes to the reporting system [89]. This can take hundreds of hours of manual database building to report PCa data. The coders have to add both topography and morphology ICD-0-3 codes for each narrative biopsy report reviewed [89]. This manually intensive approach delays the time from diagnosis to identifying PCa cases for surveillance. PCa reporting is required at least 28 months after diagnosis (~2 years) to understand changes in incidence [112]. Without timely data this is not 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 [118-120]. For our study, we decided to use the SNOMED CT codes [46]. SNOMED CT is a comprehensive and precise health terminology used globally that incorporates a structured list of health terms or concepts [46]. One of the benefits of using SNOMED is that codes can be mapped to other coding systems to facilitate interoperability [46]. All references to SNOMED relate to SNOMED CT [121]. In an AP setting, SNOMED CT is used to capture the histological finding in the form of M and T codes. The M and T codes are captured directly in the LIS by the pathologist after examining prostate cores. The same histological findings are also reported as a narrative pathology report. 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 AP 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, a biopsy of prostatic origin with an adenocarcinoma finding, the following codes are reported (description in brackets): (i) T-28000 or T-92000 (Prostatic structure), (ii) M-80003 (Malignant neoplasm, primary) and (ii) M-81403 (Adenocarcinoma, no subtype) [122].

Studies outside South Africa have used SNOMED codes to transform laboratory and other reported data for cancer registry reporting [123]. A good example is the Danish PCa registry that analysed SNOMED data for biopsies with PCa, i.e. histologically verified [123]. This study confirmed that the SNOMED codes generated clinically useful data [123].

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 may 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.

2.3 Aim and Objective

One of the objectives of this study was to describe the methodology used to report PCa data using SNOMED lookup tables.

2.4 Methods

2.4.1 Study design

The retrospective descriptive study design was used for this study.

2.4.2 Inclusion /Exclusion Criteria

Data was limited to prostate biopsies recorded for men ≥30 years.

2.4.3 Sample Population

Men who received a prostate biopsy in the Gauteng Province using convenience sampling.

2.4.4 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.

2.4.5. Data extraction and preparation

Retrospective prostate biopsy data for the period 2006 to 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, MA, USA) query tool through the use of pattern matching and using the fuzzy string search function [124]. Two data extracts were received: (i) prostate biopsy and (ii) chained M and T code/s captured for each biopsy.

The prostate biopsy extract included the following variables: (i) episode number, (ii) unique patient identifier (generated using a probabilistic matching algorithm [125]), (iii) age, (iv) gender, (v) race (where populated), (vi) facility code, (vii) facility name, (vii) reviewed date and (viii) biopsy results text (unstructured narrative report detailing histological findings). The SNOMED data extract included the following variables: (i) episode number, (ii) chained (comma separated) M code/s (comma delimited), e.g. M-00100, M-43000, M-72450 and (iii) chained T code/s, e.g. T-92000, T-74000).

A high-level overview of the SNOMED code manipulation and lookup table development is described in Figure 2.1.

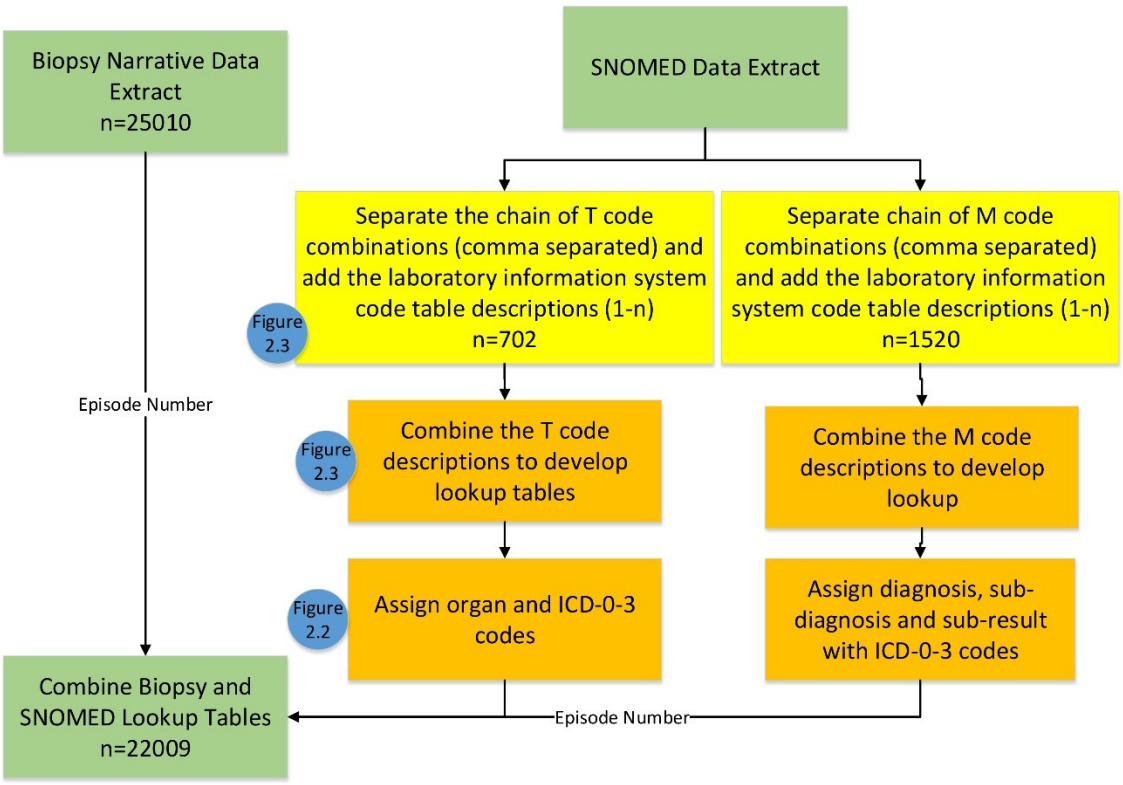


Figure 2.1: High-level overview of the steps taken to code each chained SNOMED code description combination. There were two separate SNOMED data extracts from the laboratory information system; morphology (M) and topography (T). The colour coding indicates the various processes: (i) green: data extracts, (ii) yellow: SNOMED code manipulation in preparation for lookup table development and (iii) orange: lookup tables with coded variables. The data extract was prepared using the two lookup tables to generate the following new coded variables: (i) organ, (ii) organ ICD-0-3, (iii) diagnosis, (iv) diagnosis ICD-0-3, (v) sub-diagnosis, (vi) sub-diagnosis ICD-0-3 and (vii) 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.

The lookup tables were developed in a two-step process: (i) code manipulation to combine descriptions and (ii) coding the lookup tables. The blue circles in the figure above indicate which figures provide additional details for each step, e.g., data manipulation and query to combine data extracts and lookup tables in a single database query described in the figures below.

2.4.6 Combining SNOMED 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. By chained we refer to multiple SNOMED CT codes that separated by commas and stored in one row of data, e.g. ‘M-00100, M-72400, M-43000’. This format is not

easy to analysis. Data was prepared using Microsoft Excel (Microsoft Corporation, Redmont, Washington, USA) [126]. The first step was to use the text to columns wizard such that each value was in a separate column. Next, for each code we used the VLOOKUP function to populate the matching code description, e.g., the description for M-00100 was 'Normal tissue (finding)'. Finally. we concatenated all the code descriptions. For the lookup table development, the unique SNOMED CT M and T code combinations and descriptions were extracted. The code manipulation is described in detail (Figure 2.2).

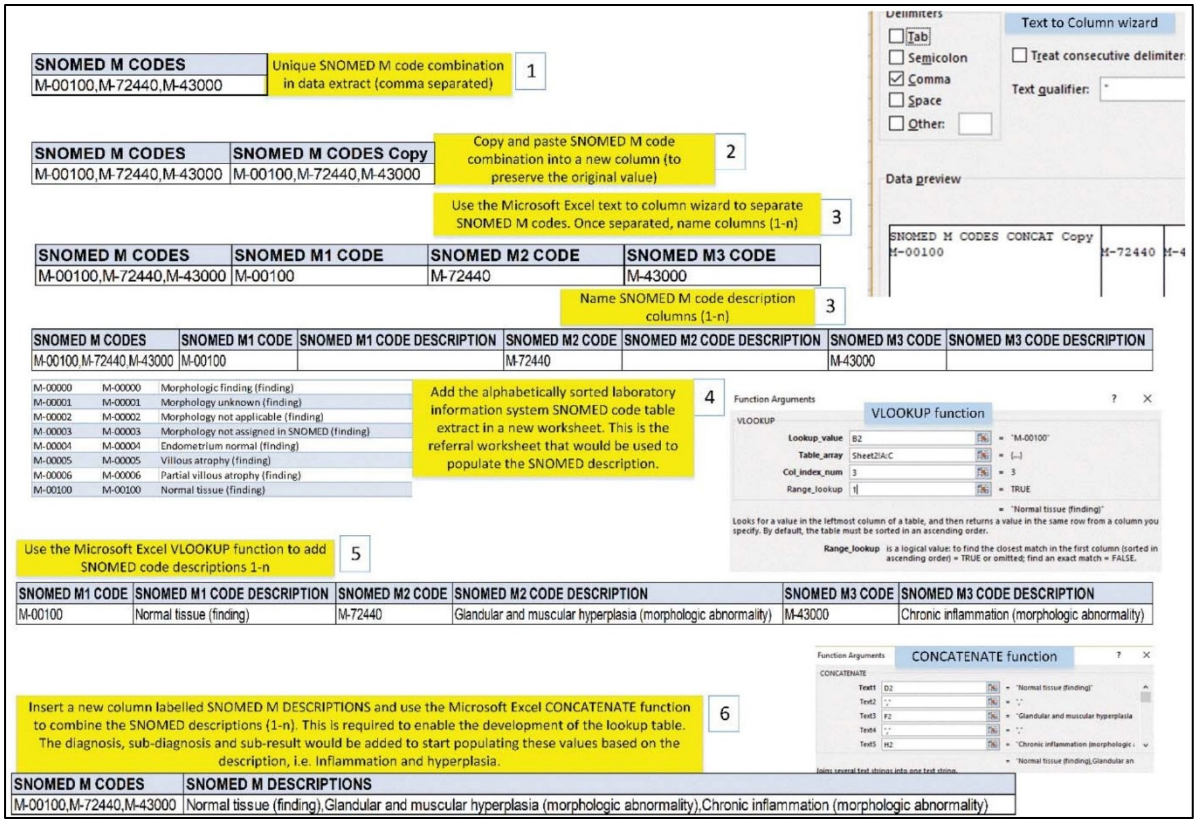


Figure 2.2: Six step procedure used to transform the chained comma separated SNOMED M codes into individual columns (leaving the original value intact) to add the laboratory information system code table descriptions (in preparation for lookup table development). The same procedure was conducted for 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 SNOMED 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-00100 code description is 'Normal tissue (finding)' and (6) Microsoft Excel CONCATENATE function was used to combine code descriptions in a new column.

2.4.7 Coding the SNOMED M and T lookup tables

We used the prepared M and T code/s 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: (i) diagnosis, (ii) diagnosis ICD-0-3 code, (iii) sub-diagnosis, (iv) sub-diagnosis ICD-03 code and (v) sub-result with guidance from an anatomical pathologist and 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 2.1. We captured the matching ICD-0-3 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-0-3: 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 reporting order to assign a benign sub-diagnosis as follows: (i) inflammation, (ii) hypertrophy, (iii) hyperplasia, (iv) oedema, (v) atrophy and (vi) adenosis. This made it easier to assign these finding in a 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-0-3 code (C61.9 for Prostate).

2.4.8 Combining the lookup tables with the prostate biopsy data

Microsoft Access (Microsoft Corporation, Redmont, Washington, USA) was used to combine the datasets in a single query: (i) prostate data extract table, (ii) SNOMED M data, (iii) SNOMED M lookup table, (iv) SNOMED T data and (v) SNOMED T lookup table. Figure 2.3 describes how these tables were combined using a left outer join in a relational database [127]. This join type ensures that all rows from the prostate data extract table were populated with only the matching values from the other tables reported using referential integrity [127, 128].

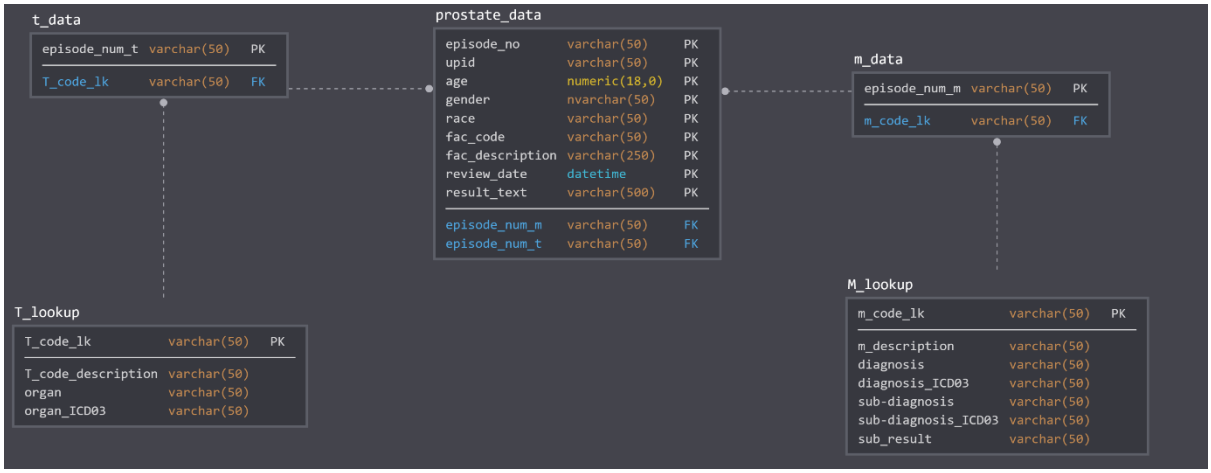


Figure 2.3: Relational database diagram describes how the various tables were joined (left outer). For each table, the primary and foreign keys are provided. The lookup tables contain only the unique SNOMED code combinations. The lines indicate a relationship join between tables. Once the table joins are implemented, variables from any table can be reported. Structured query language (SQL) could be used to combine the required data for analysis.

This query contained all the variables for the data analysis. An example of how data from the lookup tables were populated is provided in Table 2.1. The SQL code for the query has been uploaded on GitHub (https://github.com/naseemc26/SNOMED-CT-Lookup-Table).

Table 2.1: Example of four prostate biopsies where the SNOMED code descriptions were assigned a diagnosis and sub-diagnosis including the allocation of ICD-0-3 codes. The biopsy result text is also provided. Episode numbers are anonymised.

Episode no	SNOMED M Codes (T Code)	Biopsy Result Text	Date of Review	Gleason score	SNOMED M DESCRIPTIONS	Diagnosis (ICD-0-3)	Sub-Diagnosis (ICD-0-3)
A	M-72450,M-00100 (T-92000)	&	2009-10-19	N/A	Adenofibromyomatous hyperplasia (morphologic abnormality) Normal tissue (finding)	Benign/negative for malignancy (8000/0)	Hyperplasia
B	M-72000 (T-92000)	$	2013/03/12	N/A	Hyperplasia (morphologic abnormality)	Benign/negative for malignancy (8000/0)	Hyperplasia
C	M-00100,M-81403 (T-92000)	#	2014/02/16	4 + 3 =7	Normal tissue (finding) Adenocarcinoma, no subtype (morphologic abnormality)	Neoplasm, malignant (8000/3)	Adenocarcinoma (8140/2)
D	M-71000,M-40000 (T-92000)	%	2016/10/28	N/A	Hypertrophy (morphologic abnormality) Inflammation (morphologic abnormality)	Benign/negative for malignancy (8000/0)	Inflammation and hypertrophy

[&]EPISODE NUMBER: A CLINICAL HISTORY: AN 86 YEAR OLD MALE PATIENT WITH A PSA = 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: 8.5G 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 PSA = 273. HARD, MULTINODULAR PROSTATE. TRUS BIOPSY DONE. MACROSCOPY: TWELVE BIOPSIES, THE LONGEST 1.8CM. 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 PSA LEVEL OF 10,4. MACROSCOPY: THIRTEEN CORES, THE LONGEST OF WHICH MEASURED 11MM AND THE SHORTEST 2MM. 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.

2.4.9 SNOMED M and T 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, USA) [129]. The descriptive analysis was conducted for prostatic biopsies with the M code captured. Test volumes for the top ten M code combinations with their chained descriptions were reported (in descending order). The number of biopsies of prostatic origin was also reported.

2.4.10 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 ten sub-diagnoses were reported, the first ten were reported and the remaining grouped as ’Other‘.

2.4.11 Evaluating the accuracy of the SNOMED CT codes

To evaluate the accuracy of the SNOMED CT codes captured by the pathologist, we extracted 1000 random prostate biopsies with a ‘Benign/negative for malignancy’ or ‘Neoplasm, malignant’ diagnosis. We added an ID variable that assigned each biopsy a number from 1 to n. We then used a structured query language (SQL) to randomly extract 1000 narrative reports for manual coding of the diagnosis by two experts. We used a contingency table to compare the SNOMED CT and manually coded diagnosis to report the positive predictive value (PPV). The 2020 Globocon report indicated that the 5 year prevalence of PCa was 39,863 cases (7,973 per year) [130]. As this was reported for a population of 100,000, we determined a national PCa prevalence of 7.97% [130]. We reported the sensitivity, specificity and PPV.

2.5 Results

Using our methodology, we analysed 25 010 biopsy results for this study. For the lookup tables, there were 1520 M and 702 T unique code combinations. Our results are presented in the following paragraphs.

2.5.1 SNOMED M and T code descriptive analysis

M codes were provided for 22 195/25 010 biopsies (89%). The T code was provided for 24 546/25 010 (98%) biopsies. There were 22009/25010 biopsies with both an M and T code populated (88%). There were 2 815/25 010 (11%) biopsy reports could not be analysed using lookup tables because there was no M and T provided (Table 2.2).

Table 2.2: Contingency table to assess the percentage of SNOMED M and T codes populated using the mapping table for prostate biopsies between 2006 and 2016 in the Gauteng Province, South Africa

	T Code/s captured	T Code/s not captured	Total
M Code/s captured	22 009 (88%)	186 (1%)	**22 195 (89%)**
M Code/s not captured	2 537 (10%)	278 (1%)	**2 815 (11%)**
Total	**24 546 (98%)**	**464 (2%)**	**25 010 (100%)**

Descriptive analyses were conducted for 20 551/25 010 (82%) biopsies of prostatic origin. M-40000 ('Inflammation') was the most commonly reported sub-diagnosis (15.5%). This was followed by two adenocarcinoma combinations (M-81403 & M-81403, M-80003) reported for 12.1% and 9.8% of biopsies respectively (Table 2.3).

Table 2.3: Top ten most commonly requested SNOMED M code combinations from the prostate biopsy data between 2006 and 2016 in the Gauteng Province, South Africa.

M Codes	M code descriptions	Total	%
M-40000	Inflammation (morphologic abnormality)	3 440	15.5%
M-81403	Adenocarcinoma, no subtype (morphologic abnormality)	2 677	12.1%
M-81403, M-80003	Adenocarcinoma, no subtype (morphologic abnormality) Malignant neoplasm, primary (morphologic abnormality)	2 166	9.8%
M-00100	Normal tissue (finding)	1 668	7.5%
M-80003, M-81403	Malignant neoplasm, primary (morphologic abnormality) Adenocarcinoma, no subtype (morphologic abnormality)	1 596	7.2%
M-00100,M-81403	Malignant neoplasm, primary (morphologic abnormality) Adenocarcinoma, no subtype (morphologic abnormality)	1 299	5.9%
M-00100, M-72440	Normal tissue (finding) Glandular and muscular hyperplasia (morphologic abnormality)	1 081	4.9%
M-80003	Malignant neoplasm, primary (morphologic abnormality)	706	3.2%
M-09350	Morphologic description only (finding)	394	1.8%
M-72440	Glandular and muscular hyperplasia (morphologic abnormality)	370	1.7%

2.5.2 Descriptive analysis of diagnosis and sub-diagnosis

Using this approach, we noted 10 441 (51%) biopsies with a benign diagnosis (ICD-0-3:8000/0) and 9 302 (45%) biopsies with a malignant diagnosis (8000/3). Atypia was reported for 760 biopsies (4%). An uncertain (whether benign or malignant) diagnosis (8000/1) was reported for 0.2% of biopsies (Table 2.4). In addition to a malignant finding, this approach also provided more details for benign findings. Inflammation was reported as a sub-diagnosis for 4075 biopsies (39%). This was followed by no pathologic diagnosis and hyperplasia at 27% (n=2 809) and 21% (n=2 174) respectively. Inflammation and hyperplasia were reported in a

number of combinations contributing ~49% and ~29% of the top ten benign sub-diagnoses. Cumulatively, the top ten sub-diagnoses reported represented 99% (n=10 320) (Table 2.4) of all benign diagnoses. The majority of samples with a malignant diagnosis reported an adenocarcinoma (88%) sub-diagnosis. There were 408 biopsies with a carcinoma (4%) sub-diagnosis. The malignant neoplasm sub-diagnosis was reported for 693 (7%) of biopsies where it was not possible to differentiate the tissue type, reporting predominantly the M-80003 code (Malignant neoplasm, primary) [122]. The majority of biopsies with an atypia/dysplasia diagnosis reported an atypia sub-diagnosis (81%) followed by dysplasia (11%). Only 2% of biopsies with an atypia/dysplasia diagnosis reported a high-grade intraepithelial lesion. Only 0.2% of biopsies reported an uncertain sub-diagnosis (8000/1).

Table 2.4: Descriptive analysis of diagnosis and sub-diagnosis where both a SNOMED M and T code are populated of prostatic origin between 2006 and 2016 in the Gauteng Province, South Africa. Where appropriate, the ICD-0-3 codes are provided in brackets.

Diagnosis	***n***	***%***
Benign/negative for malignancy (8000/0)	10 441	50.8%
Neoplasm, malignant (8000/3)	9 302	45.3%
Atypia/Dysplasia (8000/0)	760	3.7%
Neoplasm, uncertain whether benign or malignant (8000/1)	48	0.2%
Sub-diagnosis		
Benign/negative for malignancy	**n**	**%**
Inflammation	4 075	39.0%
No pathologic diagnosis	2 809	26.9%
Hyperplasia	2 174	20.8%
Inflammation and hyperplasia	788	7.5%
Inflammation and hypertrophy	123	1.2%
Atrophy	120	1.1%
Hypertrophy	82	0.8%
Inflammation and atrophy	63	0.6%
Hyperplasia and atrophy	55	0.5%
Inflammation, hypertrophy and hyperplasia	31	0.3%
Other	121	1.2%
Neoplasm, malignant sub-diagnosis	**n**	**%**
Adenocarcinoma (8140/3)	8 201	88%
Carcinoma (8010/3)	408	4%
Malignant neoplasm (8000/3)	693	7%
Atypia/Dysplasia sub-diagnosis	**n**	**%**
Atypia	616	81%
Dysplasia	87	11%
Atypia/Dysplasia	44	6%
High grade intraepithelial lesion (8148/2)	13	2%
Neoplasm, uncertain whether benign or malignant sub-diagnosis	**n**	**%**
Neoplasm, uncertain whether benign or malignant (8000/1)	48	0.2%

2.5.3 Evaluating the accuracy of the SNOMED CT codes

For the random sample of 1 000 biopsies, 556 had a SNOMED CT lookup table assigned benign (55.6%) compared for a 444 malignant diagnosis (44.4%). This is very similar to the findings reported in Table 2.4 indicating that the SQL random selection was satisfactory. We reported a sensitivity and specificity of 1.00 and 1.00 respectively. When combined with a

prevalence of 7.97%, our analysis reported a PPV of 0.96 for 1000 random biopsies (Table 2.5).

Table 2.5: Contingency table used to determine the accuracy of the pathologist assigned SNOMED CT codes for 1000 random prostate biopsies with a 'Benign/negative for malignancy' or 'Neoplasm, malignant' diagnosis. These biopsies were randomly selected from prostate biopsies that were performed between 2006 and 2016 in the Gauteng Province, South Africa.

		SNOMED CT Assigned Diagnosis		Total
		Neoplasm, malignant (+)	Benign/negative for malignancy (-)	
7Manually Coded Diagnosis	Neoplasm, malignant (+)	442	2	**444**
	Benign/negative for malignancy (-)	2	554	**556**
Total		**444**	**556**	**1000**
Sensitivity		1.00		
Specificity		1.00		
Positive Predictive Value		0.96		

2.6 Discussion

We showed that it was possible to automate prostate biopsy reporting using commonly available tools such as Microsoft Excel and Access. Our study set out to demonstrate the use of SNOMED lookup tables generated to describe PCa in the Gauteng Province. It was possible to automate PCa reporting for 88% of prostatic biopsies. The use of lookup tables described in our study is illustrated in Figure 2.3. Our study has demonstrated that it was possible to report PCa data in an automated fashion for 88% of biopsies without the need for manual coding. The use of ICD-0-3 codes for malignant findings facilitate PCa reporting similar to cancer registries [116].

Our study has proposed a different approach to the registration and surveillance of PCa in South Africa. The first step is to automate PCa reporting at the corporate data warehouse (CDW) in the NHLS, through the use of the lookup tables developed for our study. Lookup tables are routinely used by the CDW to transform laboratory data for reporting. A good example is microbiology samples were where the multiple coded results are transformed to a single value. For example, a negative HIV serology result may be reported as 'NEG', 'N' or 'NEGATIVE' by different laboratories. The lookup table will transform all three values to 'NEG' to facilitate uniform reporting. 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, sub-diagnosis within three months of diagnosis. There is a trickle feed interface between the LIS and the CDW that replicates data between the data systems on a daily basis. All biopsies with a histological finding reported on the LIS will be replicated to the CDW within ~2-3 hours. Based on the short data transmission time, the CDW has been identified within the NHLS as a health information exchange (HIE) by providing instantaneous results to mobile Health (m-Health) applications and patient management systems such as Therapy Edge™ [131].

Using lookup tables will facilitate a constant feed of analysed PCa data to the South African NCR to ensure more timeous reporting. The approaches applied in our study could be used to provide PCa findings for laboratory data up to 2020. To achieve this, the lookup tables have to be updated for any new SNOMED code combinations not reported in our study. With the updated lookup tables, PCa data could be reported within three weeks. By providing this data at shorter intervals, it will be possible to triangulate 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 [132]. Unpublished data for our study in the Gauteng Province has revealed that between 2012 and 2016, PCa incidence per 100 000 has increased from 44.92 to 57.31 per 100 000 compared to 46.53 reported by the NCR in 2012 [133].

This highlights the importance of providing up to date PCa ASIR data at both the national, provincial and health district levels to identify hotspots where programmatic interventions are required. The data extract includes the health facility information. Based on the health facility, the Province, district and sub-district are also provided in routine laboratory data systems to facilitate reporting, for example, at the various health care levels. This has been demonstrated by Carmona *et al* that have used laboratory data to report HIV epidemiological data at the different health care levels from national to the health facility [134].

Another advantage of this approach is that PCa data from other African countries could also be analysed using the lookup tables developed as long as they are using a LIS. This will dramatically improve PCa reporting across Africa. Antoni *et al* assessed the methods used for reporting the 2018 GLOBOCON estimates [135]. For 14/51 African countries, incidence estimates were based on simple average rates from neighbouring countries (27%) [135]. For South Africa, projections of national incidence were sourced from the NCR [135]. There were only 7/54 countries that reported regional incidence rates [135]. 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 to PCa 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 table 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 population in 2018 in South Africa [22]. The number of incident cancer cases were 2 727, 14 097 and 12 983 respectively [22]. Adopting the approaches applied in our study will improve cancer reporting at both the national and regional levels in South Africa.

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 to 2011 was undertaken using SNOMED codes [123]. In the Danish cancer registry, SNOMED codes were transcoded to a meaningful and analysable format using a number of computational steps: (i) SNOMED code combinations were translated into text strings, (ii) the text strings were then recoded to assign the procedure type, tissue type, and diagnosis and finally (iii) prioritise and summarise conclusions for each biopsy, e.g. ‘bGS 3 + 4’ (prostatic biopsy with adenocarcinoma and a 3 + 4 =7 GS and bN (biopsy of the prostate without malignancy) [123]. The Danish cancer registry predominantly reported data for PCa, whereas our study reported data for negative biopsies as well. In our study, the combination of the diagnosis, sub-diagnosis and GS reported in an unpublished study could be used to report data similar to the Danish cancer registry, e.g. diagnosis of ‘Neoplasm, malignant’, sub-diagnosis of ‘Adenocarcinoma’ and a GS of 3 + 3 =6 are coded as ‘bGS 3 + 3’ [123].

SNOMED lookup tables could be used to focus programmatic interventions for geographic areas with higher PCa burden. Similar initiatives using laboratory data have been undertaken locally for HIV and TB services [125, 136, 137]. An example is the World Bank report that described a spatial clustering analysis of HIV Viral Load suppression at the national, provincial, district, sub-district and health facility levels [125]. 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 [125]. This information can be accessed by health care workers using the epidemiological dashboard developed to identify health facilities that are performing poorly [125, 136]. Similar work has also been published using CD4 data to highlight areas where HIV+ patients presenting for care have a higher burden of advanced disease [136].

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-0-3 codes are really important for PCa cases, we felt that specifically for benign findings the additional information was required especially for inflammation and hypertrophy. Inflammation has been shown to be linked with many other cancers [138, 139]. Several studies have indicated that chronic inflammation has a potential role in prostatic carcinogenesis and tumour progression [138, 140, 141]. Nelson *et al* reported that chronic or recurrent inflammation may play a role in the development of PCa [142]. To assess inflammation and hyperplasia trends, the sub-diagnosis will provide important clinical information for biopsies with a non-PCa finding. By 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.

The assimilation of health data into workable and user-friendly dashboards has had a big impact on how health data is used locally [143-145]. In the medium term, it will 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 at provincial, district and sub-district levels to identify areas where interventions are required in South Africa. 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 (LTFU) and waiting time data for patients whom had presented with an elevated PSA and whom were histologically confirmed with PCa using the CDW probabilistic matching algorithm [146].

Finally, to improve SNOMED reporting in the LIS, it is recommended that the M and T codes be defined as mandatory fields. This will ensure that biopsies may only be authorised by an AP once the requisite codes have been captured. The lookup tables developed should be loaded on the CDW and tested against national data. Any new M and T code combinations should be updated to automate the reporting of PCa data. 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.

2.7 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. They could also be applied to report data for other cancers of public health interest.

CHAPTER 3 - USING TEXT MINING TECHNIQUES TO EXTRACT PROSTATE CANCER PROGNOSTIC INFORMATION (GLEASON SCORE) FROM SEMI-STRUCTURED NARRATIVE LABORATORY REPORTS IN THE GAUTENG PROVINCE, SOUTH AFRICA

For uniformity of layout of this thesis, this chapter is laid out in the style of the thesis. This paper has been re-submitted to the BMC Medical Informatics and Decision Making after substantial revision. The proof of submission is provided in Appendix B.

Abstract

Background:

PCa is the leading male neoplasm in South Africa with an ASIR of 68.0 per 100 000 population for 2018. Local guidelines specify PCa risk categories using the PSA, GS and clinical stage. The GS is the strongest predictive factor for PCa treatment, embedded within the semi-structured narrative prostate biopsy narrative report making it difficult to extract. Hence, there is a paucity of local data with all studies manually extracting the GS. The objective of our study was to explore the use of text mining techniques to extract relevant cancer-related predictive information from irregularly reported text-intensive patient reports.

Methods:

The retrospective descriptive study design was used to explore clinical text-based prostate biopsy narrative reports generated by the NHLS for men aged >30 years between 2006 and 2016 in the Gauteng Province, South Africa. We used the associated SNOMED CT M and T to identify prostate biopsies with an adenocarcinoma histological finding. We then extracted a random sample of 1000 biopsies with a PCa diagnosis, defined as adenocarcinoma with a GS reported, using Microsoft Excel (Redmond, Washington, USA). We also randomly extracted 1000 prostate biopsies with a PCa diagnosis that were submitted from private sector laboratories to the NCR, referred to as the validation dataset. For both datasets, the GS was manually coded by two experts. We developed a text mining algorithm to extract the GS from both datasets, using the following steps: (i) data acquisition (ii) pre-processing and (iii) feature extraction, (iv) feature value representation, (v) feature selection, (vi) information extraction, (vii) classification and (viii) discovered knowledge. Pre-processing was used to convert the biopsy reports to a machine-readable document format. Next, we used the tokenization, stopwords removal and stemming natural language processing techniques. We used N-grams as our feature extraction strategy to extract the major and minor score. For feature value representation, we created a document term matrix using term frequency. This was used to

transform the document into a numeric feature vector space. We used expert driven feature selection, with the following features of interest identified: (i) episode number, (ii) major score, (iii) minor score, (iv) total score and (v) combined score. Once the features were selected, we used regular expressions representative of the GS target feature such as “gleason”, “Gleason”, “GLEASON”, “Gleeson” etc. For information extraction, we manipulated the N-grams output to extract the numerical value of the major and minor scores. The algorithm also automatically classified biopsies into the following three risk classes based on the extracted total score: (i) low, (ii) intermediate and (iii) high-risk. The discovered knowledge consisted of the episode number, major, minor, total score, standardised GS and risk category for each narrative report. A confusion matrix was used to compare the text mining predicted values against the manually coded GS for both datasets, reporting the precision, recall and F-score. The different GS alpha, numeric and alphanumeric reporting formats in the narrative report were identified and reported as a table. We reported the top five GS, with the remaining scores categorised as “Other” for both datasets, with the percentage of high-risk GS reported. For the manually coded and predicted three GS risk categories, we also reported the F-score for each level as well as the macro-averaged F-score.

Results:

The use of text mining techniques successfully enabled the accurate extraction of the GS from narrative prostate biopsy reports. The first output reported an F-score of 0.99 that improved to 1.00 after the amendment of the text mining algorithm. The algorithm was evaluated using the validation dataset and reported an F-score of 0.99. We identified ten different GS reporting formats. The five most commonly reported GS were 5 + 4 =9 (17.6%), 3 + 3 =6 (17.5%), 4 + 3 =7 (16.4%), 3 + 4 =7 (14.7%) and 4 + 4 =8 (14.2%). For the validation dataset, the most commonly reported GS were: (i) 3 + 3 =6 (37.7%), (ii) 3 + 4 =7 (19.4%), (iii) 4 + 3 =7 (14.9%), (iv) 4 + 4 =8 (10.0%) and (v) 4 + 5 =9 (7.4%). A high-risk GS was reported for 31.8% of biopsies compared to 17.4% for the validation dataset. For the three GS risk categories, a macro-averaged F-score of 0.99 was reported.

Conclusions:

We demonstrated the reliable extraction of the GS with a text mining algorithm to assess late presentation. The developed algorithm has the potential to automate GS extraction to describe late presentation and address the paucity of local data. The next steps include the use of an application programming interface to automate the routine extraction of the GS for national laboratory data across South Africa. The techniques demonstrated in our study could be applied to other cancers of public health importance. Furthermore, a supervised ML that has shown similar results to the SNOMED CT lookup tables could be used address data gaps. In

addition, deep learning approaches should be investigated, which could be moved to a cloud service to automated PCa surveillance data generation across Africa.

Abbreviations (including information block)

Abbreviation	Description	Information Block
AI	Artificial Intelligence	Simulation of human brain processes (intelligence) by computer systems.
API	Application Programming Interface	A set of programming code used to develop software applications. Installed software application a computer that executes a defined set of instructions.
ASIR	Age-standardised incidence rate	Incidence measure for a population assuming a standard age structure.
CSV	Comma separated values	Type of data extract where values are separated (delimited) by commas
EAU	European Association of Urology	Guidelines authority on Europe
GS	Gleason score	Grading system used to determine the aggressiveness of prostate cancer. An important prognostic indicator for prostate cancer
IDE	Integrated Development Environment	Programming application that provides for computer programmers to develop new software and applications. An IDE provides the facility to write, compile, debug and execute programme code.
M	Morphology	Describes microscopically observed changes in cells and tissue biopsies.
ML	Machine learning	The application of artificial intelligence (AI) that provides the ability to automatically learn and improve from experience without requiring explicit programming.
NLP	Natural Language Processing	The application of computational techniques to the analysis and synthesis of natural language and speech.
NLTK	Natural Language Toolkit	A suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the English language written in the Python programming language
PCa	Prostate cancer	Prostate cancer occurs in the prostate gland. It is the most common male cancer worldwide
PSA	Prostate specific antigen	Prostate specific antigen (PSA) is a laboratory test used for prostate cancer screening. A prostate biopsy clinically indicated by an elevated PSA result and/or abnormal digital rectal examination.
SNOMED CT	Systematized Nomenclature of Medicine clinical terms	Systematized Nomenclature of Medicine clinical terms (SNOMED) clinical terms (CT) are codes used to describe the anatomy and histological findings of tissue. Histological findings obtained from microscopically examining prostate biopsies entered by Anatomical Pathologists based on the laboratory information system (LIS).
T	Topography	SNOMED CT codes that indicate a specific location in the body of the biopsy, e.g., prostate

3.1 Introduction

Globally, PCa is an important NCD due to both population growth and a concomitant increase in life expectancy [36, 37]. It is the leading male neoplasm in South Africa with an ASIR of 68.0 per 100 000 population in 2018 [22].

Local guidelines indicate that men with PCa are assigned to risk categories using the PSA result, GS and clinical stage [47, 52]. According to the EAU guidelines, the GS is the strongest predictive factor for PCa treatment [26]. The GS consists of a primary (1 to 5) and a secondary pattern/score (1 to 5) [26]. For example, a GS of 3/5 (primary or major) and 4/5 (secondary or minor) equates to a total score of 3 + 4 =7. Local guidelines categorise PCa risk using the GS as follows; (i) GS ≤6: low-risk, (ii) GS =7: intermediate-risk and (iii) GS ≥8: high-risk [47, 52]. Patients with a high-risk GS have a poorer prognosis with an increased risk of metastatic progression and death [26]. For these patients, the PCa mortality risk is 60 to 87% compared to between 42 and 70% for an intermediate-risk GS [26].

Across the NHLS, a LIS is used to record, manage, and store patient laboratory reports and related demographic health data [147, 148]. This LIS documents all processes within the laboratory workflow including sample registration, test order generation, tracking orders and reporting results [147, 148]. For AP reporting, the assigned pathologist voice-records the biopsy narrative report for electronic capture by data typists. The pathologist reviews the captured narrative report prior to authorisation. Thereafter, the patient's laboratory report is printed and/or electronically delivered to the health facility.

Some structured laboratory results (numeric or alphanumeric) are relatively easy to extract from the LIS, e.g., a potassium result from a renal function panel. However, free text laboratory reports have limited standardization regarding content, formatting and language, i.e., they are pathologist dependent. This is the case for AP narrative biopsy laboratory reports. Table 3.1 provides an example of a semi-structured narrative biopsy report that includes the headings clinical history, macroscopy and pathological diagnosis (highlighted in bold).

Table 3.1: Example of the semi-structured narrative prostate biopsy report. The narrative biopsy report included the headings clinical history, macroscopy and pathological diagnosis.

Category	Biopsy Report
Biopsy report	EPISODE NUMBER: ABC1234 **CLINICAL HISTORY**: A 67 YEAR OLD MALE PATIENT WITH A PSA OF 7,9UG/L. PROSTATE BIOPSIES HAVE BEEN DONE. **MACROSCOPY**: SIXTEEN CORES OF TISSUE, THE LONGEST MEASURING 15MM AND THE SHORTEST MEASURING 7MM. **PATHOLOGICAL DIAGNOSIS**: PROSTATE CORE BIOPSIES SHOWING THE FOLLOWING FEATURES: AN INVASIVE PROSTATIC ADENOCARCINOMA. TWO CORES ARE INVOLVED AND <5% OF THE TISSUE. GLEASON 4, 3. PERINEURAL AND LYMPHOVACULAR INVASION ARE NOT IDENTIFIED. IMMUNOHISTOCHEMISTRY: IN THE PRESENCE OF ADEQUATE POSITIVE CONTROLS, IMMUNOHISTOCHEMICAL STAINS HAVE BEEN DONE AND THE FOLLOWING RESULTS OBTAINED: P63 AND CK5/6: BASAL CELLS ARE NOT DEMONSTRATED IN THE ATYPICAL GLANDS.

The GS is reported as embedded text within semi-structured narrative biopsy reports in alpha, numeric as well as alphanumeric formats. As a result, the GS score could be captured in a variety of patterns based on the local AP practices. For example, a GS of 4+4=8 maybe captured in a variety of formats such as: (i) 4 + 4 = 8, (ii) 8 (4+4), (iii) 4;4 and (iv) major 4, minor 4. Spacic *et al* have reported that the linear structure of the GS makes it amenable to modelling using regular expressions [149]. In contrast, various cancer specific vocabularies and classification systems as well as ontologies have been used with text mining to extract structured information from narrative biopsy reports [149]. These vocabularies and ontologies work well with coding systems such as ICD-O-3, SNOMED CT and ICD-10 for example [149]. Such vocabularies and ontologies do not exist for the GS. Therefore, the extraction of this information is time-consuming requiring manual coding employed by most local studies, resulting in small sample sizes [1-3]. Hence, there is a paucity of local GS data describing late presentation for PCa in South Africa.

The NCR conducts annual national cancer surveillance in South Africa using laboratory data [114]. The registry manually codes voluntarily submitted cancer reports from both public and private laboratory service providers [114]. Trained coders read the narrative biopsy reports and extract relevant cancer information and add these details to the NCR reporting database [114]. Details extracted by the coders include the patient's primary anatomical cancer site and tumour morphology according to the ICD-O-3 [115]. Manual coding results in delays from cancer reporting to surveillance data generation with the 2016 NCR report only available by 2020 [150]. The NCR reports the annual PCa ASIR; however; the predictive GS is not

reported. Therefore, accessing and reviewing such narrative reports for review of, or research into, patient cohorts, would entail printing and physically reading and manually extracting and capturing clinical information. This is both time consuming and delays reporting.

Globally, AI has been used to automate decision making by mimicking human cognitive function by using mathematical, statistical, logical, and computer programming approaches [32, 103]. The AI model can be trained using existing data; it could be applied to new data to automate decisions [103]. AI can also be applied to semi-structured healthcare data using techniques such as natural language processing (NLP) [32]. This is achieved by employing computational techniques to extract semantic meaning from text [107, 151]. In essence, these NLP procedures convert text to machine-readable structured data [32]. These computational approaches such as tokenisation help to identify words and punctuations within a sentence [151]. In summary, NLP can be used to extract clinical information from unstructured data to supplement and enrich structured medical data [32].

There is a need to develop automated algorithms that are able to extract the GS from narrative prostate biopsy reports. The objective of our study was to explore the use of text mining techniques and regular expressions to extract the predictive GS from irregularly reported text-intensive patient reports. The aim of the algorithm is to extract the GS to improve statistical reporting and surveillance rather than inform clinical action which is based on the narrative report delivered to the health facility. A limitation of the algorithm is that without the associated clinical information, it is not able to inform care. Furthermore, the algorithm should be validated prior to use for routine surveillance. In some cases, there is no need to use machine learning approaches when text mining using regular expressions would be sufficient to extract the GS.

3.2 Materials and methods

3.2.1 Study design

The retrospective descriptive study design was used.

3.2.2 Ethical considerations

Ethics clearance was obtained from the University of the Witwatersrand (M170419). Our study did not contain any patient identifiers.

3.2.3 Text mining algorithm development

We used the Python Spyder integrated development environment (IDE) for the development of the text mining algorithm because of its robustness in advanced editing, debugging,

profiling, data exploration and interactive execution [152, 153]. An IDE is software that is used to build and develop applications. The Python code for this algorithm has been uploaded on GitHub (https://github.com/VictorO2/text-mining-gleason-score). The following Python modules were imported: (i) operating system, (ii) pandas, (iii) time, (iv) matplotlib, (v) seaborn, (vi) WordCloud and (vii) Natural Language Toolkit (NLTK). The table below provides a description for the purpose for each library in the text mining algorithm.

Table 3.2: Description of the various python libraries used for the text mining algorithm to extract the Gleason score from narrative prostate biopsy reports.

Python library	Purpose for using this library in the text mining algorithm
Operating System (OS)	To define working directory of all Python files.
Pandas	Used for data cleaning, analysis and manipulation.
Time	For time operations and measuring of script efficiency.
Matplotlib/Seaborn	Used to visually represent algorithm findings.
WordCloud	Used to create a basic word cloud/s from the documents.
Natural Language Toolkit	Used for language processing.

We followed the text mining pipeline as depicted in the flowchart below (Figure 3.1). The logical steps for the text mining algorithm were as follows: (i) data acquisition (ii) pre-processing, (iii) feature extraction, (iv) feature value representation, (v) feature selection, (vi) information extraction, (vii) classification and (viii) discovered knowledge.

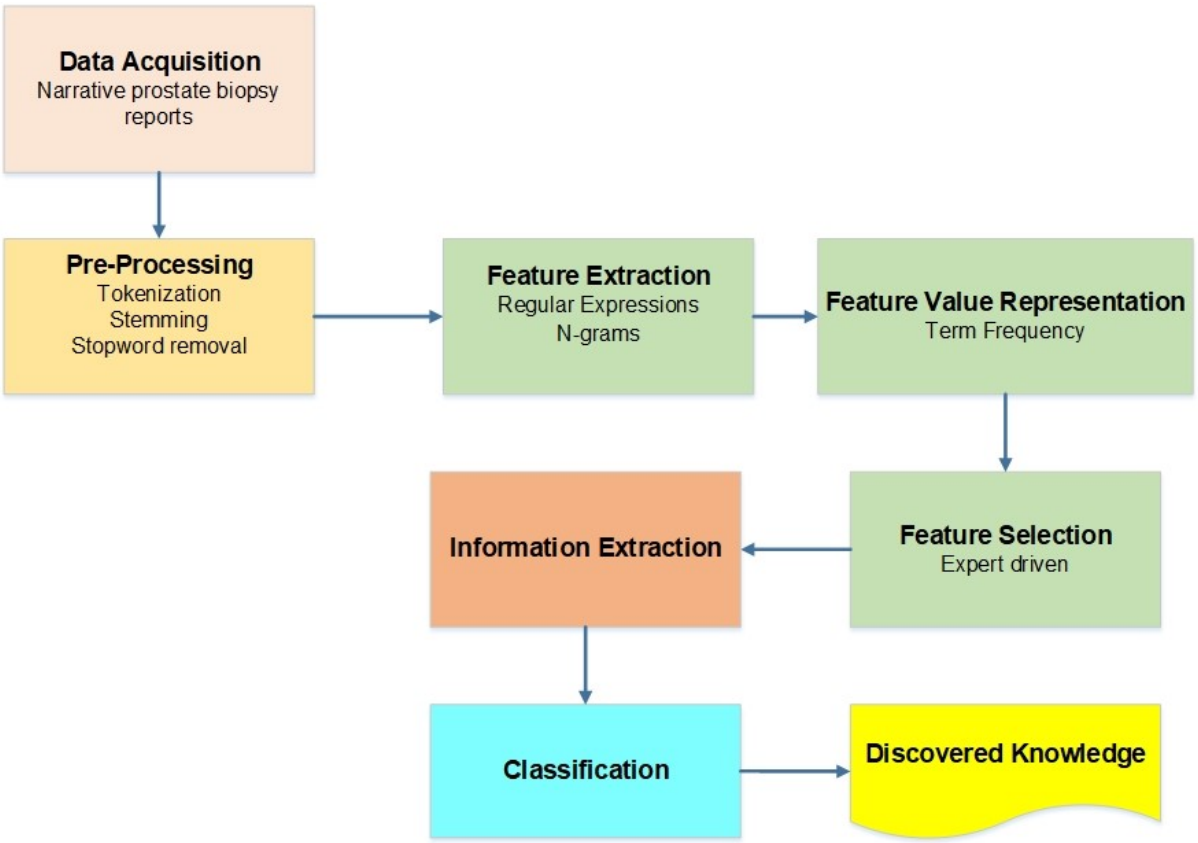


Figure 3.1: Diagram describing the logical processes used to analyse the raw narrative prostate biopsy report to generate the discovered knowledge. The steps were as follows: (i) data acquisition (ii) pre-processing and (iii) feature extraction, (iv) feature value representation, (v) feature selection, (vi) information extraction (vii) classification and (viii) discovered knowledge.

Please refer to Appendix J for additional visualisations generated for the text mining algorithm.

3.2.4 Data acquisition

We extracted all prostate biopsies performed for men aged >30 years between 1 January 2006 and 31 December 2016 that were referred to the NHLS for pathology evaluation in the Gauteng province, South Africa. Two data sets were extracted from the national laboratory data repository that houses LIS collated patient laboratory reports [133]. These data sets were: (i) narrative prostate biopsy reports and (ii) the associated SNOMED CT code(s). Data were provided as comma-separated values (CSV) extracts from the national laboratory data repository. The narrative prostate biopsy reports are captured as free-text reports in the LIS and stored in the national laboratory data repository as a single variable. As described in Chapter 2, we used the SNOMED CT M and T codes to identify biopsies with an adenocarcinoma histological finding (n=8201) [154]. Once the biopsies with PCa were identified (adenocarcinoma histological findings with a reported GS), we extracted a random sample of 1000 biopsies using Microsoft Excel (Redmond, Washington, USA) [155].

To evaluate the text mining algorithm, we also randomly extracted 1000 prostate biopsy narrative reports with a PCa diagnosis that were submitted from private sector laboratories to the NCR. This data is generated using different LIS's, with the narrative biopsy report coded at the NCR by trained coded using ICD-03. We received only the narrative biopsy reports. This is referred to as the validation dataset.

For both datasets, the GS were manually coded by two experts. Following this, a random sample of 369 biopsies were independently verified to validate the manual coding.

3.2.5 Pre-processing

Text pre-processing is used to transform text into a computer readable format. This includes aspects such as the removal of punctuations, transforming all text to lower case and the removal of stopwords (e.g., 'the'). We used pre-processing to ensure that the narrative biopsy reports were in a machine-readable format. The first step was to convert the narrative reports to a document format (also referred to as corpus), which is defined as large and unstructured text. This is required to convert the narrative reports into a structured format that is required for text mining [152, 156, 157]. Next, the data cleaning process involved using the NLP tokenization, stopwords removal and stemming techniques [153, 157]. We used tokenization to condense the streams of text into smaller meaningful elements (called tokens) that comprised of words, phrases and symbols. We employed stemming to create various variants of words into a common representation known as the stem. Stemming takes words or a set of words to their root form, e.g. root of "gleasen" is "gleason". We also standardised the word Gleason, major, minor, score, etc. Finally, we used the NLTK toolkit stop words to filter and remove irrelevant words before text processing. This removed all possible English stopwords. We also converted text to lowercase for standardisation. Figure 3.2 depicts the visualisation of the corpus before (A) and after cleaning (B).

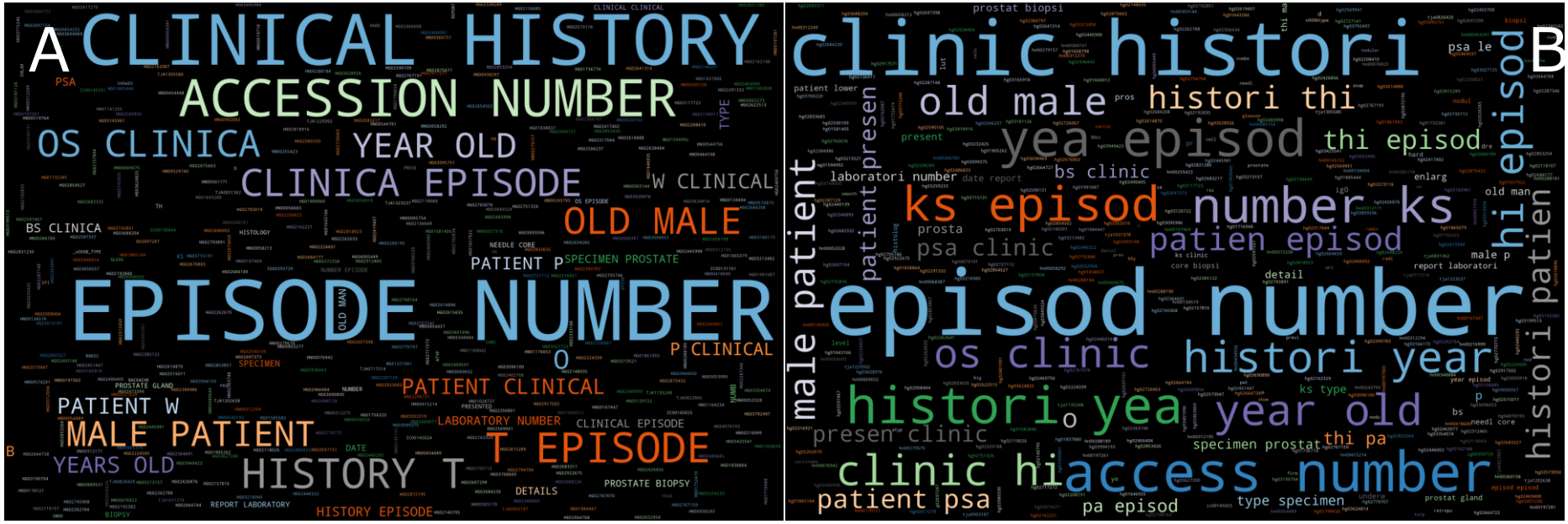


Figure 3.2: Diagram depicting the visualisation of the corpus before (A) and after cleaning (B). The larger the text the more important and frequent the term is in the narrative biopsy reports.

3.2.6 Feature extraction

Feature extraction is a process that identifies important features in the corpus. We used feature extraction to extract the features of interest from narrative prostate biopsy reports. We used regular expressions representative of the GS target feature such as “gleason”, “Gleason”, “GLEASON”, “Gleeson”, etc for feature extraction. Regular expressions can be used to define a sequence of characters that are associated with a feature. Each of these text patterns can be used as a rule-based approach to extract a feature. Similar approaches have been described by Napolitano and Spacic *et al* [149, 158].

Next, we used N-grams as our feature extraction strategy to extract the major and minor scores. We created unigrams, bigrams, trigrams and quadgrams which generated these scores. N-grams is a methodology that looks at sequences of words which are most occurring depending on the size of n. N-grams are a set of co-occurring terms that were reported in a sentence or paragraph in the corpus [159, 160]. For example, when n=1 (unigram) this represents single words in a sentence [160]. Similarly, when n is equal to 2 (bigram), 3 (trigram) or 4 (quadgram) this is represented as two, three and four words in a sentence respectively [160]. From the N-grams generated, we extracted the GS feature for each biopsy. The table below depicts the N-grams for a sample of biopsies.

Table 3.3: N-grams feature extraction output for a sample of biopsies.

['major 4 minor 5', '4+5']	['4 + 4', '4 + 4']	['3 + 3', '3 + 3']	['4 + 4', '4 + 4']	['major 4 + minor 3']	['4 + 5']
['major 4 minor 5', '4 + 5']	['major 4 minor 4']	['4+4', '4+4']	['4 + 4', '4 + 4']	['major 5 minor 4']	['major 3 minor 5']
['3+3', '3+3']	['2+2']	['3+4', '3+4']	['3+2', '3+3', '3+3']	['major 4 + minor 5']	['major 3 minor 4']
['3+5']	['3+3', '3+3']	['2 + 2', '2 + 2', '2 + 2']	['3 + 5', '3 + 5']	['4 + 3']	['3 + 5', 'major 4 + minor 5']
['major 4 minor 5']	['4+3', '4+3']	['3+2', '3+5', '3+5']	['3+3']	['major 5 minor 4']	['major 3 minor 5']
['4 + 3']	['2+3', '2+3']	['2 + 2']	['3+2', '3+2']	['major 5 + minor 4']	['major 5 + minor 4']
['major 4 minor 3']	['3 + 3', '3 + 3']	['4 + 4', '4+4']	['3+4', '3+4']	['major 3 minor 4']	['major 4 minor 5']
['3+2', '3+2']	['4 + 3', '4 + 3']	['2 + 3', '3 + 4', '3 + 4']	['4 + 3', '4 + 3']	['major 5 minor 5']	['major 3 minor 4']
['major 4 + minor 3']	['3 + 3', '3 + 3']	['4 + 4', '4 + 4']	['3+3', '3+3']	['major 5 minor 4']	['major 4 minor 5']
['5+5', '5+5']	['3 + 3', '3 + 3']	['5+4', '5+4']	['4 + 5']	['major 3 minor 3']	['major 5 minor 3']
['major 3 minor 4']	['major 5 + minor 5']	['major 5 minor 4']	['major 3 minor 5']	['major 4 minor 5']	

3.2.7 Feature value representation

The primary purpose of feature value representation was to represent words as feature vectors. This consists of a numerical representation of objects in order for the text mining algorithm to perform processing and statistical analysis. For feature representation, we

created a document term matrix using term frequency. This was used to transform the document into a numeric feature vector space. Term frequency refers to the number of times the term t appeared in the document d. The weight is denoted as $tf_{t,d}$. We reported the twenty most frequently occurring unigrams, bigrams, trigrams and quadgrams as horizontal bar graphs (Figure 3.3).

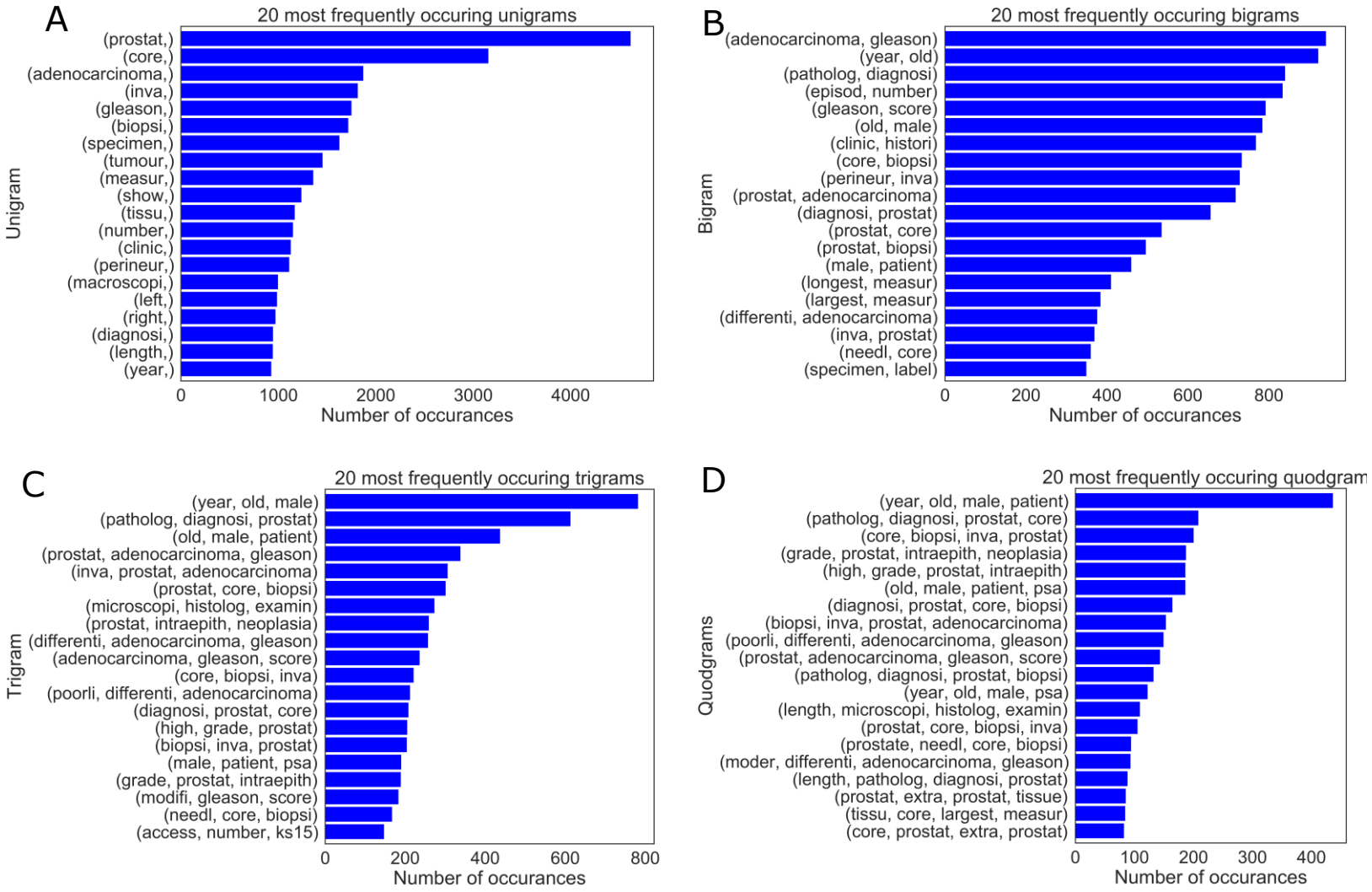


Figure 3.3: Horizontal bar graph depicting the top twenty occurring unigrams (A), bigrams (B), trigrams (C) and quadgrams (D). The number of occurrences is displayed on the x-axis.

3.2.8 Feature selection

Feature selection is the process used to either automatically or manually select those features of interest. Manual feature selection is usually expert driven. For feature selection, we used pathologists (experts) who identified the features of interest in the narrative prostate biopsy reports. As part of expert driven feature selection, we manually selected the following features: (i) episode number, (ii) major score, (iii) minor score, (iv) total score and (v) combined score. Because we used expert driven feature selection, we only choose relevant features and reduced the feature space (without using dimensionality reduction). Reducing the number of features selected improved the model performance. Even though the feature space was reduced, there was no loss of information [161].

3.2.9 Information extraction

Information extraction is used to select specific entities and relationships of interest [149]. For information extraction, we manipulated the N-grams output to extract the numerical value of the major and minor scores. Next, we calculated the total score and reported the GS in a standardised format, e.g. 4 + 4 =8.

3.2.10 Classification

We classified biopsies into the three risk categories using a rule-based approach that was implemented within the algorithm.

3.2.11 Discovered knowledge.

The discovered knowledge included the episode number, major score, minor score, total score, standardised GS and risk category. For each biopsy, the algorithm extracted a single row of structured data. From the narrative biopsy report depicted in Table 3.1, the following discovered knowledge was reported: (i) ABC1234, (ii) 4, (iii) 3, (iv) 7, (v) 4 + 3=7 and (vi) intermediate.

3.2.12 Text Mining Algorithm Evaluation

A confusion matrix was used to compare the text mining algorithm predicted against the manually coded values [162]. The confusion matrix consists of four values: (i) True Positives (TP): correctly predicting the GS, (ii) True Negatives (TN): correctly predicting a biopsy without a GS, (iii) False Positive (FP): falsely predicting a GS and (iv) False Negative (FN): falsely predicting the manually coded GS [162]. The precision and recall are calculated using these four values as follows: (i) $\frac{TP}{TP+FP}$ and (ii) $\frac{TP}{TP+FN}$ respectively. The F-score is the harmonic mean of precision and recall and is calculated using the formula $\frac{2*(Recall\ *Precision)}{(Recall+Precision)}$. The manually

coded values were assumed to be the gold standard. We reported the number of correctly predicted values.

3.2.13 Statistical analysis

We reported the top ten GS alpha, numeric and alphanumeric reporting formats as a table, i.e. how they were captured in the narrative prostate biopsy report. We also reported the top five GS reported, with the remaining scores reported as 'Others'. The percentage of a top five GS categorised as high-risk (≥8) is also indicated. As we reported data for a multi-class problem, we reported the frequencies for the predicted and manually coded values for a low, intermediate and high-risk GS as a table. Next, we calculated the macro averaged F-score (F-score for each GS risk category added up and then divided by the number of measurements) [163].

3.3 Results

The random sample taken from 1 000 prostate biopsies showed no manually coded GS misclassification errors for both datasets.

3.3.1 Text Mining Algorithm performance

For 1 000 narrative biopsies, the text mining algorithm extracted the GS in a time of under 10 minutes for both the study and validation datasets. The word cloud before and after cleaning revealed which text was more important. After using trigrams and quadgrams, the algorithm had both extracted all the GS and exhausted the sequence of words. Therefore, there was no need to use more than four grams, i.e., we had exhausted all word combinations. Our dataset was also small and logical extraction of n-Grams could only go up to four. With a larger corpus, we would have to explore using more n-Grams, e.g. 10. The term frequency analysis revealed that the Gleason score appeared as the fourth most common term for unigrams (n=1754). For the bigrams, the term Gleason score appeared in position one (n=942) and four (n=793). Similarly, the Gleason score appeared four times in trigrams compared to thrice for Quadgrams.

3.3.2 Text Mining precision and recall

The first text mining algorithm output reported an F-score of 0.99 (recall: 0.98 precision: 1.00) (Table 3.4). On manual inspection of the N-grams, we identified that two different GS were reported in both the clinical history and pathological diagnosis for 16 biopsies (diagnosis (example '3+2', '3+3', '3+3'). The algorithm was amended to report the latter GS resulting in an F-score of 1.00 (recall: 1.00 and precision: 1.00). The text mining algorithm was tested on the validation dataset and reported an F-score of 0.99.

Table 3.4: Performance of the text mining algorithm to automate the extraction of the Gleason score from narrative prostate biopsy narrative reports. A contingency table was used to compare the manually coded and algorithm predicted values. We reported the precision, recall and F-score reported for the first and updated text mining algorithm output as well as for the validation dataset.

First algorithm output		**Manual coding (Actual)** **Actual: Yes**
Predicted	**Actual: Yes**	984
	Actual: No	16
Precision = 1.00 Recall = 0.98 F-score = 0.99		
Updated algorithm output		**Manual coding (Actual)** **Actual: Yes**
Predicted	**Actual: Yes**	1000
	Actual: No	0
Precision = 1.00 Recall = 1.00 F-score = 1.00		
Validation dataset output		**Manual coding (Actual)** **Actual: Yes**
Predicted	**Actual: Yes**	988
	Actual: No	12
Precision = 1.00 Recall = 0.988 F-score = 0.99		

3.3.3 Gleason score formats reported

We identified ten different GS alpha, numeric and alphanumeric reporting formats (Table 3.5). The variations in reporting included: (i) use of both the equal sign as well as the word equals, (ii) use of brackets, (iii) spelling major and minor, (iv) use of both the word plus and the additional symbol and (v) use of colons and commas to separate major and minor scores.

Table 3.5: Different Gleason score formats reported. The clean extracted score reported, and the original value reported in the prostate biopsy report is indicated.

#	Extracted score	As reported in the biopsy report
1	5 + 4 =9	5, 4
2	5 + 4 =9	5 PLUS 4 EQUALS 9
3	3 + 3 =6	3 + 3 =6 OR 3 + 3
4	3 + 5 =8	MAJOR PATTERN 3, MINOR PATTERN 5
5	4 + 3 =7	MAJOR PATTERN: 4/5 MINOR PATTERN: 3/5
6	4 + 3 =7	MAJOR 4 PLUS MINOR 3 EQUALS 7
7	5 + 3 =8	SCORE 8 (MAJOR 5; MINOR 3)
8	3 + 4 =7	7 (3 + 4)
9	4 + 3 =7	(4 + 3) =7
10	3 + 4 =7	3 (MAJOR) + 4 (MINOR) =7/10

3.3.4 Gleason score frequency analysis

The most commonly reported GS were 5 + 4 =9 and 3 + 3 =6 GS for 17.6% (n=176) and 17.5% (n=175) of biopsies respectively (Table 3.6). There were 164 biopsies with a 4 + 3 =7 score (16.4%). A 3 + 4 =7 and 4 + 4 =8 GS was reported for 14.7% (n=147) and 14.2% (n=142) biopsies respectively. The remaining GS comprised 19.4% (n=196) of biopsies. A high-risk GS was reported for 31.8% of biopsies.

Table 3.6: Biopsy volumes for the five most commonly reported Gleason scores. The table reported the Gleason score and frequency for the top five reported scores with the remaining scores grouped and reported as "Others".

No.	Gleason score	n=	%
1	5 + 4 =9	176	17.6%
2	3 + 3 =6	175	17.5%
3	4 + 3 =7	164	16.4%
4	3 + 4 =7	147	14.7%
5	4 + 4 =8	142	14.2%
6	Others	196	19.6%
Total		**1000**	**100%**
High- Risk GS ≥8		**318**	**31.8%**

3.3.4 Gleason risk category analysis

For a low-risk GS, there were 199 predicted and 193 manually coded values (difference of 6), with an F-score of 0.98 (Table 3.7). Similarly, for an intermediate and high-risk GS a difference of 3 was reported for both groups with an F-score of 1.00 and 1.00 respectively. The macro-average F-score was 0.99 and macro recall and precision were 1.00 and 0.99 respectively.

Table 3.7: Comparison of low, intermediate and high-risk Gleason scores for the predicted and manually coded values. The macro-average F-score is reported.

GS Risk Category	Predicted	Manually Coded	F-score
Low-risk GS (1-6)	199	193	0.98
Intermediate-risk GS (7)	311	314	1.00
High-risk GS (≥8)	490	493	1.00
p-value[&]			0.9439
Macro-average F-score			0.99
Macro recall			1.00
Macro precision			0.98

3.3.5 Validation analysis

For the validation dataset, the most commonly reported GS were: (i) 3 + 3 =6 (37.7%), (ii) 3 + 4 =7 (19.4%), (iii) 4 + 3 =7 (14.9%) and (iv) 4 + 4 =8 (10.0%) and (v) 4 + 5 =9 (7.4%). A high-risk GS was reported for 17.4% of biopsies (Table 3.8).

Table 3.8: Biopsy volumes for the five most commonly reported Gleason scores for the validation dataset. The table reported the Gleason score and frequency for the top five reported scores with the remaining scores grouped and reported as "Others".

No.	Gleason score	n=	%
1	3 + 3 =6	377	37.7%
2	3 + 4 =7	194	19.4%
3	4 + 3 =7	149	14.9%
4	4 + 4 =8	100	10.0%
5	4 + 5 =9	74	7.4%
6	Others	106	10.6%
Total		**1000**	**100%**
High- Risk GS ≥8		**174**	**17.4%**

3.4 Discussion

The objective of our study was to explore the use of text mining techniques to extract the GS from irregularly reported text-intensive narrative prostate biopsy reports. The first text mining algorithm output reported that 16/1000 biopsies GS (1.6%) were inaccurately predicted. On inspection of the N-grams, we identified that these biopsies had two reported GS, once in clinical history and again in the biopsy report. We amended the text mining algorithm, resulting in all 1 000 GS accurately extracted with an F-score of 1.0. The attained F-score suggests that our feature engineering process was effective as we managed to pull out discriminative features that were most representative of our dataset. The text mining algorithm was further evaluated against a validation dataset, with good overall accuracy and precision. The F-score reported for both datasets is similar to a Perl routine that also used regular expressions to extract the GS [158]. Similar approaches using regular expressions have been reported by two other studies [149, 158].

Our findings reveal that despite the variability in the GS reporting, the text mining algorithm was able to extract the GS. This indicates that in settings with different AP reporting styles, the text mining algorithm would still be able to extract the required features. This is a promising finding that indicates that the text mining algorithm is able to handle varying reporting formats.

We noted a difference in the top five reported GS for our study and the validation dataset. We reported a high-risk GS for 31.8% of biopsies compared to 17.4% for the validation dataset. This indicates that late presentation differed between the public and private sector. All the data

reported for our study was for the public sector, with the majority being Black African. The racial variation in medical aid coverage, with only 10.5% of Black Africans having insurance, could explain the difference in the percentage of high-risk biopsies [164]. However, due to the small sample sizes for our study, these findings may differ for a larger study.

As we reported data for a multi-class problem and compared the predicted and manually coded values categorised as low, intermediate and high-risk [52]. The analysis revealed an acceptable macro-averaged F-score indicating that the text mining algorithm was able to accurately classify the GS risk category. This is a promising finding, indicating that the algorithm would be able to assess late presentation. With minor modifications to the algorithm, it would also be possible to report the updated GG [26]. This would make it possible for the text mining algorithm to report both the GS and GG routinely.

Our findings indicate that the text mining algorithm could be used to reliably extract the GS from laboratory data in similar settings where the narrative data is stored in a LIS. Given the paucity of local PCa data, this algorithm would make it easier to conduct studies for larger sample sizes. This would be achieved by implementing the text mining algorithm as an application programming interface (API) [165]. The text mining algorithm code can be packaged as an executable application that can be applied to routinely extract data from narrative laboratory reports. Such an approach could be used to facilitate the generation of important predictive clinical information for PCa using any LIS based data to derive both retrospective and prospective health information. This would dramatically improve the availability of the GS data for local studies and routine surveillance.

The text mining tools employed in our study could be used to extract clinical information for other cancers of public health interest. For example, breast cancer biopsies are graded using the modified Bloom and Richardson system [166]. This grading system is similar to the GS as it reports the cytology, tubule formation, nuclear pleomorphism, and the mitotic count to determine the grade (I, II or III) [166]. Therefore, the techniques employed in our study could be applied to other narrative laboratory data such as immunophenotyping, fluorescence in situ hybridization (FISH) and leukaemia reports, to extract important clinical, diagnostic and predictive information.

Furthermore, to address the remaining 12% of biopsies without a SNOMED CT code, our text mining algorithm could be supplemented by ML to extract an adenocarcinoma histological finding. When combined with the developed text mining algorithm, it would be possible to

assign a PCa diagnosis from narrative biopsy reports in an automated fashion. This has the potential to offer near real-time to cancer registry type information removing the need for manual coding. This would also dramatically reduce the time from reporting to generating surveillance data. In addition, the extraction of the GS would make it possible to better assess trends in late presentation.

In addition to ML approaches, we would also recommend using deep learning approaches [167]. Deep learning is composed of multiple processing layers that learn representations of data with multiple levels of abstraction [168]. This approach has dramatically improved AI approaches for visual object recognition and object detection [168]. Deep learning models are able to extract information from large datasets and will continue to improve the knowledge discovery as more data is generated [169]. This enables deep learning to outperform classical ML approaches [169]. One of the benefits is that deep learning is able to extract the feature without the need for supervision required by ML. A good example is representation learning, a deep learning approach that automatically discovers the representations needed for detection or classification from raw data [168].

Once these ML and deep learning algorithms have been developed, it would be possible to move the extraction of an adenocarcinoma histological finding with the GS to a cloud service. This would make possible for narrative prostate datasets to be uploaded using an internet connection and the extracted knowledge delivered as a data extract. Similar approaches have been demonstrated for breast cancer [170]. This has the potential for cancer registries across Africa to load their narrative data and obtain coded data for incidence and late presentation surveillance activities.

3.5 Conclusion

Our study has shown that a text mining algorithm can be used to extract the predictive GS from narrative biopsy reports. This could be used to better assess late presentation by extracting the GS in an automated fashion. These tools have the potential to describe PCa in an African context with a paucity of data. The potential exists to automate GS extraction for national laboratory data using an API. This approach is applicable to other cancers of public health interest. Furthermore, ML could be used to reduce the delays in the publication of cancer registry data. In addition, deep learning approaches should be investigated. These algorithms could be moved to a cloud service to extend automated PCa surveillance data generation across Africa.

CHAPTER 4 - RETROSPECTIVE ANALYSIS OF PROSTATE-SPECIFIC ANTIGEN TESTING TRENDS AND PRIMARY HEALTH CARE PUBLIC-SECTOR GUIDELINES IN THE GAUTENG PROVINCE, SOUTH AFRICA, BETWEEN 2006 AND 2016

For uniformity of layout of this thesis, this chapter is laid out in the style of the thesis. This paper is currently under review at the BMJ Open. The proof of review is provided in Appendix C.

Abstract

Objectives:

The objective of this study was to use laboratory data to describe PSA testing trends for Primary Health Care (PHC) services from a single province.

Design:

Retrospective descriptive study design.

Methods:

We analysed national laboratory data between 2006 and 2016 for men ≥30 years in the Gauteng Province. We used the probabilistic matching algorithm to create the first-ever PSA cohort. We used hot-deck imputation to assign missing race group values and the district health information system (DHIS) descriptors to include clinics and community health centres (CHC), and reported patient numbers by calendar year, age category and race group as well as descriptive statistics. We also analysed patient number by age category for each race group. We used logistic regression to assess any association for race group and age with a PSA ≥4µg/L.

Results:

Between 2006 and 2016, numbers of men tested increased from 1 782 to 67 025 respectively, with 186 984/239 506 (78.1%) tests from clinics. The majority of testing was for Black African men in the 50 - 59 age category. The majority of testing in the 40-49 was for Black Africans (25.6%). In comparison, White men were tested predominantly in age categories 60-69 (29.9%) and ≥70 years 18.1%). We reported a median of 0.9µg/L that increased with age. A PSA ≥4µg/L was reported for 11.7% of men, increasing to of whom 35.5% in those aged 70 or greater. The logistic regression reported that the odds of having a PSA ≥4µg/L was significantly lower for Indian/Asians, Coloureds and Whites than for Black Africans.

Conclusions: Our study shows a marked increase in PSA testing from PHC facilities, suggestive of screening for PCa. The approaches developed can be used to describe national PSA testing data.

Strengths and limitations of this study

This study analysed laboratory data for a first-ever PSA across a province for public-sector PHC facilities serving 80% of the population. Such a large sample size dramatically improves the generalisability and validity of the study to enable an inference for the population. Most local studies reported data for one or more health facilities in a urology clinic with small sample sizes.
Our study identified that patients presented with a first-ever median PSA well within the 4µg/L cut-off with up to 20% of low-value PSA testing.
Our study demonstrated the value of laboratory data to describe PSA testing for PHC facilities across a province. It will be feasible to implement these approaches for national data in other settings with a laboratory information system.
Due to the absence of a national unique patient identifier, this study used the probabilistic matching algorithm to de-duplicate patients. De-duplication may be sub-optimal for patients with incomplete demographic information.
The paucity of race group data was a limitation for our study.

4.1 Introduction

PCa ranks as the second most frequent neoplasm and the fifth leading cause of cancer death in men globally [22]. PCa was the most common male neoplasm in South Africa in 2012 [77]. The 2018 global burden of cancer report indicated that PCa incidence is highest in developed countries such as Australia, New Zealand and North America [22, 171]. PCa was reported to be the leading cause of death among black men particularly in sub-Saharan Africa and the Caribbean [22]. Globally, it is estimated that there will be 1.3 million new cases of PCa and about 359 000 deaths in 2018 [22]. A meta-analysis of PCa incidence rates across Africa by Adeloye *et al* reported an estimated pooled PCa incidence rate of 22.0 (95% CI: 19.93 – 23.97) per 100 000 population [83]. This study reported that the incidence rate is higher in South Africa with reports of rates ranging from of 47.7 to 67.8 per 100 000 population [83]. In 2014, the local cancer registry reported a PCa incidence of 43.7 per 100 000 population, which is lower than estimates provided by Adeloye *et al* [7, 83]. This incidence is expected to increase with a concomitant improvement in life expectancy.

The high incidence in African populations may be due to a genetic predisposition to PCa specifically for race group and family history [26, 41]. The preponderance of PCa is much higher among men of African descent who also have advanced stage of disease at presentation [39]. A disproportionate burden of PCa has been reported for men of African descent in the United States of America, United Kingdom and the Caribbean [42, 172]. Rebbeck *et al* reported that the overall weighted prevalence of PCa was 19.9% in men of Asian descent, 26.7% in men of European descent and 26.2% in men of African descent [173]. Despite adjusting for PSA testing, men of African descent from these regions still present with advanced disease at diagnosis [174].

Several local studies conducted in urology settings have evaluated PSA levels, specifically for PCa cases at hospitals and have shown that African men present late to hospitals with very high PSA levels. PSA is a glycoprotein with a close structural relationship to the glandular kallikreins, functioning as a serine proteinase [175]. PSA levels may be elevated in BPH as well as in PCa [26]. Despite these limitations, PSA testing has improved the early detection of PCa for many countries [74]. Schroder *et al* reported that PSA-based screening reduced PCa mortality by 20% but was associated with a high risk of over- diagnosis [74]. This requires that HCW carefully identify patient cohorts that will benefit most from individualised early diagnosis, taking into account the potential benefits and harms involved [26]. However, as an independent variable, PSA is a strong predictor of PCa [26]. PSA may also be used to describe

risk e.g. the EAU guidelines defines risk groups for the biochemical recurrence of localised and locally advanced PCa with a PSA ≥20µg/L classified as high risk [26].

A local study which assessed PSA testing for Black Africans showed that cases presented with advanced disease as evidenced by a median PSA of 98.8µg/L compared to 9.1ug/L for controls [3]. Similar findings have been reported by other investigators. For example, a study from the KwaZulu-Natal Province reported that Black African men (n=81) diagnosed with PCa have a median PSA of 154µg/L (IQR 39 - 448) at presentation [2]. A similar study in the Western Cape Province (n=901) for patients diagnosed with PCa reported that the mean PSA was significantly higher for Black Africans (766.1µg/L) compared to 673.3µg/L and 196.1µg/L for Coloureds and Whites respectively [1]. On the other hand, men presenting at PHC facilities have been shown to have much lower PSA levels [85].

South Africa does not have a national screening policy and there is a paucity of data on PSA testing trends from PHC services, which act as a gatekeeper to higher levels of care [176]. These services are offered by clinics and community health centres (CHC) [177]. Clinics offer a range of basic services for 8 hours compared to CHC which in addition offer 24 hour maternity, accident and emergency services [177]. The objective of this study was to describe first-ever PSA testing trends for PHC services in the Gauteng Province. These findings will contribute important information for a country with high rates of PCa incidence and mortality.

4.2 Methods

4.2.1 Study design

We used a retrospective descriptive study design with convenience sampling. Our inclusion criteria included tests requested by PHC facilities and a patient age ≥30 years in the Gauteng Province, South Africa.

4.2.3 Ethical considerations

We obtained ethical clearance from the University of the Witwatersrand (M170419). Our study did not use any patient identifiers.

4.2.4 Data extract

Data were extracted for the 2006 to 2016 period from national laboratory data, provided as password-protected data files.

4.2.4 Data preparation

The PSA data extract included the following variables: (i) episode number, (ii) age, (iii) race group, (iv) facility description, (v) reviewed date of the PSA result, (vi) unique patient identifier (for the patient), (viii) numeric PSA result, and (ix) non-numeric PSA result. The data was extracted from the CDW. The CDW houses all laboratory data for public-sector testing in South Africa.

Data were prepared using Microsoft Access and Excel (Redmond, WA, USA) and analysed using SAS 9.4 (Cary, NC, USA) [129, 155]. Age, gender and race group are captured from details provided on the laboratory request form by HCW. We classified data into the following age categories: (i) 30 - 39, (ii) 40 - 49, (iii) 50 - 59, (iv) 60 - 69, (v) ≥70 and (vi) unknown.

We converted PSA values both below and above the analytical range reported as text values to numbers, e.g. <0.01 and > 100 reported as 0.01 and 100.0 µg/L respectively [175]. We used the district health information system (DHIS) organisational hierarchy to identify testing by PHC facilities, which are the first point of contact between the population and the health system [176].

Race group was identified as an important PCa risk factor with higher incidence and late presentation reported specifically for Black Africans [1-3]. For our study, race group was poorly recorded (94.5% missing), we opted to use a locally developed hot-deck imputation method to assign missing data [178]. Well populated cancer registry data with patient-reported race grouping was used to construct the imputation reference panel [179]. Andridge *et al* described hot-deck imputation as a way of handling missing data that is replaced with an observed response from a “similar” unit [180]. Hot deck imputation has been applied in epidemiologic and medical settings [181]. Tang *et al* reported the various imputation methods to address missing survey responses [182]. This study reported that the hot-deck imputation method showed small bias and good coverage for all variables [182]. In an unpublished local study, Chen *et al* used a hold-out test to compare imputed *versus* cancer registry patient-reported race group values for 406 642 unique surname-race group pairings [179]. The historically well-populated cancer registry data was split into patient-reported race group values and imputed values [179]. This study compared the imputed and original values and reported a 94.3% agreement [179]. We used the Statistics South Africa (STATS SA) 2011 Census reported race groups: (i) Indian/Asian, (ii) Black African, (iii) Coloured and (iv) White [88].

The inclusion criteria for the study was testing requested by PHC facilities for men 30 years and older. All PSA testing requested by hospitals was excluded. For figures 4.4 and 4.5 as well as Table 4.2, data for an unknown age and race were excluded.

We used the unique patient identifier that is generated by a probabilistic matching algorithm developed by the CDW that includes fuzzy logic matching. This algorithm uses a set of demographic variables to de-duplicate patients. We used this algorithm to generate a de-duplicated PSA presentation (first-ever result) cohort. This cohort consists of patients who presented for care or were screened for PCa, using only the earliest recorded PSA result. Throughout this chapter, any reference to patients is for men. Bassett *et al* have shown that using the CDW unique patient identifier, they were able to identify a cohort of HIV patients transferred from a hospital to PHC facilities with 90% accuracy [183].

4.2.5 Statistical Methods

We reported the number of men that had a first-ever PSA per year, reporting the percentage year on year increase. As lower test volumes were reported in 2012 and 2013, the 2014 and 2015 calculations were based on 2011 numbers. We are not able to determine what caused the lower test volumes for 2012 and 2013. Annual test volumes were reported for clinics and CHC, indicating the annual PSA median and 75th percentile. We reported the number of patients tested by year for both age category and race group, with the fold increase reported between 2006 and 2016. The inset reported total volumes, percentage contribution as well as the fold increase. We reported descriptive statistics for age category and race group as well as the PSA median and interquartile range (IQR). We performed the Cochran-Armitage test to assess/test for PSA level trend across the years 2004-2016. We reported what percentage of men were in each age category for the four race groups. We reported the proportion of men with a PSA result in three risk categories stipulated by urological guidelines as well as for a PSA ≥4µg/L [52]. We used multiple logistic regression to determine the risk factors of a PSA ≥4µg/L controlling for age and race group. With PSA testing not recommended below 40 years of age, we excluded the 30 - 39 age category for this analysis. We reported the odds ratio (OR), p-values, and a 95% confidence interval (CI). A p-value of less than 0.05 was used to indicate significant associations between the categorical variables of interest.

4.3 Results

The data extract included 580 714 tests, of which 277 983 both met the inclusion criteria and were requested by PHC facilities (Figure 4.1). Following de-duplication, there were 239 506 (86.2%) men with a first-ever PSA. There were 186 984 men that tested at clinics compared to 52 522 for CHC. Patient age reported a skewness of 0.22 with values provided for 98% of

men. There was a paucity of race group data, with values provided for only 1.7% of samples. We populated all race values using hot-deck imputation, with 7.1% classified as unknown.

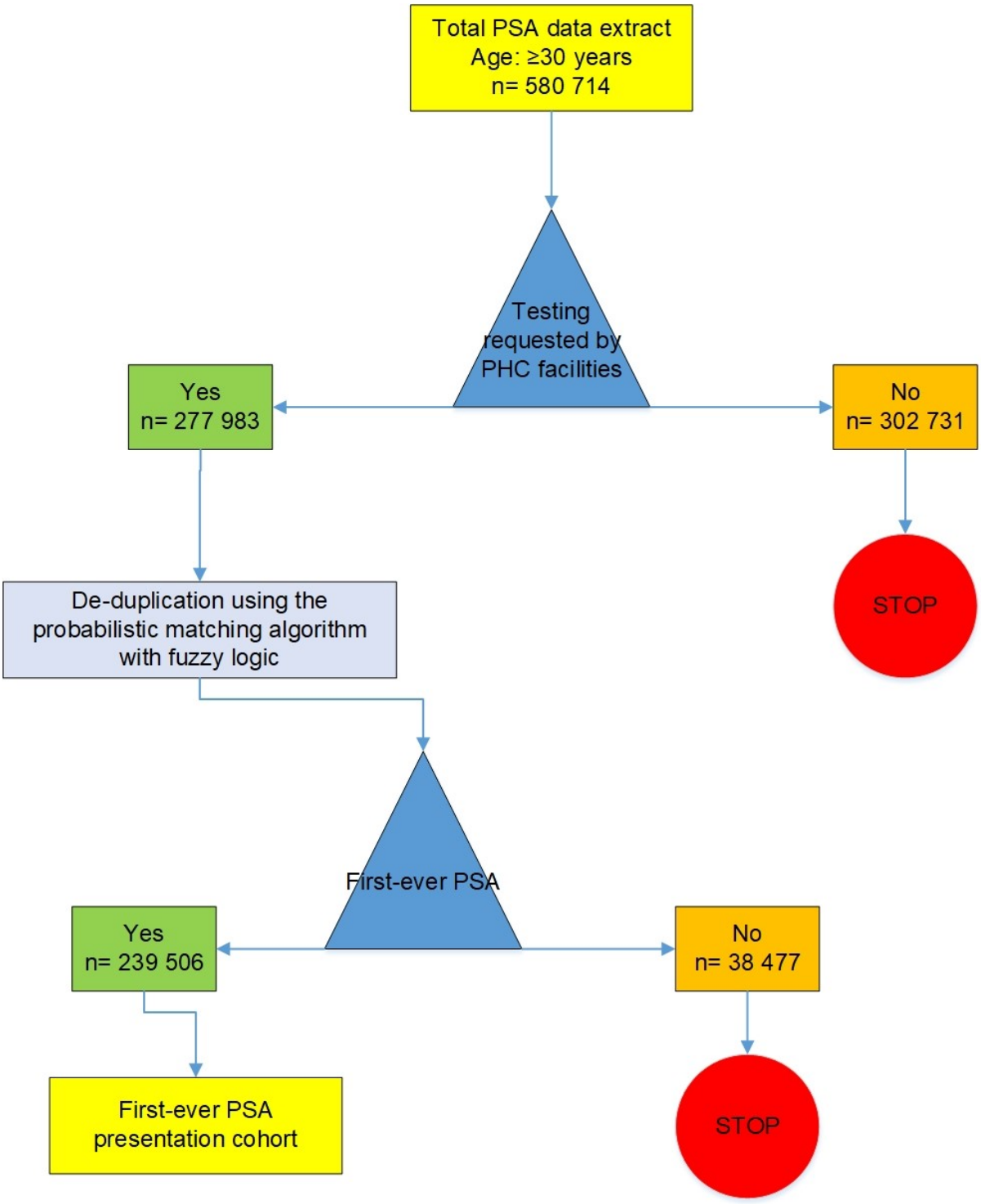


Figure 4.1: Flow chart depicting all the data steps to generate the data used to create the first-ever PSA cohort for PHC services between 2006 and 2016 in the Gauteng Province, South Africa

4.3.1 Number of patients receiving first-ever PSA test

The annual number of men receiving a first-ever PSA increased from 1 782 in 2006 to 67 025 by 2016. The percentage year on year change ranged from -5% to 145%. An unexplained decrease was noted for 2012 and 2013, with a percentage year on year change of -5% and -25% respectively (Figure 4.2). Patient numbers increased from 24 220 in 2014 to 67 025 by 2016.

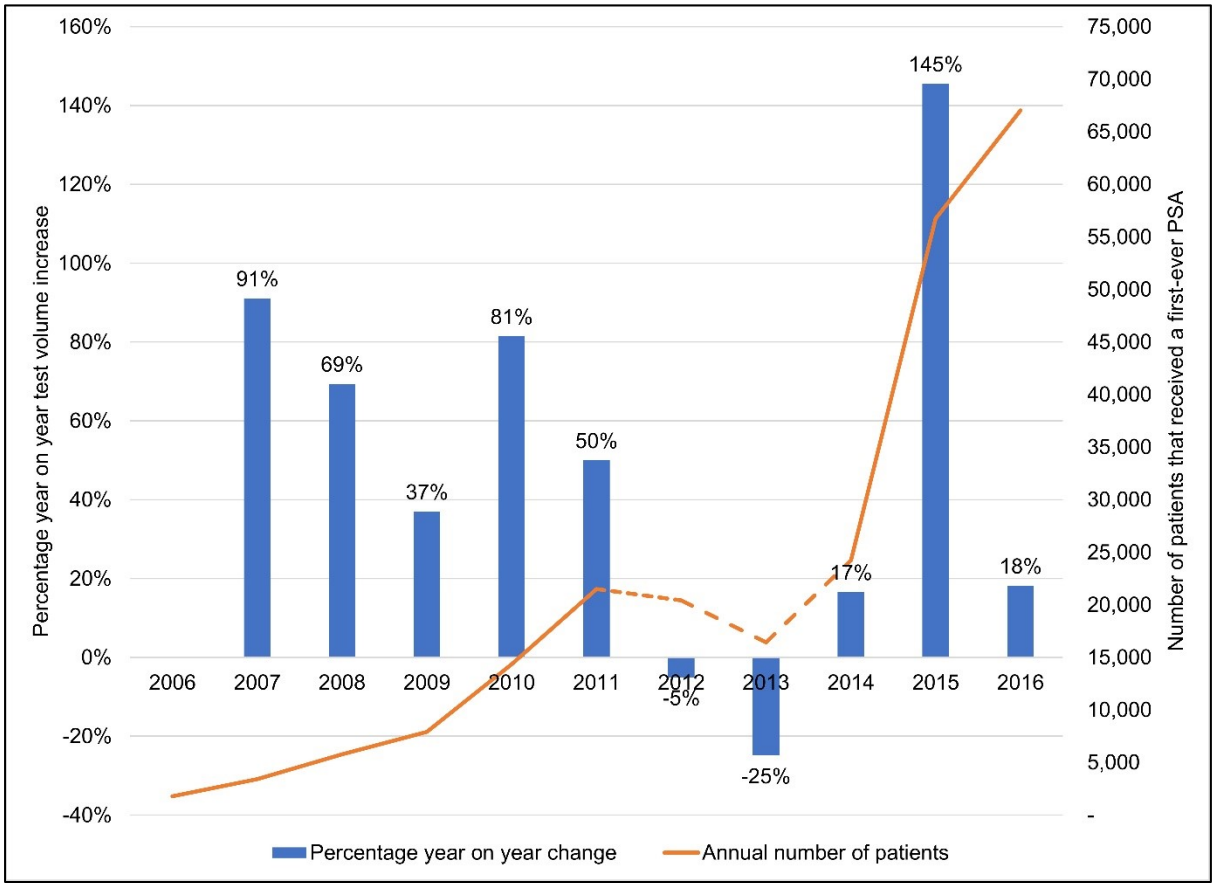


Figure 4.2: Percentage year on year change reported as a bar chart for patients with a first-ever total PSA test at PHC services in the Gauteng Province, South Africa. The annual test volumes were reported as a line chart. As lower test volumes were reported in 2012 and 2013 (indicated by the dotted line), the 2014 and 2015 percentage year on year change calculations were based on the 2011 numbers.

In 2006, the number of patients tested were similar for both clinics (n=924: 51.9%) and CHC (n=858: 48.1%). Between 2007 and 2016, clinic patient numbers increased from 2 316 (68.0%) to 53 939 (80.5%). In contrast, CHC patient numbers increased from 1 089 (32.0%) to 13 086 (18.7%) (Figure 4.3). Between 2014 and 2015, clinic test volumes increased dramatically from 19 172 to 46 144 (2.4 fold), respectively. Similarly, CHC test volumes increased from 5 048 in 2014 to 10 588 by 2015 (2.1 fold).

The annual median PSA decreased from 1.1 in 2006 to 0.9µg/L by 2016 (p-trend<0.001). Similarly, the 75th percentile PSA decreased from 2.6µg/L in 2006 to 1.7µg/L by 2016.

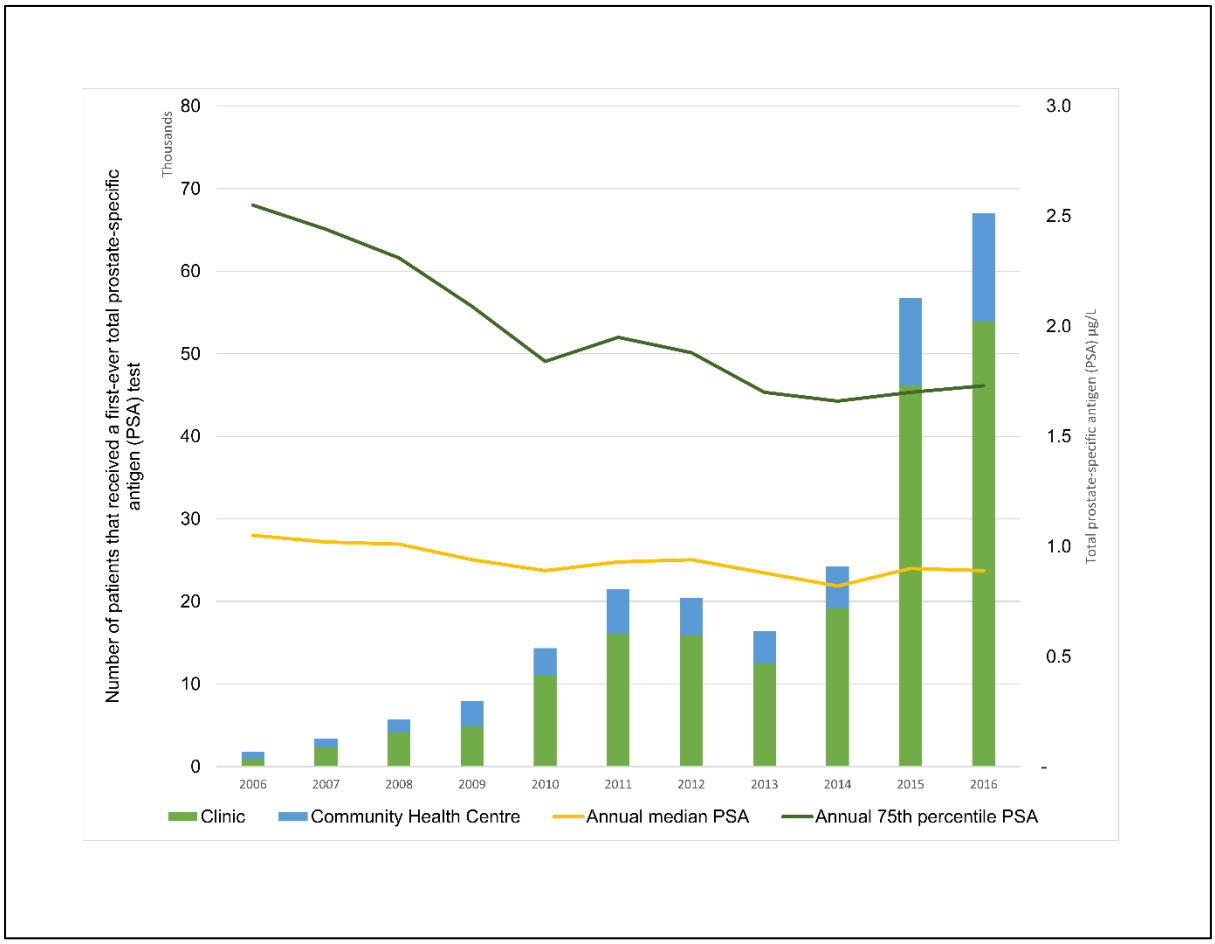
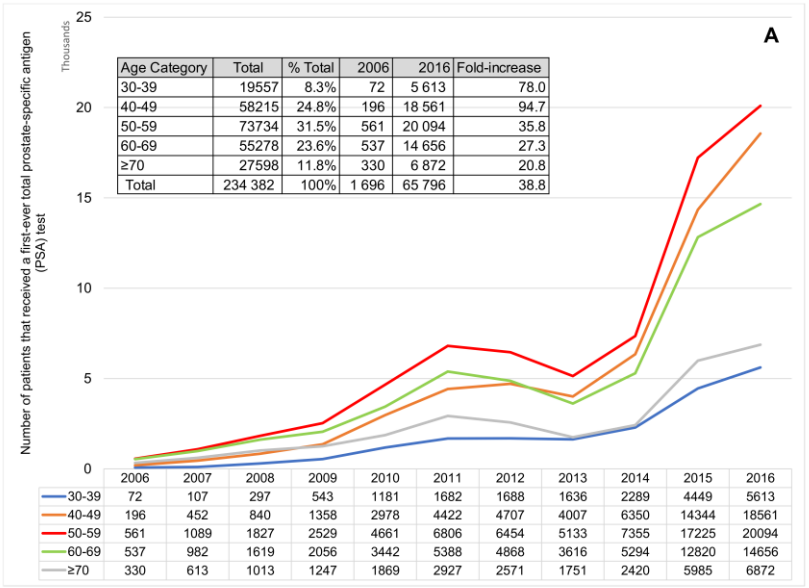


Figure 4.3: Annual PSA test volumes by unit type for PHC services. The annual PSA volumes are reported for clinics (blue bars) and community health centres (CHC) (dark grey bars) between 2006 and 2016 in the Gauteng Province, South Africa.

4.3.2 Number of patients receiving a first-ever PSA test by age category and race group

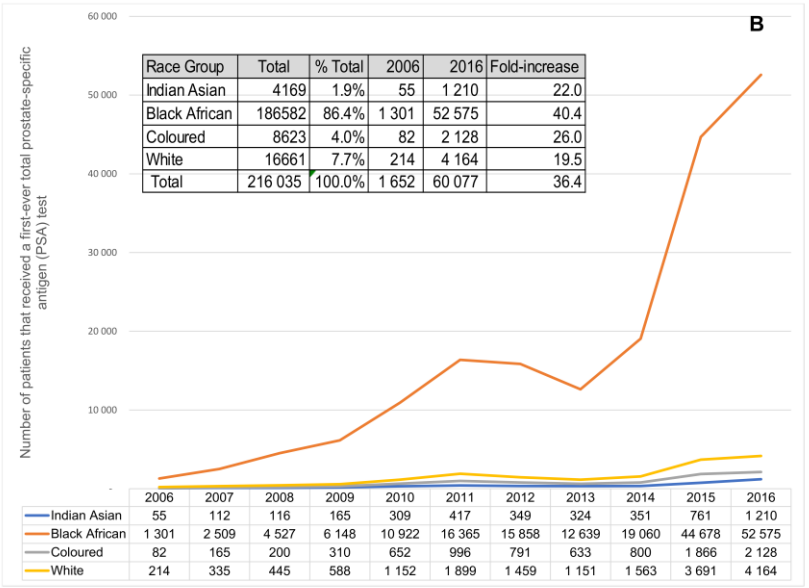
Across all eleven years, the majority of testing was performed for men aged 50 to 59 (31.5%), followed by the 40 - 49 (24.8%) and 60 - 69 (23.6%) age categories (Figure 4.4A). We noted a 94.7-fold increase in patient numbers between 2006 and 2016 for the 40 - 49 age category (from 196 to 18 561). The 30 - 39 age category reported a 78.0-fold-increase, with patient numbers increasing from 72 in 2006 to 5 613 by 2016. For the 50 - 59 and 60 - 69 age categories, we noted a 35.8- and 27.3-fold increase, with patient numbers increasing from 561 to 20 094 and 537 to 14 656 respectively. The ≥70 age category reported a 20.8-fold increase.

We reported a 40.4-fold rise in patient numbers for Black Africans between 2006 and 2016 increasing from 1 301 in 2006 to 52 575 by 2016 (Figure 4.4B). All other racial groups showed increased test volumes as follows: 26.0, 22.0- and 19.5-fold increase for Coloureds, Indian/Asians, and Whites respectively.



Age Category	Total	% Total	2006	2016	Fold-increase
30-39	19557	8.3%	72	5 613	78.0
40-49	58215	24.8%	196	18 561	94.7
50-59	73734	31.5%	561	20 094	35.8
60-69	55278	23.6%	537	14 656	27.3
≥70	27598	11.8%	330	6 872	20.8
Total	234 382	100%	1 696	65 796	38.8

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
30-39	72	107	297	543	1181	1682	1688	1636	2289	4449	5613
40-49	196	452	840	1358	2978	4422	4707	4007	6350	14344	18561
50-59	561	1089	1827	2529	4661	6806	6454	5133	7355	17225	20094
60-69	537	982	1619	2056	3442	5388	4868	3616	5294	12820	14656
≥70	330	613	1013	1247	1869	2927	2571	1751	2420	5985	6872

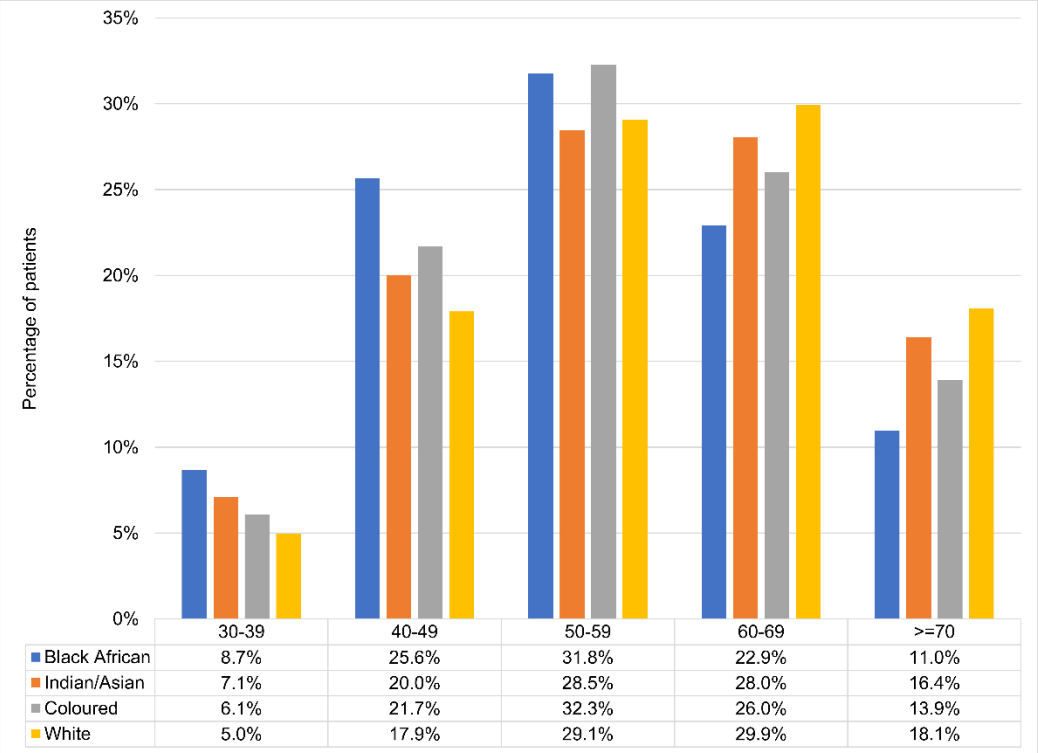


Race Group	Total	% Total	2006	2016	Fold-increase
Indian Asian	4169	1.9%	55	1 210	22.0
Black African	186582	86.4%	1 301	52 575	40.4
Coloured	8623	4.0%	82	2 128	26.0
White	16661	7.7%	214	4 164	19.5
Total	216 035	100.0%	1 652	60 077	36.4

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Indian Asian	55	112	116	165	309	417	349	324	351	761	1 210
Black African	1 301	2 509	4 527	6 148	10 922	16 365	15 858	12 639	19 060	44 678	52 575
Coloured	82	165	200	310	652	996	791	633	800	1 866	2 128
White	214	335	445	588	1 152	1 899	1 459	1 151	1 563	3 691	4 164

Figure 4.4: Line chart reporting the number of patients with a first-ever total prostate-specific antigen (PSA) test by year and age category (A) and race group (B) for PHC facilities between 2006 and 2016 in the Gauteng Province, South Africa

For age categories 30-39 and 40-49, 8.7% and 25.6% of men were Black African (Figure 4.5). For the 50-59 age category, a range of 28.5% to 32.3% was reported for Indian/Asians and Coloureds. More White men were tested in the 60-69 and ≥70 years, contributing 29.9% and 18.1% respectively.



	30-39	40-49	50-59	60-69	>=70
Black African	8.7%	25.6%	31.8%	22.9%	11.0%
Indian/Asian	7.1%	20.0%	28.5%	28.0%	16.4%
Coloured	6.1%	21.7%	32.3%	26.0%	13.9%
White	5.0%	17.9%	29.1%	29.9%	18.1%

Figure 4.5: Bar chart reporting the number of patients with a first-ever total prostate-specific antigen (PSA) test by race group and age category for PHC facilities between 2006 and 2016 in the Gauteng Province, South Africa.

4.3.3 Descriptive statistics by age and race group

Our study reported a median PSA of 0.9µg/L (IQR: 0.5 - 1.8). Median PSA increased with age group from 0.7µg/L (IQR 0.4 - 1.0µg/L) for the 30 - 39 age category to 2.3µg/L (IQR 1.0 - 6.4µg/L) for men seventy years and older. The median PSA ranged from 0.9µg/L (IQR 0.5 - 1.8) for Black Africans to 1.0µg/L (IQR 0.5 - 2.0) for Whites. Our findings indicate that we classified 2.3% and 2.7% of samples as intermediate and high-risk, respectively (Table 4.1).

A high-risk PSA result (≥20µg/L) was reported for 0.1%, 0.2%, 1.2%, 4.2% and 10.5% of men for age categories 30 - 39, 40 - 49, 50 - 59, 60 - 69 and ≥70 respectively. For Black African

men, 11.6% reported a PSA ≥4µg/L compared to 12.6% for Coloureds. A high-risk PSA result was reported for 1.6% of Black Africans compared to 2.7%, 2.4% and 1.6% for Coloureds, Whites and Indian/Asians respectively.

Table 4.1: Descriptive statistics for a first-ever total prostate-specific antigen (PSA) test reported for age category and race group between 2006 and 2016 in the Gauteng Province, South Africa. The overall median and interquartile range (IQR) as well as the proportion of men with a with a first-ever total PSA test ≥4, <10 (low-risk), 10 - 19.9 (intermediate-risk) and ≥20µg/L (high-risk) are reported.

Category	Median (IQR)	% ≥4µg/L	% <10µg/L (Low risk)	% 10 - 19.9µg/L (Intermediate risk)	% ≥20µg/L (High-risk)
Overall	**0.9 (0.5 - 1.8)**	**11.7%**	**227 629 (95.0%)**	**5 439 (2.3%)**	**6 438 (2.7%)**
Age Category					
30 - 39 (n=; %)	0.7 (0.4 - 1.0)	1.7%	19 481 (99.6%)	50 (0.3%)	26 (0.1%)
40 - 49	0.7 (0.4 - 1.1)	2.3%	57 891 (99.4%)	195 (0.3%)	129 (0.2%)
50 - 59	0.8 (0.5 - 1.5)	7.5%	71 842 (97.4%)	1 009 (1.4%)	883 (1.2%)
60 - 69	1.3 (0.7 - 2.9)	18.5%	51 002 (92.3%)	1 946 (3.5%)	2 330 (4.2%)
≥70	2.3 (1.0 - 6.4)	35.5%	22 644 (82.0%)	2 070 (7.5%)	2 885 (10.5%)
Unknown	1.0 (0.6 - 2.1)	14.8%	4 769 (93.1%)	169 (3.3%)	186 (3.6%)
Race Group					
Indian/Asian	10.7%	10.7%	4 016 (96.3%)	86 (2.1%)	67 (1.6%)
Black African	11.6%	11.6%	177 049 (94.9%)	4 271 (2.3%)	5 262 (2.8%)
Coloured	12.6%	12.6%	8 183 (94.9%)	208 (2.4%)	232 (2.7%)
White	1.0 (0.5 - 2.0)	12.5%	15 965 (95.8%)	380 (2.3%)	316 (1.9%)
Unknown	0.9 (0.5 - 1.8)	11.2%	22 416 95.5%	494 (2.1%)	561 (2.4%)

4.3.4 Association between an elevated PSA with age and race group

The logistic regression reported that the risk of a PSA ≥4µg/L for Indian/Asians, Coloureds and Whites were significantly lower than that of Black Africans (Table 4.2). We reported an OR of 0.70 (CI: 0.63 - 0.78) for Indian/Asians compared to an OR of 0.94 (CI 0.88 - 1.02) and 0.77 (CI 0.74 - 0.82) for Coloureds and Whites respectively. We reported a significant association for Indian/Asians and Whites (p-value <0.0001) but not for Coloureds. These results indicate that Indian/Asians, Coloureds and Whites are 30%, 6% and 23% less likely to have a PSA ≥4µg/L.

The ≥70 age category was determined to have an increased risk of a PSA ≥4µg/L with an OR of 23.90 (CI 22.45 - 25.45) when compared to the 40 - 49 age category. The 50 - 59 and 60 -

69 age categories reported an OR of 3.41 (CI 3.20 - 3.64) and 9.73 (CI 9.15 - 10.35), respectively.

Table 4.2: Logistic regression to assess the association between samples with a first-ever total PSA ≥ 4µg/L and race group as well as age category between 2006 and 2016 in the Gauteng Province, South Africa. We controlled for race group and age category.

Characteristic	Odds ratio	95% CI	p-value
Race Group			
Black African	1		
Indian/Asian	0.70	0.63 - 0.78	<0.0001
Coloured	0.94	0.88 - 1.02	0.129
White	0.77	0.74 - 0.82	<0.0001
Age Category			
40 - 49	1		
50 - 59	3.41	3.20 - 3.64	<0.0001
60 - 69	9.73	9.15 - 10.35	<0.0001
70+	23.90	22.45 - 25.45	<0.0001

For the logistic regression, the Odds Ratio [OR], 95% confidence interval (CI) and p-value were reported. An alpha of 0.05 was assumed.

4.4 Discussion

The purpose of this study was to describe first-ever PSA testing trends for PHC services in the Gauteng Province between 2006 and 2016. We reported a gradual increase in patients receiving a first-ever PSA test between 2006 and 2013. Between 2014 and 2015, a striking increase in patient numbers was reported. Furthermore, we noted an increase in testing across all age groups with 33% of tests conducted in those below the age of 50 years, and almost 12% in those above 70. The median PSA for a first-ever test decreased over the study period.

Whilst South Africa does not have a screening policy, the fact that there was a huge increase in patients being tested at PHC facilities suggests that some the testing may have been for screening purposes, which is further supported by a low median PSA very similar for PCa screening for men without PCa reported by Capitanio *et al* [184]. An unexplained decrease was noted for 2012 and 2013 followed by a dramatic increase in 2015. We are not sure what could explain the lower test volumes in 2012 and 2013. Possible explanations for the lower test volumes could include industrial action, fewer men presenting for testing or other factors we are not aware of. Our analysis for CDW extracts for HbA1c, creatinine, glucose and

cholesterol show similar decreases in test volumes for the same period (data not shown). This suggests that there were broader systems issues in the health care system that resulted in lower volumes for multiple tests. The increase between 2006 and 2011 could be due to the inclusion of the PSA test in the essential test list (ETL), public-sector campaigns, introduction of clinicians at PHC facilities or other factors we are not aware of. However, the dramatic increase in 2015 is most likely due to the introduction of PHC public-sector guidelines in 2014. For many years, it was debatable as to whether PSA screening did, in fact, reduce PCa mortality. The European Randomised Study of Screening for Prostate Cancer (ERSPC), clearly showed a reduction in PCa mortality with screening [74, 185]. On the other hand, screening can result in harm such as detection of cancers which otherwise may never have been diagnosed during the life of the patient, with subsequent over-treatment [74].

As described in Chapter 5, there was a substantial increase in prostate biopsies between 2006 and 2016 (1.7 fold increase). In addition, the ASIR increased from 44.92 per 100 000 population in 2006 to 57.31 by 2016. This is indicative of more men with an elevated PSA being referred to a urologist and diagnosed with PCa. Despite the increase in prostate biopsies, we did not see a corresponding increase in more men presenting at PHC facilities with an elevated PSA.

Various clinical guidelines for PCa assessment have been released in South Africa. Public sector guidelines introduced in 2014 for PHC services state that PCa occurs in men >50 years and is most often asymptomatic [57]. These guidelines for the first time recommended PSA testing for men over 50 years of age [57]. At the same time, guidelines issued by the Urology Society of South Africa in 2013 recommended targeted screening. They recommend testing in men with a life expectancy of >10 years, and suggest race and age based cut off values for PSA, with testing for Black African men from the age of 40 years [52]. These are not dissimilar to European guidelines which recommend an individualised risk-adapted strategy for early detection of PCa, to a well-informed men with a life-expectancy of at least 10 - 15 years [26].

The development of conflicting guidelines is seen across the globe and most likely arise as a result of three large prospective randomised controlled trials (RCTs) that published data on PCa screening [72-74]. Andriole *et al* compared annual PCa screening *versus* the standard of the care group at 10 American study centres [72]. Following 7 to 10 years of follow-up, the PCa mortality rate was very low and did not differ significantly between the two study groups [72]. Another study randomised men in a 1:1 ratio, to either a screening group invited for PSA testing every two years compared to a control group that were not screened [73]. This study

showed that PCa mortality was reduced almost by half over 14 years in the intervention arm with the risk of over-diagnosis. To complicate matters, a Cochrane review reported that PCa screening is not likely to result in a significant reduction in cancer-specific and overall mortality [75]. In the United States, a definite reduction in PSA screening was noted following the release of the 2012 United States Preventive Services Task Force (USPSTF) recommendations against PSA screening in all men [186, 187]. This was followed by a decline in prostate biopsy and PCa incidence with a trend to higher grade and stage of cancer at diagnosis [188]. It is important to note that none of these studies were conducted in Africa and hardly any African men were included. It is, therefore, necessary to conduct African studies that will enable the development of appropriate targeted screening or testing algorithms to identify those that would benefit the most.

Our finding of significant increases in testing for the age categories 30 - 39 and ≥70 can be described as low value testing, which has been defined as use of a health service whose costs or harms exceed its benefits [189]. Our findings reveal that predominantly Black African men aged 30 to 49 were offered PSA testing. In contrast, predominantly older White men 60 years and older presented for testing. This appears to indicate that PSA testing is benefitting the high-risk racial group in the Gauteng province. Earlier testing for Black African men is in line with local urological guidelines [52]. Low value care has been shown to be a major driver of wasteful health care expenditure [190]. Testing of elderly men without a history of PCa is not unique to our setting. It has been reported in the United States where the Choosing Wisely Campaign was initiated to reduce low value care [191].

PCa is the leading cause of cancer death among men particularly in SSA [22]. In spite of the increase in PSA testing across all ages, our data from the same laboratory shows that over the same period the PCa age-standardised incidence rate has increased from 44.9 (2006) to 57.3 (2016) per 100 000 population [133]. More Black African men were diagnosed with PCa than men from other racial groups and more Black African men presented with high-risk PCa compared to men from other racial groups [133]. While we were not able to assess mortality due to PCa, it is of concern that Black men in particular present with more advanced disease. There is a clear need for education about when and who to test for PSA and how follow-up should be carried out. Clinical guidelines provide an efficient mechanism to introduce evidence-based clinical practices to offer a standard package of services to the entire population serviced by public health facilities [192]. There are multiple HIV and TB examples that demonstrate that national guidelines are able to extend a package of services to the most

remote communities [193, 194]. It is similarly necessary to provide clear guidelines for PSA testing in our setting.

The findings of this current study of a low median PSA, differ substantially from local urological hospital studies [1-3]. These studies offered PSA testing to men with a clinical suspicion of PCa, which could explain why they reported much higher median PSA values [1-3]. Furthermore, this study's data suggest that in spite of increased testing for PSA, early PCa is not detected. The reasons for this are not yet known and should be investigated.

We have emphasised the value of laboratory data to describe PSA testing for PHC facilities across a province. It will be feasible to implement these approaches for national data across South Africa. We could use this data to link patients who require follow-up to a urology service. Such an approach has been used effectively locally for HIV [195]. Laboratory data provide a ready resource to generate real-time PCa screening data to target health care interventions to reduce PCa incidence. In order to enrich PCa screening, the DRE results captured electronically will also be needed.

CHAPTER 5 - PROSTATE CANCER AGE-STANDARDISED INCIDENCE INCREASE BETWEEN 2006 AND 2016 IN THE GAUTENG PROVINCE, SOUTH AFRICA: A LABORATORY DATA-BASED ANALYSIS

For uniformity of layout of this thesis, this chapter is laid out in the style of the thesis. This paper has been accepted for publication in the South African Medical Journal. The proof of acceptance is provided in Appendix D.

Abstract

Background:

PCa is the leading male neoplasm in South Africa and is the second most frequently diagnosed cancer among men globally. ASIR vary by up to 189-fold globally, with an ASIR of 68.0 per 100 000 in 2018 in South Africa.

Objectives:

The aim of our study was to describe PCa in the Gauteng Province of South Africa among men undergoing prostate biopsies.

Methods:

This was a retrospective descriptive study using prostate biopsy data collected from the NHLS database between 2006 and 2016. The SNOMED CT M and T codes were extracted to assign histological findings using the ICD-O-3. PCa was defined as adenocarcinoma with a reported GS. The new GG based on the GS is defined as follows: (i) GG 1 for a GS ≤6, (ii) GG 2 for a GS of 3 + 4 =7, (iii) GG 3 for 4 + 3 =7, (iv) GG 4 for a GS of 8 and (v) GG 5 for GS ≥9. Higher-grade disease was defined as a GG ≥4 (GS ≥8) in line with local guidelines. We reported associations of PCa with a GG ≥4 with age and race. We used the number of incidence cases, provincial and world standard population data to determine annual ASIR.

Results:

We identified 22 937 biopsies referred to NHLS between 2006 and 2016. Of the 6 448 (39%) biopsies with a PCa finding for Black Africans, 46% were diagnosed with high-risk PCa compared to 36 - 40% for other race groups (p <0.0001). Black Africans were more likely to have PCa with a GG ≥4 than Whites (Odds Ratio =1.45: 95% CI: 1.27 - 1.67). The ASIR increased from 44.9 (2006) to 57.3 (2016) per 100 000 population.

Conclusion:

Black African men are significantly more likely to present with PCa with a GG ≥4 compared to the other racial groups in the Gauteng Province. The ASIR increased dramatically during the

study period, perhaps due to increased screening and awareness. There is a need for additional research to better understand why Black African men present with higher grade disease.

5.1 Introduction

PCa is a major public health concern as life expectancy is increasing globally [36, 37]. In 2018, global PCa incidence was the second highest after lung cancer [38]. PCa incidence varied by up to 189 fold across 186 countries, with ASIR ranges recorded varying widely from 1.0 in Bhutan to as high as 189.1 per 100 000 population in Guadeloupe in 2018 [22]. Comparatively higher ASIR were reported for some less developed regions such as Barbados (129.3), Bahamas (85.8), South Africa (68.0) and Reunion (63.7) per 100 000 population respectively [22]. PCa ASIR is highest amongst men of African descent, reflecting racial and genetic predisposition in these countries [196]. This could also be due to varying PCa screening and testing patterns that affect incidence [40].

Limited local data reported that Black Africans present with more aggressive PCa with 36% reporting a GS ≥7 compared to 17% for African Americans [3]. The GS is the strongest prognostic factor for PCa treatment [26]. Local guidelines categorise PCa risk using the GS as follows: (i) 2 - 6 low, (ii) 7 intermediate and (iii) 8 - 10 high [47, 52]. The new GG is based on the GS as follows (i) GG 1 for a GS ≤6, (ii) GG 2 for a GS of 3 + 4 =7, (iii) GG 3 for 4 + 3 =7, (iv) GG 4 for a GS =8 and (v) GG 5 for GS ≥9 [47, 49]. Limited PCa biopsy data, however, exist for Black African men, with most local studies reporting data for selected health facilities [1-3]. Currently, aside from local cancer registry data, which reported national 2016 data by 2020, there is no comprehensive public sector PCa data for the Gauteng Province [117].

In South Africa, PCa is the leading neoplasm in men [22, 38]. Risk factors include age, race (Black African) and family history [30, 36, 47, 52]. The NHLS provides diagnostic services to 80% of the public [197]. Routinely accumulated laboratory data through analysis of diagnostic samples offers a unique opportunity to describe PCa histological characteristics with the related demographic detail in local populations. The SNOMED CT T and M codes captured routinely based on the histological findings could be used to represent relevant clinical information consistently and comprehensively as an essential part of producing electronic health information [46]. The T code identifies anatomic terms for each organ and the M code describes the microscopic changes observed [198].

We chose the Gauteng Province for this study as it represents a typical dense urban area that comprises 24.1% of the national population [199]. It is the economic hub of the country with both the highest numbers of international and domestic migrants, the latter originating from provinces such as Limpopo, KwaZulu-Natal and Eastern Cape [199, 200]. These domestic

migrants are aged between 15 and 64 years, with males outnumbering females by 111 to 100 [201]. Over 70% of migrants are Black African [201].

The aim of our study was to describe PCa in the Gauteng Province of South Africa among men undergoing prostate biopsies using data fields directly extracted from patient laboratory records.

5.2 Methods

5.2.1 Study design

We extracted all occurrence of prostate biopsies between (01/01/2006) and (31/12/2016) for men over the age of 30 years referred to the NHLS for pathology evaluation in the Gauteng Province. Two datasets were extracted from the national laboratory data repository which houses the LIS collated patient laboratory reports [197]. These datasets were, (i) narrative prostate biopsy data and (ii) SNOMED CT code(s). Ethical clearance was obtained from the University of the Witwatersrand (Certificate number M170419).

5.2.2 Inclusion criteria

We included all prostate biopsies for men aged 30 years and older. We excluded biopsies identified as post mortem.

5.2.3 Biopsy data preparation

The narrative biopsy extract included the following variables: (i) episode number, (ii) unique patient identifier, (iii) age, (iv) gender, (v) race group, (vi) facility name, (vii) reviewed date and (viii) biopsy narrative report. The episode number identifies samples for all laboratory processes. The unique patient identifier was generated using the organisation probabilistic matching algorithm that includes fuzzy logic matching minimizing the duplication of PCa case reports [202]. For duplicate records, only the first record was used to calculate ASIR. A sequence number was generated for chronologically sorted patient data. The GS was manually coded, with random samples independently verified to validate manual coding (n= 369).

The SNOMED CT data extract included the episode number, M and T codes. The M code describes the microscopic changes noted in cells, tissues, and organs [198]. The T code describes the anatomical site of origin (prostate) [198]. We used the SNOMED CT codes to develop lookup tables to report: (i) histological findings (PCa/Not PCa), (ii) biopsy findings, (iii) tumour type and (iv) organ with the assistance of an experienced anatomical pathologist and

urologist. We used the ICD-O-3 to assign PCa findings, i.e. adenocarcinoma tumour type with the associated tumour grading (GS) [116]. For missing M codes, we manually coded biopsies using the narrative result text.

Race group was identified as an important PCa risk factor [1-3]. As race group was poorly recorded (94.5% missing), we used a locally developed hot-deck imputation method to assign missing data [178] described in an unpublished study [179]. Well populated cancer registry data with patient-reported race grouping was used to construct the imputation reference panel [179]. The race groups reported by STATS SA in the 2011 Census are Indian/Asian, Black African, Coloured, and White [88].

We used a relational database to combine the respective tables using a left outer join with the narrative biopsy data extract as the master table [126]. Data were analysed using SAS 9.4 (Cary, NC, USA) and Stata 12 (College Station, TX, USA)[203, 204].

Descriptive statistics by race group included: (i) mean age (with range), (ii) biopsy numbers, (iii) biopsy findings and (iv) number of biopsies reporting a PCa histological finding. We compared the mean age across race groups by one-way analysis of variance (ANOVA). We used the Chi-square test (χ) to identify whether there was a statistical difference for age category and biopsy findings for the four race groups. An alpha of 0.05 was used for all association tests.

We used a pyramid chart to report the number of biopsies with a PCa and non-PCa finding for age categories (excluding an unknown age). Age was reported on the X-axis and percentage of biopsies on the Y-axis. Annual biopsy numbers were reported as a bar chart.

Multiple logistic regression method was used to determine the factors associated with PCa with a GG ≥4 controlling for age and race group. This is in line with local guidelines for a GS ≥8 classified as high-risk [47, 52]. We reported odds ratio, p-values, and confidence intervals. Biopsies with a PCa finding were analysed by GS and GG risk group for the four race groups reported as a stacked bar chart [47, 52]. We used a stacked bar chart to report the percentage of biopsies per year for the GS and GG risk categories, with GG 4 and GG 5 reported together. The ASIR per 100 000 population and annual average percentage change (AAPC) were calculated for defined age categories for each calendar year using the Joinpoint trend analysis software [205]. Data variables for the model included: (i) gender, (ii) year of diagnosis, (iii) age categories/groups, (iv) number of incident cases (count variable), (v) population estimates and

(vi) Segi world standard population (WSP) [205-207]. The latter was used to enable comparison with cancer registry reports [206].

5.3 Results

There were 25 010 biopsies extracted from the LIS. The SNOMED CT codes were used to assign 88% of histological findings, with 10% manually coded [154]. We excluded 285 biopsies that did not meet the inclusion criteria and excluded a further 2 073 (8.3%) non-prostatic biopsies. Thus, the total sample size was 22 937. It was not possible to determine a GS for 455/8 390 biopsies (5.4%) with a PCa finding. We randomly reviewed GS for accuracy and noted no misclassification errors.

5.3.1 Prostate biopsy outcomes by age and race group

The mean age by race group was 66.3 (range: 31 - 92) for Indian/Asians, 67.4 (range: 30 - 99) in Black Africans, and 67.7 (range: 30 - 99) for Whites (Table 5.1). A statistically significant difference for mean age for the four race groups was determined by one-way ANOVA [p= <0.001]. There was a significant difference in biopsy frequency by age category across the four race groups (p-value <0.0001). A malignant biopsy finding was reported for 46% of Black Africans compared to 39%, 37% and 30% for Coloureds, Whites and Indian/Asians respectively. For Black Africans, 39% of biopsies reported a PCa finding, compared to 23%, 29% and 31% for Indian/Asians, Whites and Coloureds respectively (p-value <0.0001).

Table 5.1: Prostate biopsy descriptive statistics by race group. Prostate biopsy descriptive data for each race group reported as a table for the Gauteng Province between 2006 and 2016. Mean age (range), biopsy numbers and percentages reported for age category, race group, biopsy results and prostate cancer results. One-way analysis of variance (ANOVA) reported for age as a continuous variable and Chi-square test used to identify whether there was a statistical difference for age category and biopsy findings for the four race groups.

	Indian/Asian	Black African	Coloured	White
Number of biopsy samples	487	16 491	1 129	3 422
Mean Age (Range)	66.3 (31 - 92)	67.4 (30 - 99)	67.4 (35 - 99)	67.7 (30 - 99)
p-value[&]				<0.0.001
Age Category				
30 - 39(n, %)	4 (1%)	49 (0%)	2 (0%)	7 (0%)
40 - 49	13 (3%)	298 (2%)	22 (2%)	44 (1%)
50 - 59	76 (16%)	2418 (15%)	160 (14%)	436 (13%)
60+	380 (78%)	12502 (76%)	890 (79%)	2830 (83%)
Unknown	14 (3%)	1224 (7%)	55 (5%)	105 (3%)
p-value[$]				<0.0001
Biopsy findings				
Atypia/Dysplasia	21 (4%)	575 (3%)	43 (4%)	129 (4%)
Benign/negative for malignancy	317 (65%)	8 321 (50%)	641 (57%)	2 031 (59%)
Neoplasm, malignant	147 (30%)	7 561 (46%)	442 (39%)	1 255 (37%)
Uncertain whether benign or malignant	2 (0.5%)	34 (0%)	3 (0%)	7 (0%)
p-value[$]				<0.0001
Prostate Cancer				
Prostate Cancer[#]	111 (23%)	6 448 (39%)	354 (31%)	1 009 (29%)

[&] One-way analysis of variance (ANOVA) [$]Chi-square test

[#] Adenocarcinoma histological finding with Gleason score reported

Twenty-three percent of biopsies from the 65 - 69 age category reported a PCa finding opposed to 19% and 20% for the 60 - 64 and 70 - 74 age categories respectively (Figure 5.1). The percentage of non-PCa biopsies were lowest in older ages with 12% reported in the 75 - 79 and 8% in 80+ age categories. Eleven and five percent of biopsies reported a PCa finding for the 50 - 54 and 55 - 59 age categories, respectively. A non-PCa finding was reported for 22% of biopsies in the 60 - 64 age category, decreasing to 19% and 11% for the 65 - 69 and 70 - 74 age categories respectively.

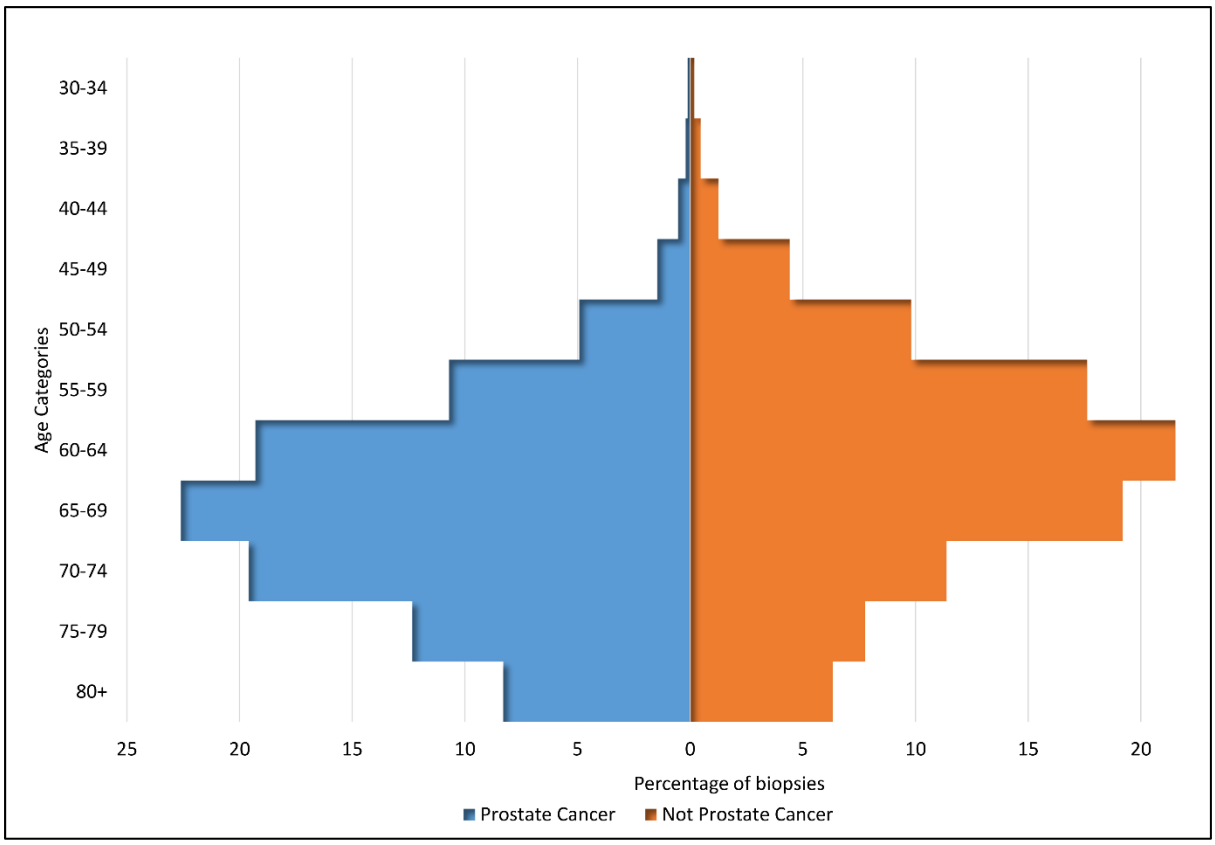


Figure 5.1: Age distribution for prostate cancer findings. Biopsy numbers reported by age category for both prostate cancer and non-prostate cancer histological findings in the Gauteng Province between 2006 and 2016 as a population pyramid.

Black Africans were more likely to have PCa with a GG ≥4 than Whites (Odds Ratio [OR] = 1.45, 95% CI: 1.27 - 1.67) (Table 5.2). Similarly, Coloureds and Indian/Asians were also more likely to have PCa with a GG ≥4 than Whites with an Odds Ratio of 1.14 (95% CI: 0.89 - 1.46) and 1.06 (95% CI: 0.71 - 1.59) respectively.

Men <70 years were determined to have a reduced probability of being diagnosed with PCa with a GG ≥4 than those older than 70 with OR = 0.99 (CI: 0.69 - 1.42) for ages 30 - 49, OR = 0.90 (95% CI: 0.79 - 1.03) for ages 50 - 59, and OR = 0.87 (CI: 0.79 - 0.96) for ages 60 - 69. Controlling for race group, males ≥70 years are at a significantly greater risk of having PCa with a GG ≥4.

Table 5.2: Multiple logistic regression to assess the association between prostate cancer outcomes and Gleason score and race group and age category. For the dependent variable, a grade group (GG) of 4 or 5 was coded as 1 (GS ≥8) and GG 1-3 as 0 for a prostate cancer histological finding with a Gleason score (GS) reported.

Characteristic	Odds ratio	95% CI	p-value
Race Group			
White	1		
Indian/Asian	1.06	0.71 - 1.59	0.764
Black African	1.45	1.27 - 1.67	<.0001
Coloured	1.14	0.89 - 1.46	0.311
Age Category			
30 - 49	0.99	0.69 - 1.42	0.948
50 - 59	0.90	0.79 - 1.03	0.131
60 - 69	0.87	0.79 - 0.96	0.005
70+	1		

For the logistic regression, the 95% confidence interval (CI), p-value and age- (for analysis of racial group) or race-adjusted (for analysis of age group) Odds Ratio [OR] are reported.

Between 2006 and 2016, the number of biopsies increased from 1 630 to 2 761 (Figure 5.2). We noted lower volumes in 2008 (n=1 168), decreasing from 1 610 in 2007 (27% decrease). A 25% increase in biopsies was reported between 2010 (n=1 868) and 2011 (n=2 335), compared to 19% between 2014 (n=2 357) and 2015 (n=2 805).

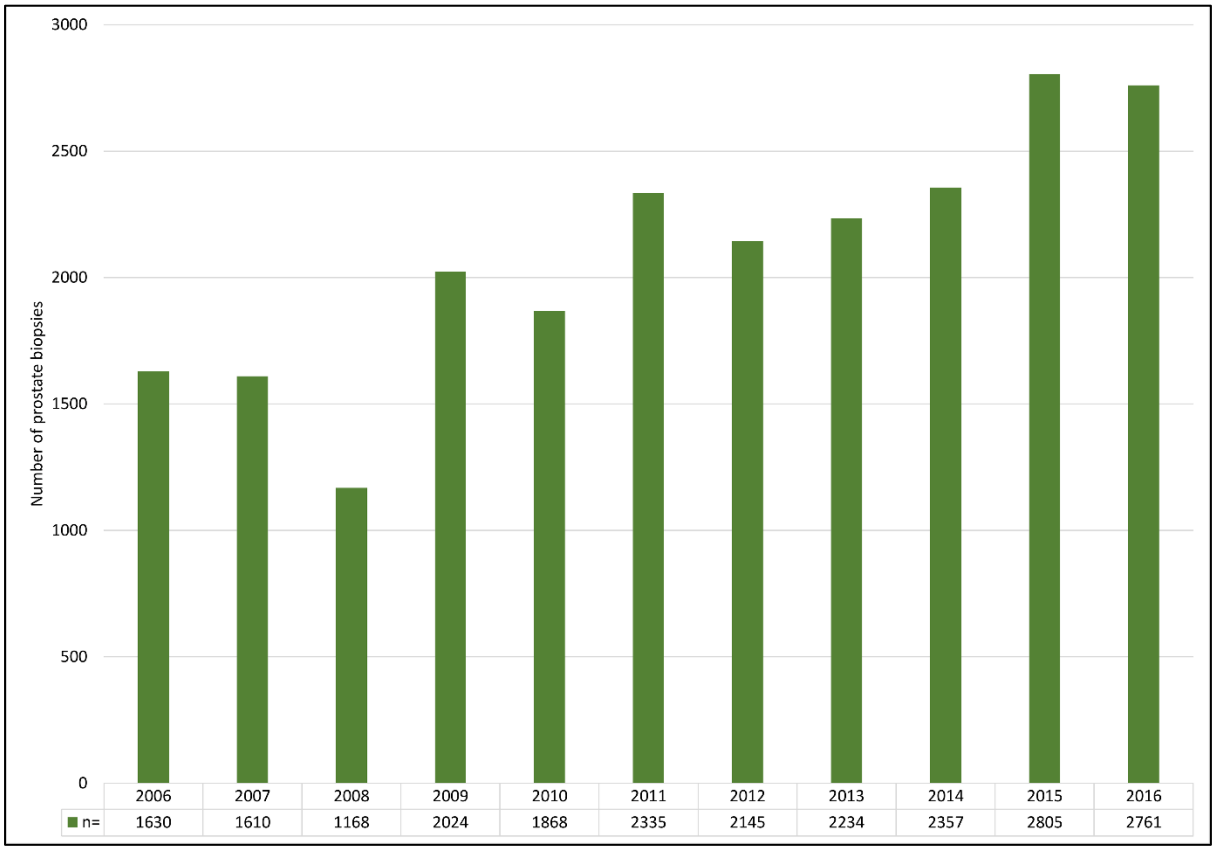


Figure 5.2: Line chart reporting prostate biopsy numbers for the Gauteng Province between 2006 and 2016. Stacked bar charts indicating annual prostate biopsy numbers. The calendar year reported on the X-axis and the number of prostate biopsies on the Y-axis.

5.3.2 Analysis of low, intermediate and high-risk Gleason scores

A high-risk GS (≥8) was reported for 46% (n=6 448) of biopsies with a PCa finding for Black Africans compared to 40%, 38%, and 36% for Coloureds, Indian/Asians and Whites, respectively (Figure 5.3). The percentage of biopsies with a GG 2 ranged from 17% for Whites to 24% for Coloureds. Similarly, the percentage of biopsies with a GG 3 ranged from 10% for Coloureds to 15% for Indian/Asians. There were 18% and 13% of biopsies for Black Africans that reported a GG 2 and GG 3 respectively.

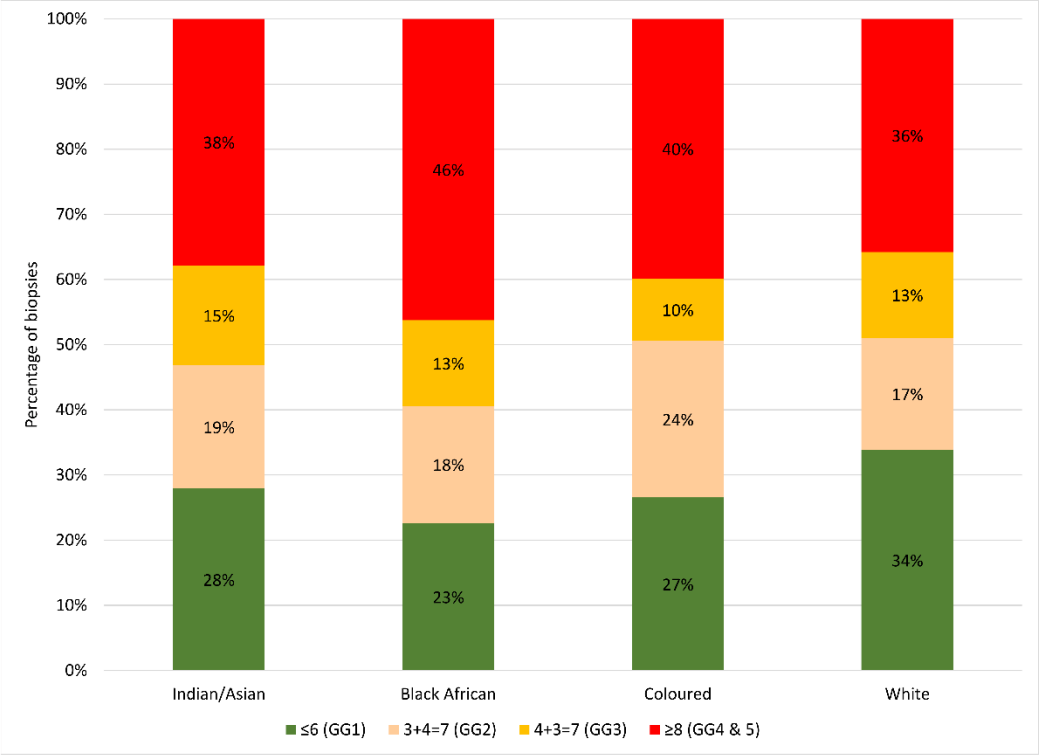


Figure 5.3: Percentage of prostate biopsies with a low (GS ≤6), intermediate (GS =7: 3 + 4 =7 and 4 + 3 =7), and high (GS ≥8) risk Gleason score by race group. Bar chart used to report on the percentage of prostate biopsies with prostate cancer finding by GS risk categories. The grade group (GG) is indicated in brackets, with 4 and 5 combined to report a GS ≥8. The percentages reported as data labels per race group.

Between 2006 and 2016, the percentage of biopsies with a GG 1 decreased from 27% to 18% respectively (Figure 5.4). Conversely, the percentage of biopsies with a GG ≥4 increased from 42% in 2006 to 48% by 2016. Biopsies with a GG 3 increased from 12% in 2006 to 27% by 2016. In contrast, the percentage of biopsies with a GG 2 decreased from 20% to 18% respectively.

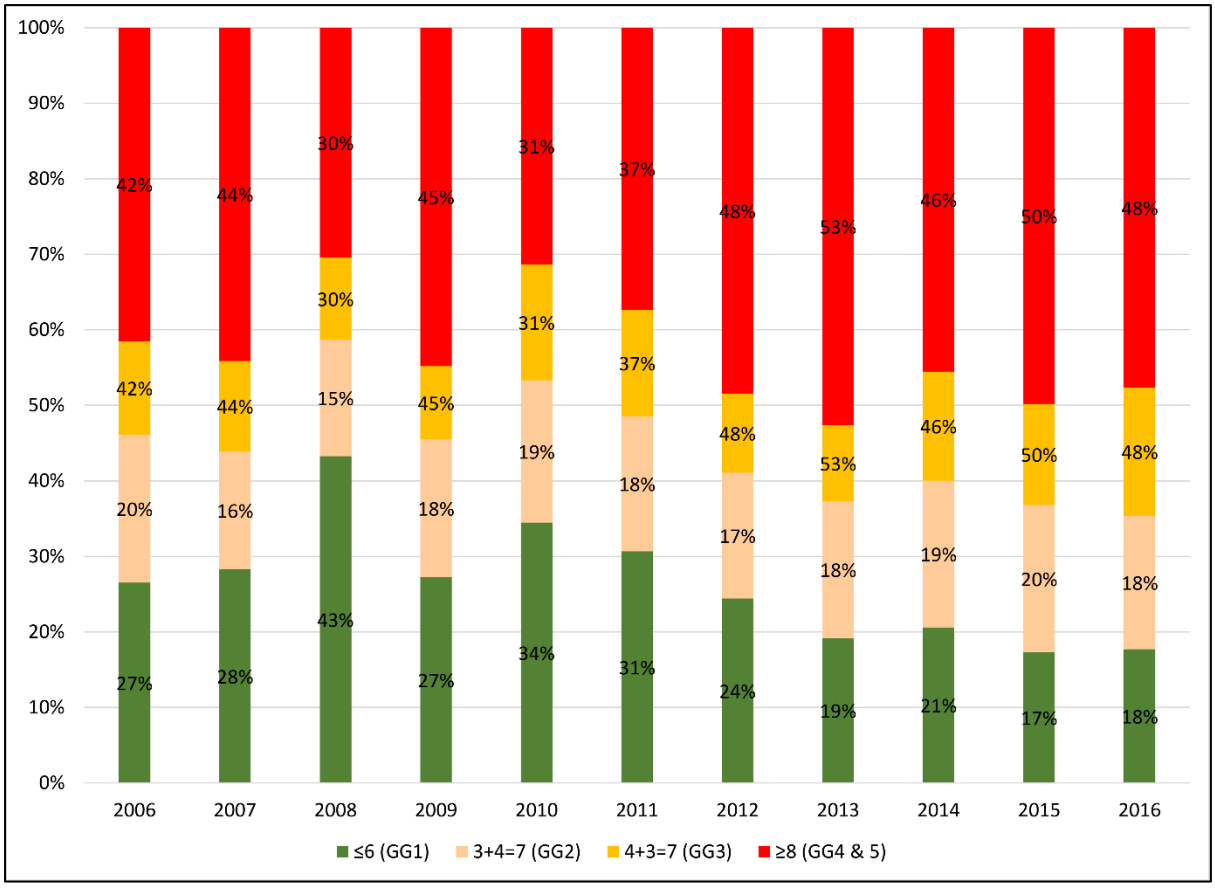


Figure 5.4: Percentage of prostate biopsies with a low (GS ≤6), intermediate (GS =7: 3 + 4 =7 and 4 + 3 =7), and high (GS ≥8) risk Gleason score by calendar year. Bar chart used to report on the percentage of prostate biopsies with prostate cancer finding by GS risk categories. The grade group (GG) is indicated in brackets, with 4 and 5 combined to report a GS ≥8. The percentages reported as data labels per calendar year.

The ASIR increased from 44.92 per 100 000 population in 2006 to 57.31 by 2016 (Figure 5.5). Between 2006 and 2016, we reported an AAPC of 2.7% (95% CI: -1.2 to 6.7) that was significant (p-value ≤0.0001). A lower ASIR was reported for 2008 (27.0 per 100 000 population). Between 2014 and 2015, the ASIR increased from 40.8 to 52.6 per 100 000 population respectively (largest year on year increase).

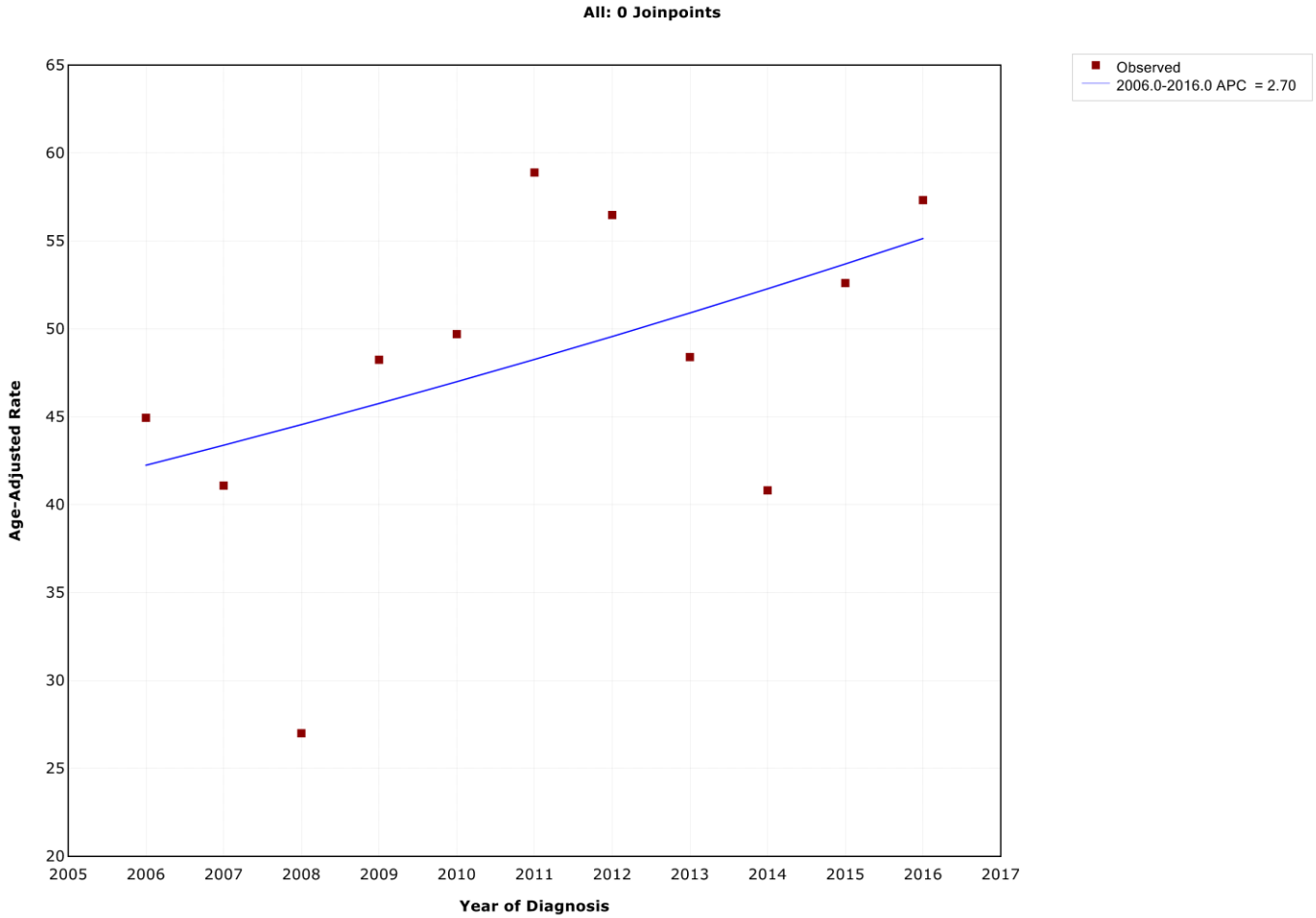


Figure 5.5: Age-standardised prostate cancer incidence rates between 2006 and 2016 reported as orange dots with the average annual percentage change (AAPC) reported as a blue line. The year of diagnosis reported on the X-axis and the age-adjusted rate on the Y-axis.

5.5 Discussion

Our study reported an increase in biopsy numbers between 2006 and 2016, with PCa diagnosed predominantly in older men. The analysis also reveals that the ASIR in the province has increased substantially from 44.9 (2006) to 57.3 (2016) per 100 000 population. We also found that Black Africans were statistically significantly more likely to have advanced disease with a GG ≥4 compared to other racial groups.

The large increase in prostate biopsies reported in our study may indirectly reflect improved access to care. This may include more patient referrals to a urologist following an elevated PSA and/or abnormal DRE. Urology services which were previously limited to academic centres have been introduced in several provincial hospitals [208]. This means that patients can be managed locally removing the need to travel to tertiary hospitals.

The findings of an increasing ASIR are in keeping with Babb *et al* who reported a national rise from 17 in 1986 to 27 per 100 000 population by 2006 in South Africa [91]. Similarly, between 2007 and 2013, the NCR reported ASIR increasing from 29 to 44 per 100 000 population, respectively [93, 209]. We have shown a slightly higher increase compared to Babb *et al* and the NCR reported national values [91, 210]. Between 2007 and 2013, PCa ASIR in the Gauteng Province was substantially higher than national estimates; 41.1 *versus* 29 and 48.4 *versus* 44 per 100 000 population respectively [93, 209]. By reporting provincial trends, we are able to highlight areas with an ASIR increasing faster than nationally reported data. Some of the reasons for the reported increase in incidence include improved urological guidelines, better access to screening, enhanced patient awareness and improved coverage for PCa diagnosis [47, 52].

The 2013 PCa urology treatment guidelines recommend screening from the age of 40 for Black Africans and 45 for all males [52]. Studies have shown that PSA screening is associated with increased diagnosis and decreased mortality [26]. Unpublished data by the authors indicate that PHC PSA testing has increased substantially in the province between 2014 and 2016 [211]. It is reasonable to assume that improved access to screening in the Gauteng Province, specifically at PHC facilities, may have resulted in the increased PCa diagnosis. Screening and education both play important roles in identifying men at risk of developing PCa [26, 47, 52, 212].

It is also possible that PCa awareness is a cause for the increased ASIR noted. Public awareness of female cancers started in 2009, promoting early detection of breast cancer with

mobile mammography units [213]. In contrast, local awareness campaigns for PCa started much later with initiatives such as “Movember,” “Daredevil run” and “Suit up September” beginning in 2015 operating predominantly in metropolitan areas [214-216].

PCa exhibits the highest reported heritability of any major cancer and several genetic susceptibility loci have been identified in European and Asian men [42, 196]. Another reason for the increased ASIR may be due to a combination of genetic and environmental risk factors which are at present not known.

We noted an increase in the percentage of biopsies with a high-risk GS between 2006 and 2016. From 2012, approximately fifty percent of biopsies reported a high-risk GS, highlighting late presentation in the Gauteng Province. A system review by Ilic *et al* reported that PSA testing increased the detection of early-stage disease and slightly decreased the detection of advanced PCa [217]. Similarly, other studies have shown that access to PSA testing has both increased PCa detection rates combined with a migration to lower-stage disease throughout the United States, Canada and Europe [218]. Unpublished data from our Province showed that PHC PSA testing increased dramatically in 2015 [26]. We would, therefore, have expected to see an increase in early-stage disease (GS ≤6). It is possible that it is too early to realise the benefit of increase PSA testing.

We noted that Black African men were statistically significantly more likely to present with a GG ≥4 compared to Whites and Coloureds. Whilst the reasons for this are uncertain, racial differences in health care seeking behaviour have been reported, with Whites and Coloureds respondents more likely to seek care than Black Africans [219]. In South Africa, males have been shown to be less likely to access health care than females and sought care first from traditional healers [219]. Male under–utilization of health care services has been identified in multiple countries [220]. One of the challenges men face include are the working hours of health services [220]. Although, increasing numbers of men presented for diagnosis in our study, it may be that the lack of knowledge is one of multiple factors for late presentation [214-216].

Our findings that older age was a risk factor for high-grade PCa are in keeping with some studies from Canada and the United States [221, 222], whilst local studies have produced conflicting results [1-3]. Although local guidelines recommend PCa screening from the age of 40 years, our findings show that most patients were diagnosed in their mid-sixties [47, 52]. This is similar to local and other published data from South Africa, Ghana and Nigeria [1-3,

223, 224]. Local data indicate that mean age for PSA testing at PHC facilities was 55.6 for Black Africans, highlighting possible delayed diagnosis [85]. The same study showed that only 7% of Black African men who had a PSA ≥4µg/L had a follow-up biopsy [85].

Our study builds on previous work that provides evidence of the value of laboratory data as an invaluable repository of health data, specifically, this work has extended PCa reporting using laboratory data. Applying similar methods of reporting such as the use of SNOMED CT lookup tables could provide national and provincial PCa statistics across South Africa to complement cancer registry reporting. These approaches can be complemented and supplemented with the application of text mining and predictive analytics that have demonstrated the reliable extraction of the prognostic GS data from narrative biopsy reports [225]. These tools have the potential to generate automated PCa reporting leading to the development of PCa epidemiological reporting using laboratory data.

There is a need to improve our understanding of the contribution of genetic, environmental and lifestyle factors for PCa in African men. There is also a need to improve our understanding of late presentation in Black African men. Late presentation is associated with limited treatment options and an increased mortality rate. Broader public health initiatives are required to identify men earlier by improving awareness and health-seeking behaviour as well as providing better access to urology services that meet the needs of men. Here, the provision of tailor-made male-only health care services could play an important role in reducing PCa incidence.

CHAPTER 6 – DISCUSSION

This chapter summarises the results of the study, highlighting key points.

6.1 Summary of findings

For each objective, evidence is presented in Table 6.1 below.

Table 6.1: Consolidated finding

Objectives	Chapter	Evidence
Develop lookup tables to automate the reporting of prostate biopsy data.	2	We used SNOMED CT lookup tables to assign histological findings for 88% of prostate biopsies. The T code identified biopsies of prostatic origin and the M code assigned histological findings.
Develop a text mining algorithm to extract the GS	3	Text mining and predictive analytics were used to extract the GS, which is embedded within the narrative report. The GS was reliably extracted, despite the variations in reporting this information. The manually coded GS values were compared to the extracted values using a confusion matrix, with an F-score of 0.99 reported. The algorithm was evaluated using a validation dataset and reported a similar F-score.
Describe first-ever PSA testing trends for PHC services in the Gauteng Province	4	We reported the number of men receiving a first-ever PSA at PHC facilities, which increased from 1 782 in 2006 to 67 025 by 2016 respectively. The majority of men presented at a clinic (78.1%). Most men were Black Africans (86.4%) and in the 50 - 59 age category (35.8%). A median PSA well within the laboratory cut-off of 4µg/L was reported. A PSA ≥4µg/L was reported for 11.7% of men, increasing to 35.5% for the ≥70 age category.
Describe the number of prostate biopsies performed annually.	5	The number of biopsies performed between 2006 and 2016 increased from 1 630 to 2 761 respectively. Overall, we reported data for 22 937 prostate biopsies.
Describe the number of prostate biopsies by race group and age category.	5	The number of prostate biopsies requested by race group and age category was reported. There were 16 491 biopsies for Black Africans compared to 3 422, 1 129 and 487 for Whites, Coloureds and Indian/Asians respectively. The majority of biopsies were performed for men aged 60 years and older.
Describe the number of prostate biopsies with a PCa histological finding (Adenocarcinoma and GS reported).	5	There were 6 448 biopsies with a PCa finding for Black Africans compared to 111, 354 and 1 009 for Indian/Asians, Coloureds and Whites respectively.

Objectives	Chapter	Evidence
De-duplicate the prostate biopsy data to identify incident patients that were diagnosed with PCa to report ASIR.	5	Following de-duplication, we identified incident PCa cases. The Joinpoint regression program was used to report the ASIR and AAPC. Variables used to calculate ASIR included year of diagnosis, gender (male), age category, number of incident cases (count variable), Gauteng male population estimates and Segi world standard population (WSP). The ASIR increased from 44.92 per 100 000 population in 2006 to 57.31 by 2016, with an AAPC of 2.7% (95% CI: -1.2 to 6.7) that was significant (p-value ≤0.05).
Describe the number of patients presenting with advanced disease (GS ≥8).	5	The number of patients that presented with a GS ≥8 (high-risk) was reported annually and by race group. A high-risk GS was reported for 46% of Black Africans compared to 40%, 38%, and 36% for Coloureds, Indian/Asians and Whites respectively. The percentage of biopsies with a GS ≥8 increased from 42% in 2006 to 48% by 2016.

In summary this study has highlighted a number of points:

- The SNOMED CT lookup tables assigned histological findings for 88% of biopsies.
- The reliable extraction of the GS was demonstrated using text mining.
- A gradual increase in first-ever PSA testing was noted prior to 2015, after which a striking rise was reported most likely due to the 2014 public-sector guidelines. Despite this increase, a median PSA of 0.9µg/L suggests that some screening took place. A PSA ≥4µg/L was reported for only 11.7% of men.
- The PCa ASIR in the Gauteng Province was higher than national estimates reported by the local cancer registry.

6.1.1 SNOMED CT Lookup Tables

No local study has reported using SNOMED CT codes to assign histological findings for prostate biopsies. We have reported comprehensive public-sector data for the Gauteng Province and have demonstrated the use of SNOMED CT codes to extract histological findings for 88% of prostate biopsies. Furthermore, the ASIR reported is consistent with both local cancer registry and GLOBOCON estimates for South Africa [7, 22, 38]. Helgstrand *et al* developed a computer algorithm to transcode SNOMED codes into an analysable format that included the procedure, diagnosis and date of diagnosis [123]. This study reported that SNOMED codes had to be transcoded into a meaningful and analysable format, extracting only PCa findings [123]. In contrast, we have extracted both benign and malignant findings. Furthermore, we have demonstrated the computational steps taken to develop lookup tables. The value of the SNOMED CT codes is that they make it possible to code health information at the appropriate level of detail, without forcing the information to be either too general or too specific [226]. In particular, pathologists can use SNOMED CT codes to index surgical pathologic diagnoses, report the incidence and severity of cancer cases to central registries [226]. Our findings demonstrate that SNOMED CT codes can be used to represent relevant clinical information consistently and comprehensively as an essential part of producing electronic health information [198]. Furthermore, SNOMED CT codes can ensure the availability of more complete PCa data for surveillance purposes and to report incidence [226]. This will shorten the time lag between incident case identification and the availability of statistical reports for analysis [226]. Improved timelines for reporting PCa incidence will substantially improve surveillance efforts.

6.1.2 Using text mining and predictive analytics to extract the GS

All local studies have manually extracted the GS to assess late presentation [1-6, 85]. This explains the small sample sizes for these studies given the effort to manually code narrative reports [1-6, 85]. We extracted the GS using text mining and predictive analytics, which is embedded within the narrative semi-structured prostate biopsy report as alpha, numeric as well as alphanumeric formats. This data meets all six criteria of big data. PCa histological findings provide important clinical value. This data is generated by laboratories using different formats of reports across the NHLS on a regular basis and thus meets the volume, velocity and variety criteria. The narrative biopsy report is captured in a format that is not easy to interpret, meeting the veracity criteria. Finally, the data generated from the narrative text provides valuable information about prognosis. One thousand randomly selected prostate biopsies with an adenocarcinoma finding, derived using SNOMED CT codes with a manually coded GS, were used. The text mining algorithm consisted of the following logical steps: (i)

data acquisition (ii) pre-processing and (iii) feature extraction, (iv) feature value representation, (v) feature selection, (vi) information extraction, (vii) classification and (viii) discovered knowledge. The knowledge discovery output included the episode number, age, major score (3/5), minor score (4/5), total score (7), the combined score (3 + 4 = 7) and risk category (intermediate). When compared to the manually coded GS value, the confusion matrix reported an F-score of 0.99. Our findings show that text mining and predictive analysis can be used to automate the extraction of the GS reliably, removing the need for manual coding. Furthermore, the algorithm was tested against a separate dataset and performed well. A Norwegian study reported using a rule-based algorithm for 25 biopsy reports with an F-score of 0.94 [227]. Similarly, a Northern Ireland study used a Perl routine to extract the GS from 2 263 prostate biopsies [158]. This study reported a precision and recall of 0.99 and 0.98 respectively, that equates to an F-score of 0.98 [158]. The F-scores reported for our study is both acceptable and in line with what has been reported elsewhere [158, 227].

6.1.3 Primary health care first-ever PSA trends

This is the first local study looking at PSA public-sector testing trends over an eleven-year period specifically for PHC facilities for an entire province. Our findings highlighted that prior to the 2014 public-sector guidelines, a gradual increase in first-ever PSA testing was reported compared to a striking increase from 2015.

Most men were aged ≥50 years which is in line with local guidelines [52]. A median PSA of 0.9µg/L was reported which is similar to findings reported by Maphayi *et al [85].* PSA testing in a PHC setting offered by urology registrars reported a median PSA of 20.39µg/L, which is higher than our findings [228]. Results of PSA testing from hospitals reported much higher median PSA values [1, 84]. This supports the fact that men are being seen very late by urologists at hospitals [1, 84]. PSA testing was offered to men attending urology services with a clinical suspicion of PCa or have PCa for the majority of local studies, which explains a higher median PSA [1, 3-6, 84]. For these studies, the PSA values are all very high (>33.5µg/L) suggesting metastatic disease [1, 3-6, 84]. The differences in the data reported for studies from South Africa are possibly due to differences in the study design, racial distribution, geographic setting and the inclusion criteria applied [1, 3-6, 84]. In support of possible screening is the fact that we reported a very low median PSA. We believe that our findings show that there is a need for clear guidelines for testing and referral. Based on the current urology guidelines, some of the testing reported for our study at PHC facilities may be inappropriate as 8.3% was in men aged 30 to 39 years and 11.8% in men ≥ 70 years. Furthermore, these age categories showed substantial increases over the study period.

6.1.4 Prostate biopsy

Our study differs from other local studies that reported data for a hospital-based urology setting for patients that were either diagnosed with PCa or have a clinical suspicion of PCa [1-3]. In contrast, our study reports data for all prostate biopsies requested by public health facilities in the Gauteng Province over an eleven-year period. We reported data for a large sample size of 22 937 biopsies. This is a much larger sample size than all local urological studies that reported data for between 81 and 1 016 patients [1, 2, 4, 5, 85]. Between 2006 and 2016, the number of biopsies increased from 1 630 to 2 761 respectively. Such a large sample size increases the generalisability of our study. Furthermore, we were able to provide trends at the provincial level not previously reported. This will make it possible to identify provinces with higher incidence.

We reported that the majority of biopsies were for Black Africans (77.6%). Local studies outside of the Western Cape Province reported similar data ranging between 69.8% and 79.0% [4, 85]. In contrast, studies in the Western Cape Province reported between 7.9% and 25.9% of biopsies were for Black Africans [1, 5]. This may reflect differences in clinical practices and the population distribution for men accessing care.

We reported a mean age of 67.4 year for Black Africans. This finding is comparable with data reported by local studies who reported that Black Africans at presentation were predominantly in their late sixties [1-4, 6]. In Europe, PCa is the most common neoplasm in elderly males (> 70 years) [26]. Therefore, although men are presenting late with a GS ≥8 they are presenting at an age that is similar to data reported elsewhere [26, 229].

Thirty-nine percent of Black African men in our study had a PCa histological finding which is lower than previous local studies [3, 5, 171]. Local studies reported that between 49.1% and 62.8% of biopsies had PCa [3, 5, 171]. Possible reasons for the discrepancies could include the selected study population (e.g. patients with a urological complaint at a hospital), the study design and sample size. We reported data for an entire province that would differ from data reported in a urological setting for a single hospital.

We reported that 46% of Black Africans had a GS ≥8. Local studies reported that between 39% and 43% of Black Africans reported a GS ≥8 outside of the Western Cape Province [1-3]. Two Western Cape studies reported that 25.9% and 39% of Black Africans had a GS ≥8 [42]. While it is not possible to explain these differences, we postulate that they may arise from differences in the study settings and design, health care service provision, racial and

population differences. In addition, there may be genetic differences that contribute to the findings reported. A family history of PCa increases the risk for PCa in first degree relatives. Genome wide association studies (GWAS) have identified common single nucleotide polymorphisms (SNP) that confer a cumulative risk for PCa [230]. An increased number of SNPs have been identified to show risk for more advanced PCa [230].

Lower first-ever PSA testing between 2012 and 2015 did not correspond to a reduction in prostate biopsies performed for the same period. This indicates that the administrative disruption for PSA testing did not appear to affect PCa diagnosis. This is a positive finding that indicates the urology services continued to diagnose patients despite some challenges with PHC PSA testing.

6.1.5 Prostate cancer age-standardised incidence

We reported an ASIR increasing from 44.92 per 100 000 population in 2006 to 57.31 by 2016. The AAPC was 2.7% (95% CI: -1.2 to 6.7) and was statistically significant. Between 2006 and 2014, the NCR reported a national ASIR increasing from 30.08 to 43.73 per 100 000 population respectively [7, 92]. These findings are at odds with the ASIR we reported for the Gauteng Province. Our findings show that the ASIR reported for the Gauteng Province is higher than national data. The reasons for these differences are not clear at present. It may be that there are indeed regional differences. We are, therefore, unable to assess whether more men were being diagnosed with PCa in each province. Other factors that should be investigated include genetics, environmental exposure and population density (urban/rural divide). PCa risk has been linked to various occupational exposures [45]. Mullins *et al* reported that the environmental and occupational factors are linked to PCa [45].In our context, these factors include pesticides, sunlight/ultraviolet radiation and trace minerals used in tire and battery manufacturing [45]. Local data shows that there are differences across the country in solar radiation and pesticide use, which may account for some of the differences noted [231, 232]. In 2012, the GLOBOCON ASIR estimate for South Africa was 67.9 compared to our finding of 56.5 per 100 000 population [38]. By 2018, the GLOBOCON estimate increased to 68.0 compared to our finding of 57.3 per 100 000 population in 2016 [22]. These findings indicate that the GLOBOCON estimates are generally higher than our findings. Adeloye *et al* reported that the PCa ASIR ranged from 4.4 in an Eastern Cape study to 30.8 per 100 000 population for the local population-based cancer registry (PBCR), that was specifically set up to focus on oesophageal cancer [83].

6.1.6 Implications of our findings

For PHC facilities, a PSA testing scale was noted, especially from 2015, with a very low median PSA and greatest fold increase reported for younger men. The current recommendation of offering testing from 50 years is supported by our findings. If the DRE findings were integrated into the PSA laboratory data, it will be possible to infer which men should have received a prostate biopsy. This will make the laboratory data more powerful, with the potential to generate longitudinal data to follow-up men from PSA testing to diagnosis.

Our findings also show that despite the increased PSA testing, men are still presenting with advanced disease with GS ≥8. A United Kingdom study reported that population-based PSA testing resulted in a downward trend to a lower grade PCa [233]. We reported a 10% difference in the proportion of biopsies with advanced disease between Black Africans and Whites. This highlights the importance of racial differences in late presentation. There are no local data that describe genetic or environmental factors that may play a role in the racial difference in late presentation. This is an area that requires further research.

There is some evidence to suggest that men have poorer health seeking behaviour compared to women. For example, Carmona *et al* reported that compared with females, fewer males access care and more present with an advanced disease [234]. Similarly, Glencross *et al* reported that a disproportionately high number of males compared with females had advanced and very advanced HIV disease for patients accessing care in the Lephalale municipality [235]. Men have been reported to view public health facilities as a place for women and antenatal care [236]. Based on this perception, men indicating that they would prefer attending single-sex health facilities is understandable [236-238]. Therefore, there is a need for men-friendly health services. The dramatic increase in PSA testing in 2015 may indicate that some of these changes were already taking place.

We reported an ASIR that increased substantially between 2006 and 2016. Our findings are higher than NCR estimates but lower than GLOBOCON data [8, 22, 38]. This highlights that national ASIR estimates may not reflect provincial differences. Other studies have reported that African Americans have a higher incidence of PCa than White Americans and also have a worse prognosis [15]. Our findings confirm that almost 50% of Black Africans present with advanced disease. This confirms that different interventions are required for these men. The techniques used to automate PCa findings need to be extended to national data. This could facilitate PCa surveillance data for nine provinces and 52 health districts to identify areas with a higher burden of disease. The figure below describes the implications of our findings.

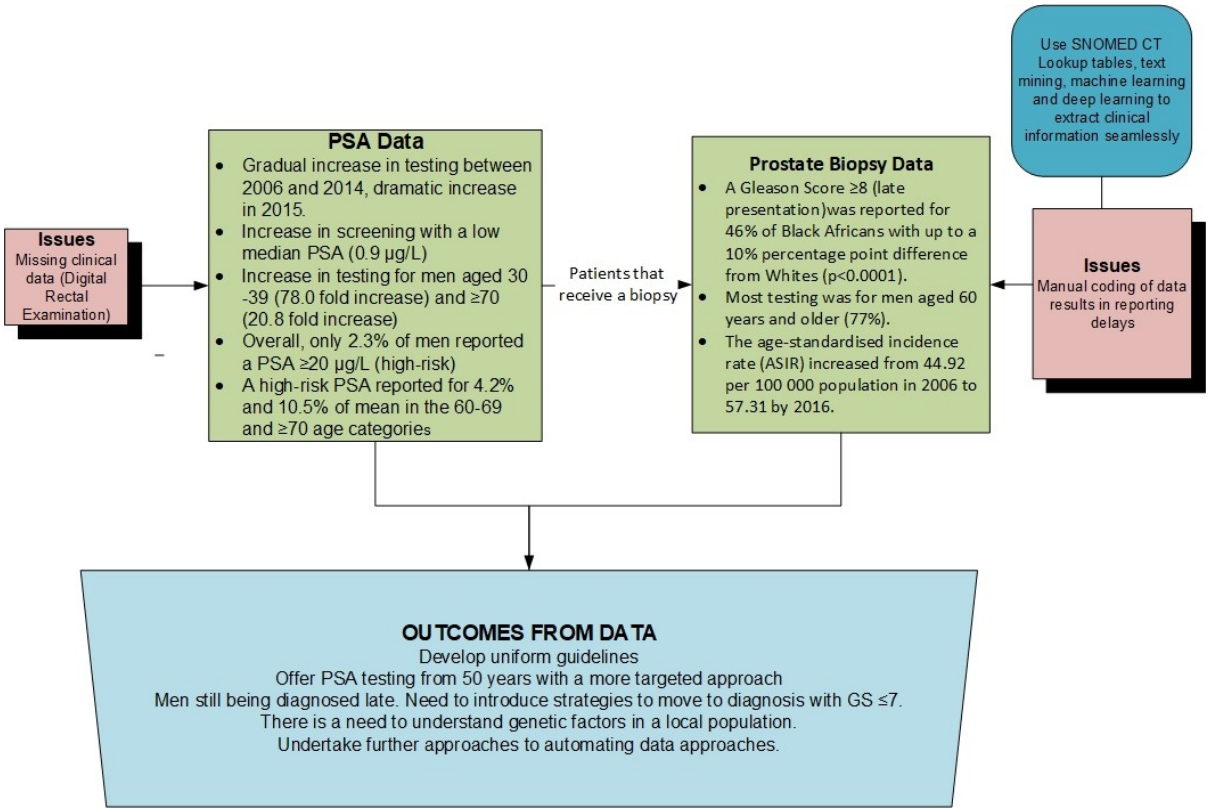


Figure 6.1: Flow chart highlighting the outcomes from the data generated by this study as well as data challenges.

A key outcome of our findings is the need for uniform national guidelines across all levels of care. Currently, urological recommendations differ markedly from PHC public sector guidelines [52, 57]. A key omission in public sector guidelines is that they should clearly indicate that Black African men with a positive family history have a higher risk of developing PCa and may require testing earlier [52].

Furthermore, guidance should be provided on how to interpret PSA results. Urological guidelines recommend age-related cut-off values compared to a fixed cut-off of 4µg/L reported by NHLS laboratories [52]. The cut-off value plays a crucial role with the DRE findings to determine which patients should receive a prostate biopsy. The guidelines need to also be clear on who and when to offer testing and the criteria for follow-up testing. Clear evidence-based guidelines could improve how testing is offered and result in a reduction in both low value testing and a higher median PSA.

This has been achieved in HIV services that uses national consolidated guidelines for all levels of care [239]. The benefit is that they are also used in the private sector [239]. The reason for this is that these guidelines were developed by a team of experts [239]. Previously, there were separate prevention of mother-to-child transmission of HIV (PMTCT) and management of HIV in children, adolescents and adults guidelines [239]. Since 2015, these have been consolidated into a single national guideline [239]. These guidelines were developed by three technical working groups: (i) adult ART, (ii) paediatric ART and (iii) PMTCT ART [239]. They ensure that similar care is provided by all health facilities across South Africa using identical criteria to enrol patients on ART [239]. A similar approach is required for PCa guidelines.

A national technical working group should be assembled that includes NDoH senior staff, urologists, oncologists, pathologists and clinicians to review available evidence to develop national PCa guidelines. These guidelines should include all aspects of PCa management from PSA testing to treatment. This expert group should use evidence-based recommendations to prepare local guidelines. This will ensure that PSA testing is offered across all levels of care in a uniform manner. Guidelines play an important role to inform disease management. The figure below depicts how these guidelines inform health care workers on: when to start testing, assess presentation, identify who to offer testing to as well as how testing should be delivered. With regards to PSA testing, the expert committee have to decide: (i) when to start PSA testing, (ii) how to assess presentation, (iii) who to offer testing to and (iv) how testing should be offered. The prostate biopsy data will inform ASIR, presentation using the GS and when to stop offering a biopsy (if applicable).

The PSA and biopsy data stored in the CDW could be de-duplicated to develop a cohort to identify how many patients should have received a prostate biopsy. This data could provide important insights for waiting times, delayed diagnosis, referral patterns and identify patients that did not present for a biopsy (Figure 6.2). This cohort analysis will be invaluable for planning PCa interventions, drafting policies and updating guidelines. One intervention could be to offer an individualised risk-adapted strategy for early detection to a well-informed man with a good performance status and a life-expectancy of at least 10 - 15 years [26]. This involves counselling men on the potential risks and benefits before subjecting them to PSA testing [26, 240]. One of the benefits of an individualised approach is that low value PSA testing may be reduced.

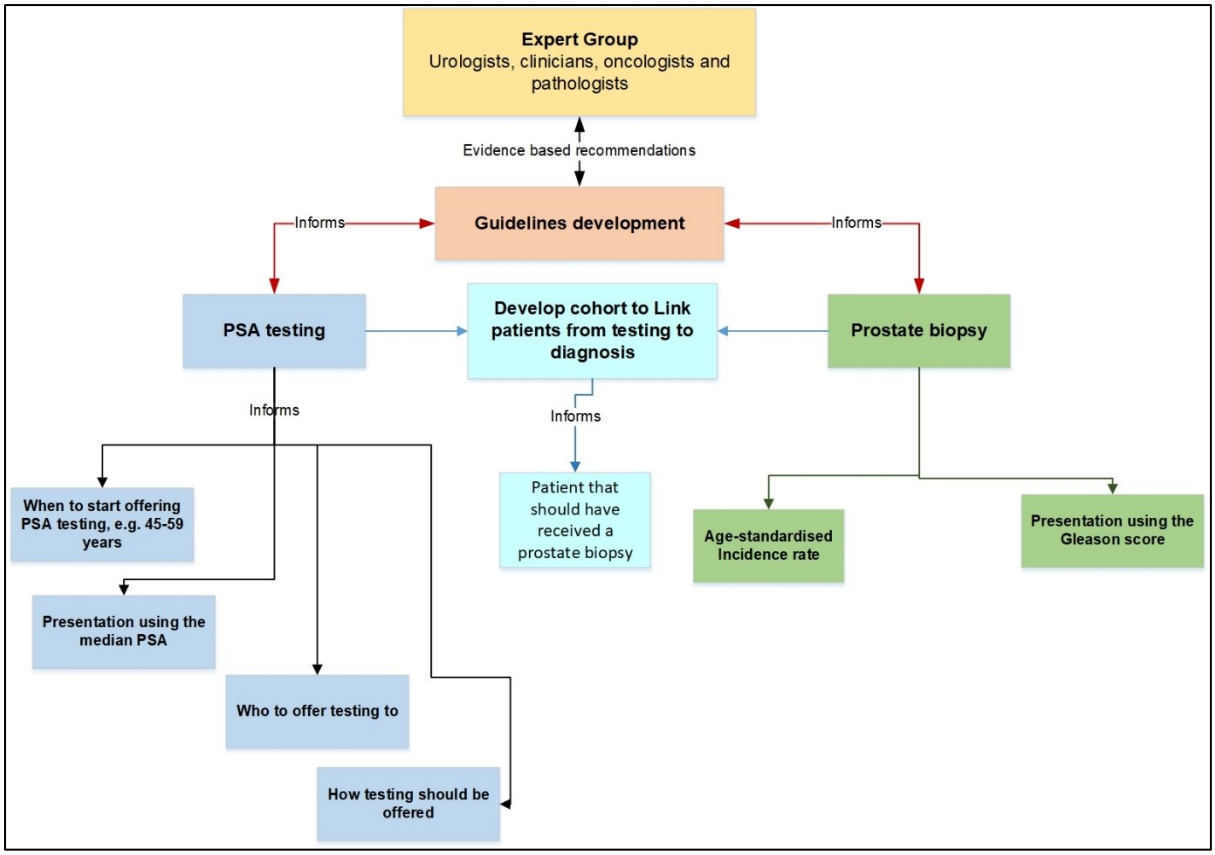


Figure 6.2: Flow chart depicting the relationship between guidelines, PSA and prostate biopsy testing. The information that PSA and prostate biopsy data informs is also indicated.

One of the key data gaps is the absence of clinical data such as the DRE findings and symptoms. The value of adding these to the PSA results is that we will be able to infer which patients should have received a prostate biopsy. In addition to DRE findings, the patient age, risk factors, weight and family history will provide critical information to assess the risk of developing PCa in our population.

We have used big data and AI approaches to automate the extraction of PCa histological findings and the prognostic GS. Histological findings were reported for 88% of prostate biopsies using SNOMED CT lookup tables. To improve PCa reporting, we recommended that the M and T codes be defined as mandatory fields. Furthermore, text mining has demonstrated that it could reliably extract the GS. To address the 12% of biopsies without SNOMED CT codes, ML and deep learning should be used to assign the missing values. The SNOMED CT coded data could be used to train the ML algorithm. A larger national dataset will be required for deep learning.

To routinely assign histological findings in the CDW, the following interventions are required.

The lookup tables developed should be loaded in the CDW. It may be necessary to add any new M and T code combinations that are reported outside the Gauteng Province. Ultimately, the lookup tables should be maintained by an expert committee as follows: (i) CDW identifies new code combinations, (ii) email alert sent to committee members, (iii) committee uses the secure CDW application to assign the missing values, (iv) CDW lookup tables updated and (v) histological findings assigned for all prostate biopsies. In addition, the private sector PCa laboratory data extract could be integrated into the CDW workflow to make it possible to provide national PCa data. Similarly, the text mining API could be loaded in the CDW to routinely extract the GS.

The raw laboratory data is stored in the CDW operational data store (ODS). Next, this data is transformed using the text mining API and SNOMED CT lookup tables to populate histological findings and the GS to the target schema (Figure 6.3). These two interventions will result in a PCa target schema that could transform data within two weeks following the authorisation of the prostate biopsy reports. The mechanism to update the SNOMED CT lookup tables is also depicted.

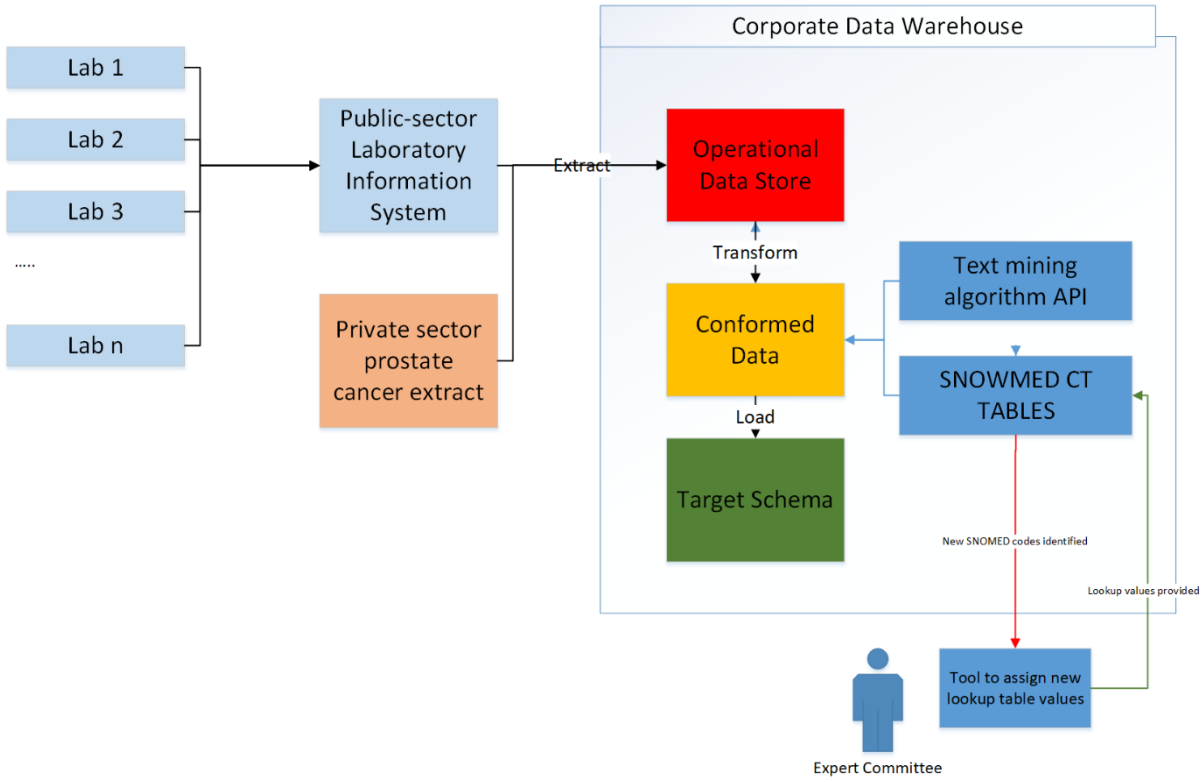


Figure 6.3: Database diagram showing how the SNOMED CT lookup tables and text mining application programming interface (API) could automate the generation of a PCa diagnosis. The private sector prostate cancer data extract could also be accommodated through the CDW process. The Expert Committee would code all new SNOMED CT M and T code combinations to update the CDW lookup tables.

Furthermore, the approaches demonstrated in our study for quantitative data could be applied to NCD. For example, the glycated haemoglobin (HbA1c) data could be analysed to indicate the level of diabetic control in our population. Other examples include chronic kidney disease (CKD), cardiovascular disease and anaemia. This data is already housed in the CDW ODS and will require some work to prepare an NCD target schema. Table 6.2 below provides some indications of the epidemiological reporting that may be possible with the NCD target developed.

Table 6.2: Table describing some of the studies that could be undertaken once the Corporate Data Warehouse target area for Non-Communicable Diseases has been developed. Details are provided for the disease, laboratory test/s, study objectives and outcomes assessed.

Disease	Laboratory Test/s	Objective	Outcome/s
Diabetes	Glycated Haemoglobin (HbA1c) Glucose	Diabetic control	Identify what proportion of patients are well controlled (standard or care *versus* self-management)? Assess glycaemic control trends in a population over time. Identify clinics that achieve better control and those that have poor control. Develop interventions that aim for better performing clinics to share best practices. In addition, they could mentor poorer clinics to improve their performance.
Cardiovascular disease	Risk factor assessment: diabetes, dyslipidaemia, hypertension Troponin T/I/C	Diabetes, lipid control. If electronic health records are captured, then add Blood Pressure control Cardiac injury	Proportion of patients have elevated Troponin results? What proportion of Troponin samples have an acceptable TAT? This is important for patients with elevated Troponin levels. Identify factors that lead to poor TAT and introduce interventions to address the root cause. Interventions could include introducing point of care testing (POCT). What proportion of patients have Acute Coronary Syndromes (ACS)? What proportion of patients with elevated Troponin levels also have hypertension? Determine their demographics. Identify hotspots with high levels of cardiovascular disease and introduce appropriate interventions
Chronic kidney disease (CKD)	Estimated glomerular filtration rate (e-GFR) Urea Creatinine	Kidney function	Identify patients with CKD. The proportion of patients with a reduced estimated GFR (eGFR) which can be used to detect chronic kidney disease (CKD)? The proportion of patients with end-stage renal disease? What proportion of patients with reduced eGFR will require dialysis or transplantation? Identify what are the main risk factors for CKD? Identify sites with poorly controlled risk factors. Introduce interventions to control risk factors such as blood pressure and diabetes control. Determine the rate of progress of GFR decline?
Dyslipidaemia	Total Cholesterol Triglycerides Low density lipoprotein (LDL)	Lipid disorders	What proportion of patients have dyslipidaemia? What proportion of patients have hypercholesterolemia, hypertriglyceridemia or low HDL? Develop ML algorithms to diagnose familial hypercholesterolemia from text intensive reports. This

Disease	Laboratory Test/s	Objective	Outcome/s
	High density lipoprotein (HDL)		has been done in a select population of patients with both clinical and laboratory data. This could be used for all patients in the CDW.
Anaemia	Haemoglobin Full Blood Count	Type of Anaemia	What proportion of patients have aplastic, megaloblastic or iron deficiency anaemia? What is the geographical, racial and age distribution for anaemia? What is the prevalence of anaemia for people accessing health care? Determine how many patients with anaemia have reported a normal Haemoglobin? (Include how long as well).

Similarly, the approaches demonstrated in our study for qualitative data could be applied to other cancers, immunophenotyping and cytogenetic narrative reports. Both text mining and ML could be used to extract findings from text intensive reports. For text mining, experts will have to provide a list of regular expressions to identify cancer findings for example. Similarly, for ML the experts will need to code a sample of reports to train the algorithm. We have started using text mining for PCa and multiple myeloma.

6.2 Conceptual framework and practical aspects for local and national implementation

The conceptual framework depicted below, describes all the aspects that inform both PCa epidemiology and testing patterns. PCa epidemiology is informed by patient demographics, modifiable and non-modifiable risk factors. In terms of laboratory data, both numeric and text-intensive biopsy reports need to be analysed to inform testing patterns. Additional aspects that affect testing patterns include the availability of evidence-based guidelines, societal factors such as awareness, attitude and knowledge of men presenting for care and the availability of health care services. The combination of epidemiology and testing patterns informs government policies, public-sector guidelines development, national cancer control programmes and ASIR estimates.

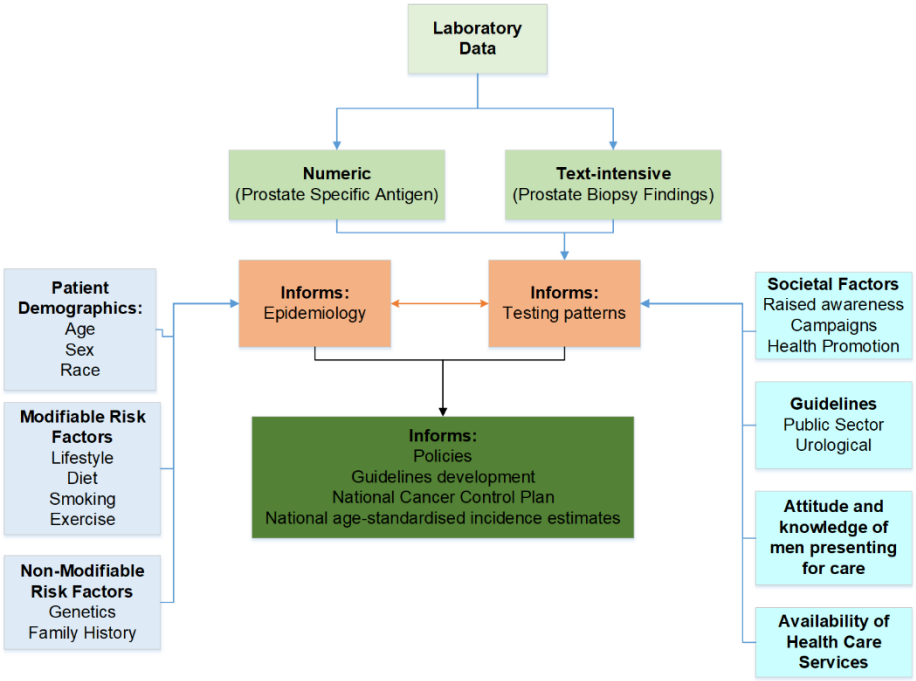


Figure 6.4: Prostate cancer conceptual framework

6.3 Areas of future research

We have attempted to use routinely collected data from the NHLS CDW to report PCa trends. These approaches could be extended to other diseases. In particular, text mining could be extended to all text intensive reports stored in the CDW.

The approaches used for PCa in the Gauteng Province should be applied to national data to assess late presentation for nine provinces. This has never been done previously in South Africa and will be achieved by using the developed SNOMED CT lookup tables and text mining algorithm.

Once uniform PCa guidelines have been implemented in South Africa, PSA and prostate biopsy data can be reviewed to assess whether there are any changes in the median PSA, late presentation and the ASIR. Improved guidelines should have reduced low value testing. These findings may highlight the need to understand PCa genetics.

The SNOMED CT coded prostate biopsy data used for our study should be used to develop a supervised ML algorithm. This will address the 12% of biopsies without a SNOMED CT code. This data should be divided into a training and testing data set. The algorithm should be trained to predict values provided by the T and M lookup tables. Once trained, a confusion matrix should be used to compare the predicted values against the lookup table values (assumed to be the gold standard). Should the ML algorithm produce a satisfactory F-score, it will be

feasible to investigate using deep learning approaches for national data over a longer time span.

Furthermore, the SNOMED CT lookup tables developed for PCa should be extended to other cancers of public health concern. Most of these cancers are reported using the SNOMED CT coding system and could be reported in a similar fashion. This will require a small expert group to code the unique combinations reported, after which the data can be analysed.

Other approaches that could be used to extract the GS from prostate biopsy data such as Apache cTAKES and MetaMap should be used.

6.4 Limitations and challenges

This study has several limitations as described below:

Due to the absence of a national unique patient identifier, we have used the probabilistic matching algorithm to de-duplicate patients. De-duplication may be sub-optimal for patients with incomplete demographic information. One of the shortcomings of our study is that our racial distribution does not match population estimates (Census 2011) [86]. In particular, Whites and Asian/Indians appear to be under-represented [86]. This could be explained by the disparity in private medical aid coverage in South Africa [241].

Due to the paucity of race group data, we used the hot-deck imputation method [178]. The accuracy of the hot-deck race imputation was also a limitation. In an unpublished study, Chen *et al* used a hold-out test to compare imputed *versus* cancer registry patient-reported race group values for 406 642 unique surname-race group pairings [179]. The historically well-populated cancer registry data was split into patient-reported race group values and imputed values [179]. This study compared the imputed and original values [179]. For the imputed dataset, a sensitivity of ≥90% was reported for Black Africans, Whites, and Indian/Asians [179]. However, a lower sensitivity of 63.8% was reported for Coloureds [179]. This confirms that the imputation may be less accurate for Coloureds, which may result in misclassification. A challenge in South Africa is that common surnames are used interchangeably by both Coloureds and Whites [242].

A limitation of our study is the absence of private sector laboratory data, affecting generalisability. There are planned discussions to receive private sector PCa data routinely. However, a local study reported that withheld private sector cancer data only accounted for a 3.8% decrease [243].

An additional limitation of our data is that 12% of biopsies did not have a SNOMED code. Unfortunately, these fields are not mandatory and were omitted in the data. By amending rules and making these fields mandatory, all biopsies will prospectively be reviewed with at least one T and M code captured. The excluded data will also affect the generalisability of our study. The only option for this data was to manually assign histological findings.

Unfortunately, alternative NLP approaches such as Apache cTAKES and MetaMap were not utilised to extract the GS for our study [244]. These approaches were not used as we had expertise with python for text mining. These NLP systems will be used against national PCa data and compared with techniques used in our study.

It was not possible to control for confounders such as socioeconomic status and environmental exposure as this data was not available.

6.5 Recommendations

We have referred to the recommendation to using ML to assign histological findings for biopsies without a SNOMED CT code. This will address the 12% of biopsies that were manually coded. This algorithm will have to: (i) identify biopsies of prostatic origin and (ii) identify biopsies with an adenocarcinoma histological finding. A confusion matrix should be used to determine the precision, recall and the F-score against the SNOMED CT assigned values.

The poor quality of race group data in the CDW should be addressed by making this variable mandatory. This will enforce compliance to provide this information for all samples. The value of this approach is that other studies where race plays a role in prevalence for example, will benefit from having this data populated in the CDW for all tests.

Once the prostate biopsy is transformed for national data, an epidemiological dashboard should be developed to report on trends by age, race group and geographical boundaries. Similar dashboards have been developed for HIV, tuberculosis (TB) and antimicrobial resistance (AMR) that reports public sector trends from the national to the health facility level [245, 246].

The HIV monitoring and evaluation (M&E) dashboard displayed below, titled 'HIV M&E Facility Indicators and Trends (All Ages)' is reporting data for the Mofolo Clinic in the City of Johannesburg for the last twelve months. This dashboard makes it possible for the facility manager to assess trends to progress towards reaching the target of 90% of patients on HIV treatment achieve viral suppression [247]. Therefore, the focus of the dashboard was to report the percentage of patients with a HIV viral load ≤1000 copies/ml, defined as viral suppression [247]. The dashboard provides information for the: (i) proportion of patients that are virologically suppressed, (ii) proportion of patients for the various HIV viral load categories, (iii) distribution of testing by gender and age, (iv) levels of viral suppression by gender and age category and (v) monthly trends (Figure 6.5). The detailed information provided by this dashboard is summarised in Appendix H. In addition, there are other HIV M&E dashboards that were developed for which details and screenshots are provided in Appendix I.

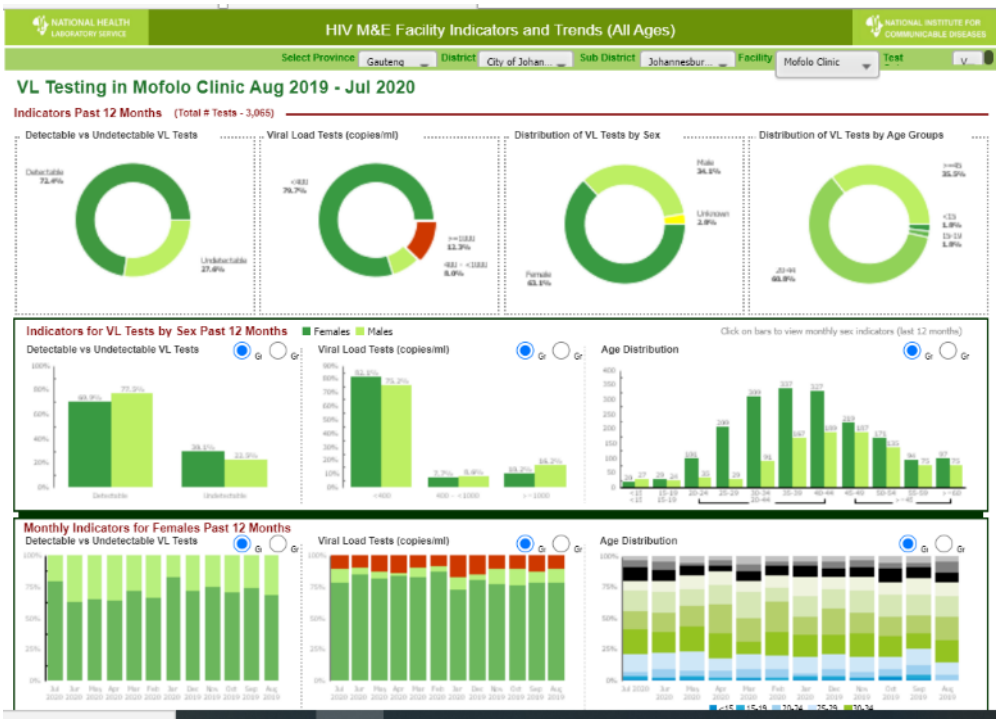


Figure 6.5: HIV Monitoring and Evaluation Facility Indicators and Trends (All Ages) dashboard screenshot (taken from the National Institute for Communicable Diseases (NICD) monitoring and evaluation online HIV programme dashboards). The dashboard displayed is titled 'HIV M&E Facility Indicators and Trends (All Ages)'. (National Institute for Communicable Disease (NICD) – HIV M&E Facility Indicators and Trends (All Ages) – August 2020).

Similarly, the transformed PCa data should be used to report the number of incident cases at the provincial, district and sub-district levels to identify areas where interventions are required in South Africa. The dashboard should 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.

To strengthen the laboratory data stored in the CDW, it will be beneficial to add information such as patient demographics, age, risk factors, weight and DRE findings. This data will be matched to the PSA and prostate biopsy data already in the CDW. This could be used to describe presentation and diagnosis. In addition, it will be possible to indicate which patients should have received a prostate biopsy. Furthermore, the CDW has created a clean target area for HIV, TB and AMR testing. Unfortunately, most NCD-related testing data has to be extracted from the ODS and then cleaned up. We strongly recommend that the CDW start working on a comprehensive NCD target area. This would make it possible to generate more local studies on the burden of NCD in South Africa. The NCD focus should be to provide epidemiological data for multiple levels of care for diabetes control for example.

Furthermore, text mining and ML should be used to extract histological findings for other cancers. Text mining is optimal for text-intensive biopsy reports. In addition, alternative approaches such cTakes and MetaMap should be tested on the prostate biopsy data.

6.6 Conclusions

This thesis set out to conduct a retrospective study to investigate prostate cancer screening and histological diagnosis using routinely collected laboratory data. It aimed to develop automated methods to reduce the burden of manual coding to extract both the histological findings and the GS. A further aim was to describe first-ever PSA testing for PHC services. The final aim was to describe late presentation specifically for Black Africans and report PCa ASIR.

For the very first time, SNOMED CT lookup tables were used in South Africa to assign histological findings for 88% of prostate biopsies. Furthermore, text mining has been shown to reliably extract the GS. This thesis provides evidence that routinely collected laboratory data can be used to describe first-ever PSA testing and PCa diagnosis. We report a dramatic increase in PSA testing between 2014 and 2015, most likely due to the 2014 public-sector guidelines. A median of 0.9µg/L was reported that is well within the laboratory cut-off of 4µg/L. This suggests that some screening took place. Prostate biopsy numbers increased substantially between 2006 and 2016, with the majority of biopsies requested for Black Africans and men aged 60 years and older. A PCa finding was reported for 39% of Black

Africans, of which 46% reported a GS ≥8. The PCa ASIR increased from 44.92 per 100 000 population in 2006 to 57.31 by 2016.

This thesis provides evidence of the value of laboratory data to report PCa trends in an automated fashion. These techniques may also be used to provide national PCa statistics across South Africa to complement cancer registry reporting. Furthermore, these approaches may be used to report data for NCD, especially other cancers of public health interest.

7. REFERENCES

1. Heyns CF, Fisher M, Lecuona A, van der Merwe A. Prostate cancer among different racial groups in the Western Cape: Presenting features and management. South African Medical Journal. 2011;101(9):267-70.doi:https://doi.org.10.7196/samj.4420
2. 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.doi:https://doi.org.https://doi.org/10.1002/pros.22806
3. Tindall EA, Monare LR, Petersen DC, van Zyl S, Hardie RA, Segone AM, et al. Clinical presentation of prostate cancer in black South Africans. Prostate. 2014;74(8):880-91.doi:https://doi.org.10.1002/pros.22806
4. Singh K, Abdel Goad EH, Ramklass SS. Waiting times for prostate cancer diagnosis in KwaZulu-Natal, South Africa. South African Medical Journal. 2015;105(6):484-6.doi:https://doi.org.10.7196/SAMJ.9192
5. Dewar M, Kaestner L, Zikhali Q, Jehle K, Sinha S, Lazarus J. Investigating racial differences in clinical and pathological features of prostate cancer in South African men. S Afr J Surg. 2018;56(2):54-8.doi:https://doi.org.10.17159/2078-5151/2018/v56n2a2324
6. Sherriff A, Da Costa N, Engelbrecht A, Li A, Price N, Joubert G. Prostate cancer profile and risk stratification of patients treated at Universitas Annex Department of Oncology, Bloemfontein, Free State, during 2008 to 2010. South African Family Practice. 2015;57(4):247-52.doi:https://doi.org.10.1080/20786190.2014.993859
7. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2014 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD) 2014 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2014-NCR-tables.pdf] (Accessed Date: 12 May 2018).
8. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2016 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD). 2016 [Available from: https://www.nicd.ac.za/wp-content/uploads/2020/04/NCR_2016_Report_updated_14April2020.pdf] (Accessed Date: 12 May 2018).
9. Budreviciute A, Damiati S, Sabir DK, Onder K, Schuller-Goetzburg P, Plakys G, et al. Management and Prevention Strategies for Non-communicable Diseases (NCDs) and Their Risk Factors. Frontiers in Public Health. 2020;8(788).doi:https://doi.org.10.3389/fpubh.2020.574111
10. Hunter DJ, Reddy KS. Noncommunicable diseases. N Engl J Med. 2013;369(14):1336-43.doi:https://doi.org.10.1056/NEJMra1109345
11. Mayosi BM, Flisher AJ, Lalloo UG, Sitas F, Tollman SM, Bradshaw D. The burden of non-communicable diseases in South Africa. Lancet. 2009;374(9693):934-47.doi:https://doi.org.10.1016/S0140-6736(09)61087-4
12. World Health Organisation (WHO). Global status report on noncommunicable diseases Geneva, Switzerland: World Health Organisation (WHO). 2014 [Available from: http://apps.who.int/iris/bitstream/handle/10665/148114/9789241564854_eng.pdf;jsessionid=A6942517CED81F199D8219808D3A2DAC?sequence=1] (Accessed Date: 29 August 2018).
13. World Health Organisation (WHO). Noncommunicable diseases Geneva, Switzerland: World Health Organisation (WHO). 2018 [Available from: http://www.who.int/en/news-room/fact-sheets/detail/noncommunicable-diseases] (Accessed Date: 27 November 2018).
14. Islami F, Goding Sauer A, Miller KD, Siegel RL, Fedewa SA, Jacobs EJ, et al. Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. CA Cancer J Clin. 2018;68(1):31-54.doi:https://doi.org.10.3322/caac.21440

15. Farmer R. Prostate cancer: epidemiology and risk factors. Trends in Urology, Gynaecology & Sexual Health. 2008;13(3):32-4.doi:https://doi.org.10.1002/tre.71
16. World Health Organisation (WHO). Global status report on noncommunicable diseases Geneva, Switzerland: World Health Organisation (WHO). 2010 [Available from: http://www.who.int/nmh/publications/ncd_report_full_en.pdf] (Accessed Date: 1 April 2018).
17. Nojilana B, Bradshaw D, Pillay-van Wyk V, Msemburi W, Somdyala N, Joubert JD, et al. Persistent burden from non-communicable diseases in South Africa needs strong action. S Afr Med J. 2016;106(5):23-4.doi:https://doi.org.10.7196/SAMJ.2016.v106i5.10776
18. Kumar V, Abul KA, Jon CA, Stanley LR. Robbins Basic Pathology Philadelphia, PA, USA: Elsevier; 2018. 952 p.
19. Torre LA, Siegel RL, Ward EM, Jemal A. Global Cancer Incidence and Mortality Rates and Trends--An Update. Cancer Epidemiol Biomarkers Prev. 2016;25(1):16-27.doi:https://doi.org.10.1158/1055-9965.EPI-15-0578
20. Afkar A, Jalilian H, Pourreza A, Mir H, Sigaroudi AE, Heydari S. Cost analysis of breast cancer: a comparison between private and public hospitals in Iran. BMC Health Services Research. 2021;21(1):219.doi:https://doi.org.10.1186/s12913-021-06136-6
21. World Health Organisation (WHO). Cancer Geneva, Switzerland: World Health Organisation (WHO),. 2019 [Available from: https://www.who.int/health-topics/cancer#tab=tab_1] (Accessed Date: 13 May 2019).
22. 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.doi:https://doi.org.10.3322/caac.21492
23. Botha S, Maphayi MR, Cassim N, George JA. Vitamin D [25(OH)D] and 1,25(OH)2D serum concentrations in patients tested at the Charlotte Maxeke Johannesburg Academic Hospital. Journal of Endocrinology, Metabolism and Diabetes of South Africa. 2021;26(2):52-9.doi:https://doi.org.10.1080/16089677.2021.1886724
24. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer. 2015;136(5):E359-E86.doi:https://doi.org.https://doi.org/10.1002/ijc.29210
25. Aaron L, Franco OE, Hayward SW. Review of Prostate Anatomy and Embryology and the Etiology of Benign Prostatic Hyperplasia. Urol Clin North Am. 2016;43(3):279-88.doi:https://doi.org.10.1016/j.ucl.2016.04.012
26. European Association of Urology. Guidelines on Prostate Cancer Aarnheim, Netherlands: European Association of Urology. 2016 [Available from: https://uroweb.org/wp-content/uploads/EAU-Guidelines-Prostate-Cancer-2016.pdf] (Accessed Date: 19 February 2018).
27. Sinnott JA, Rider JR, Carlsson J, Gerke T, Tyekucheva S, Penney KL, et al. Molecular differences in transition zone and peripheral zone prostate tumors. Carcinogenesis. 2015;36(6):632-8.doi:https://doi.org.10.1093/carcin/bgv051
28. Andersson KE. LUTS treatment: future treatment options. Neurourol Urodyn. 2007;26(6 Suppl):934-47.doi:https://doi.org.10.1002/nau.20500
29. Lepor H. Pathophysiology of lower urinary tract symptoms in the aging male population. Rev Urol. 2005;7 Suppl 7(Suppl 7):S3-S11.
30. National Cancer Institute at the National Institutes of Health (NIH). Understanding Prostate Changes: A Health Guide for Men Rockville, MD, USA: National Cancer Institute at the National Institutes of Health (NIH). 2018 [Available from: https://www.cancer.gov/types/prostate/understanding-prostate-changes] (Accessed Date: 19 February 2018).

31. Perletti G, Monti E, Magri V, Cai T, Cleves A, Trinchieri A, et al. The association between prostatitis and prostate cancer. Systematic review and meta-analysis. Arch Ital Urol Androl. 2017;89(4):259-65.doi:https://doi.org.10.4081/aiua.2017.4.259
32. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230-43.doi:https://doi.org.10.1136/svn-2017-000101
33. Dai X, Fang X, Ma Y, Xianyu J. Benign Prostatic Hyperplasia and the Risk of Prostate Cancer and Bladder Cancer: A Meta-Analysis of Observational Studies. Medicine (Baltimore). 2016;95(18):e3493.doi:https://doi.org.10.1097/MD.0000000000003493
34. Krieger JN, Lee SWH, Jeon J, Cheah PY, Liong ML, Riley DE. Epidemiology of prostatitis. International Journal of Antimicrobial Agents. 2008;31:85-90.doi:https://doi.org.https://doi.org/10.1016/j.ijantimicag.2007.08.028
35. American Cancer Society. What Is Prostate Cancer? Atlanta, Georgia, USA: American Cancer Society. 2016 [Available from: https://www.cancer.org/cancer/prostate-cancer/about/what-is-prostate-cancer.html] (Accessed Date: 19 February 2018).
36. Cooperberg MR, Chan JM. Epidemiology of prostate cancer. World J Urol. 2017;35(6):849.doi:https://doi.org.10.1007/s00345-017-2038-0
37. Neupane S, Bray F, Auvinen A. National economic and development indicators and international variation in prostate cancer incidence and mortality: an ecological analysis. World J Urol. 2017;35(6):851-8.doi:https://doi.org.10.1007/s00345-016-1953-9
38. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, et al. GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012 v1.0 Lyon, France: International Agency for Research on Cancer (IARC). 2013 [Available from: https://publications.iarc.fr/Databases/Iarc-Cancerbases/GLOBOCAN-2012-Estimated-Cancer-Incidence-Mortality-And-Prevalence-Worldwide-In-2012-V1.0-2012] (Accessed Date: 29 February 2018).
39. Tao ZQ, Shi AM, Wang KX, Zhang WD. Epidemiology of prostate cancer: current status. Eur Rev Med Pharmacol Sci. 2015;19(5):805-12.
40. Center MM, Jemal A, Lortet-Tieulent J, Ward E, Ferlay J, Brawley O, et al. International variation in prostate cancer incidence and mortality rates. Eur Urol. 2012;61(6):1079-92.doi:https://doi.org.10.1016/j.eururo.2012.02.054
41. Crawford ED. Epidemiology of prostate cancer. Urology. 2003;62(6 Suppl 1):3-12.doi:https://doi.org.10.1016/j.urology.2003.10.013
42. Rebbeck TR. Prostate Cancer Genetics: Variation by Race, Ethnicity, and Geography. Semin Radiat Oncol. 2017;27(1):3-10.doi:https://doi.org.10.1016/j.semradonc.2016.08.002
43. Foulkes WD, Cooney KA. Male Reproductive Cancers: Epidemiology, Pathology and Genetics. Foulkes WD, Cooney KA, editors. New York,. NY, USA: Springer; 2020.
44. Ferro M, Terracciano D, Buonerba C, Lucarelli G, Bottero D, Perdona S, et al. The emerging role of obesity, diet and lipid metabolism in prostate cancer. Future Oncol. 2017;13(3):285-93.doi:https://doi.org.10.2217/fon-2016-0217
45. Mullins JK, Loeb S. Environmental exposures and prostate cancer. Urol Oncol. 2012;30(2):216-9.doi:https://doi.org.10.1016/j.urolonc.2011.11.014
46. Willett DL, Kannan V, Chu L, Buchanan JR, Velasco FT, Clark JD, et al. SNOMED CT Concept Hierarchies for Sharing Definitions of Clinical Conditions Using Electronic Health Record Data. Appl Clin Inform. 2018;9(3):667-82.doi:https://doi.org.10.1055/s-0038-1668090
47. Anderson D, Barnes R, Bida M, Bigalke M, Bongers M, Chetty P, et al. South African Prostate Cancer Guideline: South African Urological Association (SAUA), South African Society of Medical Oncology (SASMO), South African Society for Clinical and Radiation Oncologists (SASCRO), South African Society of Nuclear Medicine (SASNM), South African

Oncology Consortium (SAOC) and The Prostate Cancer Foundation of South Africa (PCF). 2017 [21]. Available from: http://prostate-ca.co.za/wp-content/uploads/2017ProstateGuidelinesDraftVersion2016.pdf] (Accessed Date: 12 March 2017).
48. Sauter G, Steurer S, Clauditz TS, Krech T, Wittmer C, Lutz F, et al. Clinical Utility of Quantitative Gleason Grading in Prostate Biopsies and Prostatectomy Specimens. Eur Urol. 2016;69(4):592-8.doi:https://doi.org.10.1016/j.eururo.2015.10.029
49. Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA, et al. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. Am J Surg Pathol. 2016;40(2):244-52.doi:https://doi.org.10.1097/pas.0000000000000530
50. Pudasaini S, Subedi N. Understanding the gleason grading system and its changes. Journal of Pathology of Nepal. 2019;9:1580-5.doi:https://doi.org.10.3126/jpn.v9i2.25723
51. Moch H, Cubilla AL, Humphrey PA, Reuter VE, Ulbright TM. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part A: Renal, Penile, and Testicular Tumours. Eur Urol. 2016;70(1):93-105.doi:https://doi.org.10.1016/j.eururo.2016.02.029
52. Segone AM, Haffejee M, Wentzel S, Heyns CF, Mutambirwa SBA, Coetzee L, et al. Prosate Cancer Diagnostic and Treatment Guidelines: The Prostate Cancer Foundation of South Africa. 2013 [14]. Available from: http://prostate.acitravel.co.za/cake/app/webroot/uploads/files/Prostate_Cancer_Guidelines_2013.pdf] (Accessed Date: 12 January 2017).
53. Sobin LH, Hermanek P, Hutter RV. TNM classification of malignant tumors. A comparison between the new (1987) and the old editions. Cancer. 1988;61(11):2310-4.doi:https://doi.org.10.1002/1097-0142(19880601)61:11<2310::aid-cncr2820611127>3.0.co;2-x
54. Cheng L, Montironi R, Bostwick DG, Lopez-Beltran A, Berney DM. Staging of prostate cancer. Histopathology. 2012;60(1):87-117.doi:https://doi.org.https://doi.org/10.1111/j.1365-2559.2011.04025.x
55. D'Amico AV. Risk-based management of prostate cancer. N Engl J Med. 2011;365(2):169-71.doi:https://doi.org.10.1056/NEJMe1103829
56. European Association of Urology. Guidelines on Prostate Cancer Aarnheim, Netherlands: European Association of Urology. 2021 [Available from: https://uroweb.org/wp-content/uploads/EAU-EANM-ESTRO_ESUR_ISUP_SIOG-Pocket-Guidelines-on-Prostate-Cancer-2021-2.pdf] (Accessed Date: 29 May 2021).
57. National Department of Health (NDOH). Standard Treatment guidelines and essential medicines list for South Africa: Primary Health Care level Pretoria, South Africa: National Department of Health (NDOH). 2014 [Available from: http://www.kznhealth.gov.za/pharmacy/edlphc2014a.pdf] (Accessed Date: 12 September 2016).
58. Henry J. Kaiser Family Foundation. An Overview of Health and Health care in South Africa 1994 – 2010: Priorities, Progress and Prospects for New Gains Johannesburg, South Africa: Henry J. Kaiser Family Foundation. 2010 [Available from: https://evaluations.dpme.gov.za/evaluations/9] (Accessed Date: 1 April 2017).
59. United States Agency for International Development (USAID), World Health Organisation (WHO), Management Sciences For Health (MSH). Drug and Therapeutics Committee Training Course: Session 10 - Standard Treatment Guidelines Participants' Guide Geneva, Switzerland: United States Agency for International Development (USAID), World Health Organisation (WHO),

Management Sciences For Health (MSH),. 2007 [Available from: https://www.who.int/medicines/technical_briefing/tbs/10-PG_Standard-Treatment-Guidelines_final-08.pdf] (Accessed Date: 1 May 2016).
60. Hassan A, Abdullahi GA, Ahmed AA, Sabiu A, John U, Ezekeil G, et al. Assessing prescriber's awareness of essential medicine list, hospital drug formulary and utilization of standard treatment guidelines in a tertiary healthcare facility in North-Central Nigeria. Alexandria Journal of Medicine. 2018;54(1):81-4.doi:https://doi.org.10.1016/j.ajme.2017.03.006
61. National Department of Health (NDOH). Standard Treatment Guidelines and Essential Medicine List for South Africa Pretoria, South Africa: National Department of Health (NDOH). 2014 [508]. Available from: http://www.kznhealth.gov.za/pharmacy/edlphc2014a.pdf] (Accessed Date: 12 November 2016).
62. National Department of Health (NDOH). Standard Treatment Guidelines and Essential Medicine List for South Africa: Primary Health Care Level Pretoria, South Africa: National Department of Health (NDOH). 2008 [508]. Available from: https://www.hst.org.za/publications/NonHST%20Publications/STGandEML%20PHC%202008.pdf] (Accessed Date: 12 November 2018).
63. National Department of Health (NDOH). Standard Treatment guidelines and essential medicines list for South Africa: Hospital level Pretoria, South Africa: National Department of Health (NDOH). 2012 [Available from: http://www.kznhealth.gov.za/pharmacy/edladult_2012.pdf] (Accessed Date: 12 September 2016).
64. National Department of Health (NDOH). Standard Treatment guidelines and essential medicines list for South Africa: Hospital level Pretoria, South Africa: National Department of Health (NDOH). 2015 [Available from: https://www.gov.za/sites/default/files/gcis_document/201409/standardtreatmentguidelinesandessentialmedicineslist2012s.pdf] (Accessed Date: 12 September 2019).
65. Parker C, Castro E, Fizazi K, Heidenreich A, Ost P, Procopio G, et al. Prostate cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2020;31(9):1119-34.doi:https://doi.org.10.1016/j.annonc.2020.06.011
66. Carter HB, Albertsen PC, Barry MJ, Etzioni R, Freedland SJ, Greene KL, et al. Early detection of prostate cancer: AUA Guideline. J Urol. 2013;190(2):419-26.doi:https://doi.org.10.1016/j.juro.2013.04.119
67. Mohler JL, Antonarakis ES, Armstrong AJ, D'Amico AV, Davis BJ, Dorff T, et al. Prostate Cancer, Version 2.2019, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2019;17(5):479-505.doi:https://doi.org.10.6004/jnccn.2019.0023
68. Arnsrud Godtman R, Holmberg E, Lilja H, Stranne J, Hugosson J. Opportunistic testing versus organized prostate-specific antigen screening: outcome after 18 years in the Goteborg randomized population-based prostate cancer screening trial. Eur Urol. 2015;68(3):354-60.doi:https://doi.org.10.1016/j.eururo.2014.12.006
69. Hayes JH, Barry MJ. Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence. JAMA. 2014;311(11):1143-9.doi:https://doi.org.10.1001/jama.2014.2085
70. Loeb S. Guideline of Guidelines: Prostate Cancer Screening Guidelines. BJU Int. 2014;114(3):323-5.doi:https://doi.org.10.1111/bju.12854
71. Mason MM, Evans CP. Prostate Cancer. In: Mason MM, Evans CP, editors. eLS. 2. Hoboken, NJ, USA: Wiley; 2021. p. 103-10.doi:https://doi.org.10.1002/9780470015902.a0029270

72. Andriole GL, Crawford ED, Grubb RL, 3rd, Buys SS, Chia D, Church TR, et al. Mortality results from a randomized prostate-cancer screening trial. N Engl J Med. 2009;360(13):1310-9.doi:https://doi.org.10.1056/NEJMoa0810696
73. Hugosson J, Carlsson S, Aus G, Bergdahl S, Khatami A, Lodding P, et al. Mortality results from the Goteborg randomised population-based prostate-cancer screening trial. Lancet Oncol. 2010;11(8):725-32.doi:https://doi.org.10.1016/S1470-2045(10)70146-7
74. Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, et al. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009;360(13):1320-8.doi:https://doi.org.10.1056/NEJMoa0810084
75. Ilic D, Neuberger MM, Djulbegovic M, Dahm P. Screening for prostate cancer. Cochrane Database Syst Rev. 2013(1):CD004720.doi:https://doi.org.10.1002/14651858.CD004720.pub3
76. International Agency for Research on Cancer (IARC). GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide Geneva, Switzerland2013 [Third Edition:[Available from: https://publications.iarc.fr/Databases/Iarc-Cancerbases/GLOBOCAN-2012-Estimated-Cancer-Incidence-Mortality-And-Prevalence-Worldwide-In-2012-V1.0-2012] (Accessed Date: 25 April 2018).
77. International Agency for Research on Cancer (IARC). GLOBOCAN Cancer Fact Sheets: prostate cancer Geneva, Switzerland: International Agency for Research on Cancer (IARC). 2012 [Available from: https://gco.iarc.fr/today/data/pdf/fact-sheets/cancers/cancer-fact-sheets-19.pdf] (Accessed Date: 29 February 2018).
78. International Agency for Research on Cancer (IARC). GLOBOCAN 2018: Estimated cancer incidence, mortality and prevalence worldwide Geneva, Switzerland2018 [Third Edition:[Available from: https://gco.iarc.fr/today/online-analysis-map?v=2018&mode=cancer&mode_population=continents&population=900&populations=900&key=asr&sex=0&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=5&group_cancer=1&include_nmsc=1&include_nmsc_other=1&projection=natural-earth&color_palette=default&map_scale=quantile&map_nb_colors=5&continent=0&rotate=%255B10%252C0%255D] (Accessed Date: 25 April 2018).
79. Pilleron S, Soerjomataram I, Charvat H, Chokunonga E, Somdyala NIM, Wabinga H, et al. Cancer incidence in older adults in selected regions of sub-Saharan Africa, 2008-2012. Int J Cancer. 2019;144(8):1824-33.doi:https://doi.org.10.1002/ijc.31880
80. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2011 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD). 2011 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2011-NCR-2011-cancer-tables.pdf] (Accessed Date:
81. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in 2012 Sandringham, Johannesburg, South Africa: National Institute of Communicable Diseases (NICD). 2012 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2012-NCR-2012-results.pdf] (Accessed Date: 26 February 2018).
82. African Union (AU). Member States Addis Ababa, Ethopia: African Union (AU). 2020 [Available from: https://au.int/en/member_states/countryprofiles2] (Accessed Date: 21 September 2020).
83. Adeloye D, David RA, Aderemi AV, Iseolorunkanmi A, Oyedokun A, Iweala EE, et al. An Estimate of the Incidence of Prostate Cancer in Africa: A Systematic Review and Meta-Analysis. PLoS One. 2016;11(4):e0153496.doi:https://doi.org.10.1371/journal.pone.0153496
84. 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. South African Journal of Surgery. 2015;43(3-4):6.

85. Maphayi MR, Cassim N, Bello B, JAG G. Mining laboratory data to describe prostate specific antigen testing and prostate cancer in Johannesburg, South Africa. The Pan African Medical Journal. 2020;35(61).doi:https://doi.org.doi:10.11604/pamj.2020.35.61.21331
86. Statistics South Africa (STATS SA). Census 2011: Provinces at a glance Pretoria, South Africa: Statistics South Africa (STATS SA). 2011 [107]. Available from: http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf] (Accessed Date: 03-01-41).
87. Maphayi MR. Modalities of prostate specific antigen testing in Gauteng clinics and hospitals, South Africa [Dissertation]. Johannesburg, South Africa: University of the Witwatersrand. 2018 [49]. Available from: https://hdl.handle.net/10539/25288] (Accessed Date: 10 August 2018).
88. Statistics South Africa (STATS SA). Census 2011: Key Results Pretoria, South Africa: Statistics South Africa (STATS SA). 2011 [107]. Available from: http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Key_results.pdf] (Accessed Date: 23 January 2019).
89. National Institute for Communicable Diseases (NICD). National Cancer Registry (NCR) Methodology Sandringham, Johannesburg: National Institute for Communicable Diseases (NICD). 2017 [Available from: http://nicd.ac.za/?page=methodology&id=251] (Accessed Date: 28 February 2017).
90. Motsuku L, Chen WC, Muchengeti MM, Naidoo M, Quene TM, Kellett P, et al. Colorectal cancer incidence and mortality trends by sex and population group in South Africa: 2002–2014. BMC Cancer. 2021;21(1):129.doi:https://doi.org.10.1186/s12885-021-07853-1
91. Babb C, Urban M, Kielkowski D, Kellett P. Prostate cancer in South Africa: pathology based national cancer registry data (1986-2006) and mortality rates (1997-2009). Prostate Cancer. 2014;2014:419801.doi:https://doi.org.10.1155/2014/419801
92. National Cancer Registry. Summary statistics of cancer diagnosed histologically in South Africa in 2006 Johannesburg, South Africa: National Institute for Occupational Health (NIOH). 2006 [Available from: http://www.nicd.ac.za/wp-content/uploads/2019/12/2006-NCR_2006_TABLES_FINAL.pdf] (Accessed Date: 12 May 2018).
93. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2007 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD). 2007 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2007-NCR-2007-results.pdf] (Accessed Date: 12 May 2018).
94. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2012 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD). 2012 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2012-NCR-2012-results.pdf] (Accessed Date: 12 May 2018).
95. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2014 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD). 2013 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2013-NCR.pdf] (Accessed Date: 12 May 2018).
96. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2013 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD). 2013 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2013-NCR.pdf] (Accessed Date: 12 May 2018).
97. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in 2006 Sandringham, Johannesburg, South Africa: National Institute of Communicable Diseases (NICD). 2006 [Available from: http://www.nicd.ac.za/wp-content/uploads/2019/12/2006-NCR_2006_TABLES_FINAL.pdf] (Accessed Date: 26 February 2018).

98. Philip Chen CL, Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences. 2014;275:314-47.doi:https://doi.org.10.1016/j.ins.2014.01.015
99. Enterprise Big Data Frameworks. Data Types: Structured vs. Unstructured Data: Enterprise Big Data Frameworks. 2019 [Available from: https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/] (Accessed Date: 21 September 2020).
100. National Cancer Registry (NCR). Methodology: Data collection and data flow Sandringham, Johannesburg, South Africa: National Institute of Communicable Diseases (NICD). 2006 [Available from: https://www.nicd.ac.za/centres/national-cancer-registry/] (Accessed Date: 26 February 2018).
101. Perez JA, Poon CCY, Merrifield RD, Wong STC, Yang GZ. Big Data for Health. IEEE Journal of biomedical and health informatics. 2015;19(4).doi:https://doi.org.10.1109/JBHI.2015.2450362
102. Kitchin R, McArdle G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society. 2016;1.doi:https://doi.org.10.1177/2053951716631130
103. Benke K, Benke G. Artificial Intelligence and Big Data in Public Health. Int J Environ Res Public Health. 2018;15(12).doi:https://doi.org.10.3390/ijerph15122796
104. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. Biomedical text mining and its applications in cancer research. J Biomed Inform. 2013;46(2):200-11.doi:https://doi.org.10.1016/j.jbi.2012.10.007
105. Zhu F, Liu Q, Wang H, Zhou X, Fu Y. Unregistered Biological Words Recognition by Q-Learning with Transfer Learning. The Scientific World Journal. 2014;2014:173290.doi:https://doi.org.10.1155/2014/173290
106. Chowdhury GG. Natural language processing. Annual Review of Information Science and Technology. 2003;37(1):51-89.doi:https://doi.org.https://doi.org/10.1002/aris.1440370103
107. Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015;349(6245):261-6.doi:https://doi.org.10.1126/science.aaa8685
108. Dudhabaware RS, Madankar MS, editors. Review on natural language processing tasks for text documents. 2014 IEEE International Conference on Computational Intelligence and Computing Research; 2014 18-20 Dec. 2014.
109. World Health Organisation (WHO). National Cancer Control Programmes (NCCP) Geneva, Switzerland: World Health Organisation (WHO). 2002 [210]. Available from: https://apps.who.int/iris/bitstream/handle/10665/42494/9241545577.pdf;jsessionid=7B75F2254A4A5DC71445907B939A1709?sequence=1] (Accessed Date: 12 May 2017).
110. World Health Organisation (WHO). National Cancer Control Programme : Policies and managerial guidelines Geneva, Switzerland: World Health Organisation (WHO). 2002 [Available from: http://apps.who.int/iris/bitstream/handle/10665/42494/9241545577.pdf?sequence=1] (Accessed Date: 1 May 2016).
111. World Health Organisation (WHO). Global Action Plan for the Prevention and Control of NCDs 2013-2020 Geneva, Switzerland: World Health Organisation (WHO). 2013 [Available from: http://apps.who.int/iris/bitstream/10665/94384/1/9789241506236_eng.pdf?ua=1] (Accessed Date: 1 May 2016).
112. U. S. National Institutes of Health (NIH) NCIN. SEER Training Modules: Cancer Registry Bethesda, MD, USA: U. S. National Institutes of Health (NIH), National Cancer

Institute (NCI). 2018 [Available from: https://training.seer.cancer.gov/registration/registry/] (Accessed Date: 9 May 2018).
113. Centers for Disease Control and Prevention (CDC) NCfCDPaHP, Division of Cancer Prevention and Control, National Program of Cancer Registries,. National Program of Cancer Registries: NPCR Standards USA: Centers for Disease Control and Prevention (CDC). 1999 [Available from: https://www.cdc.gov/cancer/npcr/standards.htm] (Accessed Date: 9 May 2018).
114. 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.doi:https://doi.org.10.1016/S1470-2045(14)70310-9
115. Schonfeld SJ, Erdmann F, Wiggill T, Singh E, Kellett P, Babb C, et al. Hematologic malignancies in South Africa 2000-2006: analysis of data reported to the National Cancer Registry. Cancer Med. 2016;5(4):728-38.doi:https://doi.org.10.1002/cam4.597
116. International Agency for Research on Cancer (IARC), World Health Organisation (WHO). International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) Geneva, Switzerland2013 [Third Edition:[Available from: https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology] (Accessed Date: 25 April 2018).
117. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2014 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD). 2014 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2014-NCR-tables.pdf] (Accessed Date: 12 September 2018).
118. Thomas AA, Zheng C, Jung H, Chang A, Kim B, Gelfond J, 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.doi:https://doi.org.10.1007/s00345-013-1040-4
119. Banerjee I, Li K, Seneviratne M, Ferrari M, Seto T, Brooks JD, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. JAMIA Open. 2019;2(1):150-9.doi:https://doi.org.10.1093/jamiaopen/ooy057
120. Leyh-Bannurah SR, Tian Z, Karakiewicz PI, Wolffgang U, Sauter G, Fisch M, 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;2:1-9.doi:https://doi.org.10.1200/CCI.18.00080
121. SNOMED International. SNOMED CT Starter Guide London, UK: SNOMED International. 2017 [updated 2017-07-28. Available from: https://confluence.ihtsdotools.org/display/DOCSTART?preview=/28742871/47677485/doc_StarterGuide_Current-en-US_INT_20170728.pdf] (Accessed Date: 27 March 2018).
122. Vreede H. TrakCare SNOMED Topography and Morphology Code Tables. In: Cassim N, editor. Cape Town, South Africa: National Health Laboratory Service (NHLS); 2017.
123. Helgstrand JT, Klemann N, Roder MA, Toft BG, Brasso K, Vainer B, et al. Danish Prostate Cancer Registry - methodology and early results from a novel national database. Clin Epidemiol. 2016;8:351-60.doi:https://doi.org.10.2147/CLEP.S114917
124. IBM Netezza. Netezza Aginity Workbench Marlborough, Massachusetts, United States: IBM Netezza. 2018 [Netezza Aginity Workbench]. Available from: https://www.aginity.com/documentation/WB/NZ/Default.htm] (Accessed Date: 12 December 2018).
125. 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

[Available from: https://elibrary.worldbank.org/doi/abs/10.1596/25399] (Accessed Date: 24 April 2018).
126. Microsoft Corporation. Microsoft Office Professional Plus 2013. Microsoft Office Professional Plus 2013 ed. Redmont, Washington, United States of America: Microsoft Corporation; 2013. p. Microsoft Office Professional Plus 2013.
127. Microsoft Corporation: TechNet. Using Outer Joins Albuquerque, New Mexico, USA: Microsoft Corporation. 2018 [Available from: https://technet.microsoft.com/en-us/library/ms187518(v=sql.105).aspx] (Accessed Date: 10 April 2018).
128. Microsoft Corporation. Guide to table relationships Redmont, Washington, USA: Microsoft Corporation. 2018 [Available from: https://support.office.com/en-us/article/guide-to-table-relationships-55b8db2c-9480-4269-b1bb-f6ec09623dfd] (Accessed Date: 12 May 2018).
129. SAS Institute Inc. SAS Enterprise Guide Cary, North Carolina, United States of Americal: SAS Institute Inc. 2017 [7.15 HF2 (7.100.5.6112) (64-bit):[12 March 2019]. Available from: http://support.sas.com/software/products/enterprise-guide/index.html] (Accessed Date: 2019).
130. International Agency for Research on Cancer (IARC). GLOBOCAN 2020: Cancer Fact Sheets: South Africa Geneva, Switzerland: International Agency for Research on Cancer (IARC). 2021 [Available from: https://gco.iarc.fr/today/data/factsheets/populations/710-south-africa-fact-sheets.pdf] (Accessed Date: 7 June 2021).
131. National Health Laboratory Service (NHLS). Annual Report 2015/16 Johannesburg, South Africa: National Health Laboratory Service. 2016 [Available from: https://static.pmg.org.za/NHLS_Annual_Report_2015-2016.pdf] (Accessed Date: 23 April 2018).
132. Rutherford GW, McFarland W, Spindler H, White K, Patel SV, Aberle-Grasse J, et al. Public health triangulation: approach and application to synthesizing data to understand national and local HIV epidemics. BMC Public Health. 2010;10:447.doi:https://doi.org.10.1186/1471-2458-10-447
133. 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. South African Medical Journal 2020;111(1).doi:https://doi.org.10.7196/SAMJ.2020.v111i1.14850
134. Kufa-Chakezha T, Gita GD, Ballah NJ, Puren A, Takuva S, Carmona S, et al. Determinants of CD4 immune recovery among individuals on antiretroviral therapy in South Africa : a national analysis Washington, D.C. : World Bank Group. 2016 [62]. Available from: http://documents.worldbank.org/curated/en/851301474884707261/Determinants-of-CD4-immune-recovery-among-individuals-on-antiretroviral-therapy-in-South-Africa-a-national-analysis] (Accessed Date: 12 January 2020).
135. 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):174-84.doi:https://doi.org.10.2471/BLT.15.164384
136. 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):507-13.doi:https://doi.org.10.7196/samj.2017.v107i6.11311
137. Nanoo A, Izu A, Ismail NA, Ihekweazu C, Abubakar I, Mametja D, 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):1066-76.doi:https://doi.org.10.1016/S1473-3099(15)00147-4

138. 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):1487-92.doi:https://doi.org.10.1007/s00405-019-05381-0
139. Kobayashi S, Karube Y, Inoue T, Araki O, Maeda S, Matsumura Y, 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.doi:https://doi.org.10.5761/atcs.oa.18-00158
140. Sfanos KS, De Marzo AM. Prostate cancer and inflammation: the evidence. Histopathology. 2012;60(1):199-215.doi:https://doi.org.10.1111/j.1365-2559.2011.04033.x
141. Leitzmann MF, Rohrmann S. Risk factors for the onset of prostatic cancer: age, location, and behavioral correlates. Clin Epidemiol. 2012;4:1-11.doi:https://doi.org.10.2147/CLEP.S16747
142. Nelson WG, De Marzo AM, Isaacs WB. Prostate cancer. N Engl J Med. 2003;349(4):366-81.doi:https://doi.org.10.1056/NEJMra021562
143. Coetzee L-M, Cassim N, Glencross DK. Using laboratory data to categorise CD4 laboratory turn-around-time performance across a national programme. 2018. 2018;7(1).doi:https://doi.org.10.4102/ajlm.v7i1.665
144. 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, USA: World Bank. 2015 [updated 2016-11-21. Available from: https://elibrary.worldbank.org/doi/abs/10.1596/25399] (Accessed Date: 12 March 2020).
145. 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 Africa2018.
146. Ngobeni T. Corporate Data Warehouse lookup tables. Sandringham, Johannesburg, South Africa: National Health Laboratory Service (NHLS); 2017.
147. Sepulveda JL, Young DS. The ideal laboratory information system. Arch Pathol Lab Med. 2013;137(8):1129-40.doi:https://doi.org.10.5858/arpa.2012-0362-RA
148. Stevens WS, Cunningham B, Cassim N, Gous N, Scott LE. Cloud-Based Surveillance, Connectivity, and Distribution of the GeneXpert Analyzers for Diagnosis of Tuberculosis (TB) and Multiple-Drug-Resistant TB in South Africa. Molecular Microbiology: American Society of Microbiology; 2016.doi:https://doi.org.doi:https://doi.org/10.1128/9781555819071.ch49
149. Spasic I, Livsey J, Keane JA, Nenadic G. Text mining of cancer-related information: review of current status and future directions. Int J Med Inform. 2014;83(9):605-23.doi:https://doi.org.10.1016/j.ijmedinf.2014.06.009
150. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2014 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD). 2013 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2013-NCR.pdf] (Accessed Date: 12 May 2020).
151. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011;18(5):544-51.doi:https://doi.org.10.1136/amiajnl-2011-000464
152. Seaborn. seaborn: statistical data visualization 2018 [Available from: https://seaborn.pydata.org/] (Accessed Date: 17 December 2018).

153. The Spyder Website Contributors Spyder: The Scientific Python Development Environment: The Scientific Python Development Environment. 2018 [Available from: https://www.spyder-ide.org/] (Accessed Date: 23 January 2018).
154. Cassim N, Ahmad A, Wadee R, Glencross DK, George JA. Using Systematized Nomenclature of Medicine (SNOMED) code to assign histological findings for prostate biopsies in the Gauteng province, South Africa: Lessons learnt. Afr J Lab Med. 2020;9(1).doi:https://doi.org.10.4102/ajlm.v9i1.909
155. Microsoft Corporation. Microsoft Office Professional Plus 2013 Redmont, Washington, United States of America: Microsoft Corporation. 2013 [Microsoft Office Professional Plus 2013:[Microsoft Office Professional Plus 2013]. Available from: https://www.microsoft.com/en-us/download/details.aspx?id=42971] (Accessed Date: 12 January 2018).
156. Linguamatics. What is NLP Text Mining? Cambridge, UK: Linguamatics. 2018 [Available from: https://www.linguamatics.com/what-is-text-mining-nlp-machine-learning] (Accessed Date: 17 December 2018).
157. The Matplotlib development team. Matplotlib: The Matplotlib development team. 2018 [Available from: https://matplotlib.org/] (Accessed Date: 17 December 2018).
158. Napolitano G, Fox C, Middleton R, Connolly D. Pattern-based information extraction from pathology reports for cancer registration. Cancer Causes Control. 2010;21(11):1887-94.doi:https://doi.org.10.1007/s10552-010-9616-4
159. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Classification of forensic autopsy reports through conceptual graph-based document representation model. J Biomed Inform. 2018;82:88-105.doi:https://doi.org.10.1016/j.jbi.2018.04.013
160. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K. Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. Journal of Forensic and Legal Medicine. 2018;57:41-50.doi:https://doi.org.https://doi.org/10.1016/j.jflm.2017.07.001
161. Sorzano COS, Vargas J, AP M. A survey of dimensionality reduction techniques based on random projection Ithaca, New York: Cornell University. 2014 [Available from: https://arxiv.org/abs/1403.2877] (Accessed Date: 3 September 2020).
162. Maria Navin J R, R P. Performance Analysis of Text Classification Algorithms using Confusion Matrix. International Journal of Engineering and Technical Research (IJETR). 2016;6(4):75-8.
163. Shmueli B. Multi-Class Metrics Made Simple, Part II: the F1-score: Towards Data Science. 2019 [Available from: https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1] (Accessed Date: 3 September 2020).
164. Business Tech. The astonishing number of South Africans who do not have medical aid Pretoria, South Africa: Business Tech. 2016 [updated 5 July 2016. Available from: https://businesstech.co.za/news/lifestyle/129166/the-shocking-number-of-south-africans-who-do-not-have-medical-aid/] (Accessed Date: 28 May 2018).
165. Spyder Project Contributors. Spyder Python Integrated Development Environment (IDE) 2018 [Available from: https://www.spyder-ide.org/] (Accessed Date: 31 October 2018).
166. Walke VA, Gunjkar G. Comparative evaluation of six parametric Robinson and three parametric Howell's modification of Scarf-BloomRichardson grading method on breast aspirates with histopathology: A prospective study. Cytojournal. 2017;14:31.doi:https://doi.org.10.4103/cytojournal.cytojournal_31_17
167. Holzinger A. Machine Learning for Health Informatics: State-of-the-Art and Future Challenges. Holzinger A, editor. New York, NY, USA: Springer; 2016.
168. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.doi:https://doi.org.10.1038/nature14539

169. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nature Medicine. 2019;25(1):24-9.doi:https://doi.org.10.1038/s41591-018-0316-z
170. Khan F, Khan MA, Abbas S, Athar A, Siddiqui SY, Khan AH, et al. Cloud-Based Breast Cancer Prediction Empowered with Soft Computing Approaches. J Healthc Eng. 2020;2020:8017496.doi:https://doi.org.10.1155/2020/8017496
171. Maphayi M, Cassim N, Bello B, George JA. Mining laboratory data to describe prostate specific antigen testing and prostate cancer in Johannesburg, South Africa. Pan African Medical Journal. 2020;35.doi:https://doi.org.doi:10.11604/pamj.2020.35.61.21331
172. Odedina FT, Akinremi TO, Chinegwundoh F, Roberts R, Yu D, Reams RR, et al. Prostate cancer disparities in Black men of African descent: a comparative literature review of prostate cancer burden among Black men in the United States, Caribbean, United Kingdom, and West Africa. Infect Agent Cancer. 2009;4 Suppl 1:S2.doi:https://doi.org.10.1186/1750-9378-4-S1-S2
173. Rebbeck TR, Haas GP. Temporal trends and racial disparities in global prostate cancer prevalence. Can J Urol. 2014;21(5):7496-506.
174. Pietro GD, Chornokur G, Kumar NB, Davis C, Park JY. Racial Differences in the Diagnosis and Treatment of Prostate Cancer. Int Neurourol J. 2016;20(Suppl 2):S112-9.doi:https://doi.org.10.5213/inj.1632722.361
175. Roche Diagnostics. Total PSA kit insert: total (free + complexed) PSA - Prostate-specific antigen (tPSA) Risch-Rotkreuz, Switzerland: Roche Diagnostics. 2010 [Available from: https://www.who.int/medical_devices/diagnostics/selection_in-vitro/selection_in-vitro-meetings/00029_06_FDAclearance.pdf] (Accessed Date: 8 August 2018).
176. National Department of Health (NDOH). Integrated Clinical Services Management (ICSM) National Department of Health (NDOH): Pretoria, South Africa. 2017 [updated 27 June 2017. Available from: https://www.knowledgehub.org.za/system/files/elibdownloads/2019-07/Integrated%2520Clinical%2520Services%2520Management%2520%2520Manual%25205th%2520June%2520FINAL.pdf] (Accessed Date: 16 March 2020).
177. Kwazulu-Natal Procincial Department of Health. Definitions of health facilities Pietermaritzberg. Kwazulu-Natal province, South Africa: Kwazulu-Natal Procincial Departmenr of Health. 2001 [Available from: http://www.kznhealth.gov.za/definitions.htm] (Accessed Date: 16 March 2020).
178. Singh E, Greyling, M. A review of hot deck imputation for survey non responses: The example of population group in the national cancer registry database. [Research Article]. In press 2015.
179. Chen W, Kellett P, Greyling M, Sengayi M, E S. The use of surnames to impute missing ethnicity data in the National Cancer Registry database of South Africa. [Research Article]. In press 2019.
180. Andridge RR, Little RJ. A Review of Hot Deck Imputation for Survey Non-response. Int Stat Rev. 2010;78(1):40-64.doi:https://doi.org.10.1111/j.1751-5823.2010.00103.x
181. Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies. PLoS One. 2015;10(8):e0136179.doi:https://doi.org.10.1371/journal.pone.0136179
182. Tang L, Song J, Belin TR, Unutzer J. A comparison of imputation methods in a longitudinal randomized clinical trial. Stat Med. 2005;24(14):2111-28.doi:https://doi.org.10.1002/sim.2099
183. Bassett IV, Huang M, Cloete C, Candy S, Giddy J, Frank SC, et al. Using national laboratory data to assess cumulative frequency of linkage after transfer to community-based

HIV clinics in South Africa. J Int AIDS Soc. 2019;22(6):e25326-e.doi:https://doi.org.10.1002/jia2.25326
184. Capitanio U, Perrotte P, Zini L, Suardi N, Antebi E, Cloutier V, et al. Population-based analysis of normal Total PSA and percentage of free/Total PSA values: results from screening cohort. Urology. 2009;73(6):1323-7.doi:https://doi.org.10.1016/j.urology.2008.10.026
185. Mottet N, van den Bergh RCN, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, et al. EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer—2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. European Urology. 2021;79(2):243-62.doi:https://doi.org.https://doi.org/10.1016/j.eururo.2020.09.042
186. Cohn JA, Wang CE, Lakeman JC, Silverstein JC, Brendler CB, Novakovic KR, et al. Primary care physician PSA screening practices before and after the final U.S. Preventive Services Task Force recommendation. Urologic Oncology: Seminars and Original Investigations. 2014;32(1):41.e23-41.e30.doi:https://doi.org.10.1016/j.urolonc.2013.04.013
187. Filella X, Albaladejo MD, Allue JA, Castano MA, Morell-Garcia D, Ruiz MA, et al. Prostate cancer screening: guidelines review and laboratory issues. Clin Chem Lab Med. 2019;57(10):1474-87.doi:https://doi.org.10.1515/cclm-2018-1252
188. Eapen RS, Herlemann A, Washington SLI, Cooperberg MR. Impact of the United States Preventive Services Task Force 'D' recommendation on prostate cancer screening and staging. Current Opinion in Urology. 2017;27(3):205-9.doi:https://doi.org.10.1097/mou.0000000000000383
189. Radomski TR, Huang Y, Park SY, Sileanu FE, Thorpe CT, Thorpe JM, et al. Low-Value Prostate Cancer Screening Among Older Men Within the Veterans Health Administration. Journal of the American Geriatrics Society. 2019;67(9):1922-7.doi:https://doi.org.https://doi.org/10.1111/jgs.16057
190. Mafi JN RK, Bortz BA, Dachary M,HazelJr WA., and Fendrick AM. Low-Cost, High-Volume Health Services Contribute The Most To Unnecessary Health Spending. Health Affairs. 2017;36(10):1701-4.doi:https://doi.org.10.1377/hlthaff.2017.0385
191. Walter LC, Bertenthal D, Lindquist K, Konety BR. PSA Screening Among Elderly Men With Limited Life Expectancies. JAMA. 2006;296(19):2336-42.doi:https://doi.org.10.1001/jama.296.19.2336
192. National Department of Health (NDOH). Ideal Clinic Manual Pretoria, South Africa: National Department of Health (NDOH). 2016 [updated 27 February 2017. 264]. Available from: https://www.idealhealthfacility.org.za/App/Document/Download/204] (Accessed Date: 12 January 2020).
193. Van Schalkwyk M, Andersson MI, Zeier MD, La Grange M, Taljaard JJ, Theron GB. The impact of revised PMTCT guidelines: a view from a public sector ARV clinic in Cape Town, South Africa. J Acquir Immune Defic Syndr. 2013;63(2):234-8.doi:https://doi.org.10.1097/QAI.0b013e31828bb721
194. Johnson LF, Dorrington RE, Moolla H. Progress towards the 2020 targets for HIV diagnosis and antiretroviral treatment in South Africa. Southern African Journal of HIV Medicine. 2017;18(1):8.doi:https://doi.org.10.4102/sajhivmed.v18i1.694
195. Gauteng Department of Health (GDOH) WRHIW, Anova Health Institute,, Right To Care (RTC) TEF, Community AIDS Response (CARe),. Using NHLS Results for Action Reports: A data driven strategy to improve linkage to care Johannesburg, South Africa: Gauteng Department of Health. 2017 [updated 24 October 2017. The Wits Reproductive Health and HIV Institute
and the Paediatric Adolescent Scale-Up plan (PASP)]. Available from: https://teampata.org/wp-content/uploads/2017/11/JackieDunlop_RighttoCareSA_UsingNHLSResultsforActionReports.pdf] (Accessed Date: 29 September 2019).

196. Rebbeck TR, Devesa SS, Chang BL, Bunker CH, Cheng I, Cooney K, et al. Global patterns of prostate cancer incidence, aggressiveness, and mortality in men of african descent. Prostate Cancer. 2013;2013:560857.doi:https://doi.org.10.1155/2013/560857
197. National Health Laboratory Service (NHLS). 2017/18 Annual Report Johannesburg, South Africa: National Health Laboratory Service (NHLS). 2018 [Available from: https://nationalgovernment.co.za/entity_annual/1714/2018-national-health-laboratory-service-(nhls)-annual-report.pdf] (Accessed Date: 18 July 2019).
198. Nikiema JN, Jouhet V, Mougin F. Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts. J Biomed Inform. 2017;74:46-58.doi:https://doi.org.10.1016/j.jbi.2017.08.013
199. Statistics South Africa (STATS SA). Mid-year population estimates Pretoria, South Africa: Statistics South Africa (STATS SA). 2016 [P0302
[17]. Available from: http://www.statssa.gov.za/?page_id=1854&PPN=P0302] (Accessed Date: 12 February 2018).
200. Statistics South Africa (STATS SA). Mid-year population estimates Pretoria, South Africa: Statistics South Africa (STATS SA). 2018 [Available from: http://www.statssa.gov.za/?page_id=1854&PPN=P0302] (Accessed Date: 23 July 2018).
201. Oosthuizen M, Naidoo P. Internal Migration to the Gauteng Province Cape Town, South Africa: Development Policy Research Unit, University of Cape Town. 2004 [Available from: https://www.africaportal.org/publications/internal-migration-to-the-gauteng-province/] (Accessed Date: 7 February 2019).
202. Cohen C. The Use and Impacts of a Corporate Data Darehouse: The Case of the National Health Laboratory Service Johannesburg, South Africa2018 [Available from: http://wiredspace.wits.ac.za/bitstream/handle/10539/27827/Masters%20Report%20Final%20-%20with%20Final%20Revisions.pdf?sequence=1&isAllowed=y] (Accessed Date: 11 January 2020).
203. SAS Institute Inc. SAS Enterprise Guide. 7.15 HF2 (7.100.5.6112) (64-bit) ed. Cary, North Carolina, United States of Americal: SAS Institute Inc; 2017.
204. StataCorp. Stata 12 SE. SE ed. 4905 Lakeway Drive, College Station, Texas, USA: Stata Corp. p. Stata 12 SE.
205. Statistical Research and Applications Branch NCI. Joinpoint Trend Analysis Software. 4.6.0.0 ed. Rockville, MD, USA: National Cancer Institute; 2018. p. Joinpoint Regression Program.
206. Ahmed OB, Boschi-Pinto C, Lopez AD, Murray CJL, Lozano R, Inoue M. Age Standardization of Rates: A new WHO standard Geneva, Switzerland: World Health Organisation (WHO. 2001 [Available from: http://www.who.int/healthinfo/paper31.pdf] (Accessed Date: 4 February 2018).
207. Statistics South Africa (STATS SA). Mid-year population estimates Pretoria, South Africa: Statistics South Africa (STATS SA). 2017 [Available from: https://www.statssa.gov.za/publications/P0302/P03022017.pdf] (Accessed Date: 13 June 2017).
208. Gauteng Department of Health (GDOH). Annual Report: 2013/14 Johannesburg, South Africa: Gauteng Department of Health (GDOH). 2017 [Available from: https://cmbinary.gauteng.gov.za/Media?path=cphealth/Documents/Annual%20Reports/Department%20of%20Health%20Annual%20Report%202013-2014.pdf&Item=330&Type=Documents&Location=/cphealth] (Accessed Date: 11 March 2019).
209. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2013 Johannesburg, South Africa: National Institute for Communicable

Diseases (NICD). 2013 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2013-NCR.pdf] (Accessed Date: 12 March 2017).
210. National Cancer Registry (NCR). Summary statistics of cancer diagnosed histologically in South Africa in 2012 Johannesburg, South Africa: National Institute for Communicable Diseases (NICD). 2012 [Available from: https://www.nicd.ac.za/wp-content/uploads/2019/12/2012-NCR-2012-results.pdf] (Accessed Date:
211. Cassim N, Ahmad A, Glencross DK, George JA. Trends in prostate specific antigen (PSA) testing for primary health care facilities in the Gauteng province, South Africa, between 2006 and 2016 [Article]. In press 2021.
212. Mndzebele SL, Mogale JW. Prostate cancer awareness and screening behaviour among men in Gauteng Province, South Africa : utilization of primary health care services. African Journal for Physical Health Education, Recreation and Dance. 2015;21(Supplement 2):296-305.doi:https://doi.org.10.1186/s40064-015-0824-y
213. Pink Drive. Pink Drive: Its moving 2016 [Available from: http://www.pinkdrive.co.za/] (Accessed Date: 30 August 2016).
214. The Men's Foundation. Movember Cape Town, South Africa: The Men's Foundation. 2018 [Available from: https://za.movember.com/] (Accessed Date: 28 August 2018).
215. Hollard. Hollard Dare Devil Run Johannesburg, South Africa: Hollard. 2018 [Available from: https://www.daredevilrun.com/] (Accessed Date: 28 August 2018).
216. Prostate Cancer Foundation (PCF). Suit Up September - Prostate Cancer Awareness Campaign Johannesburg, South Africa: Prostate Cancer Foundation (PCF). 2017 [Available from: https://prostate-ca.co.za/suit-up-september/] (Accessed Date: 28 August 2018).
217. Ilic D, Djulbegovic M, Jung JH, Hwang EC, Zhou Q, Cleves A, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. BMJ. 2018;362:k3519.doi:https://doi.org.10.1136/bmj.k3519
218. Thompson IM, Ankerst DP. Prostate-specific antigen in the early detection of prostate cancer. CMAJ. 2007;176(13):1853-8.doi:https://doi.org.10.1503/cmaj.060955
219. Abera Abaerei A, Ncayiyana J, Levin J. Health-care utilization and associated factors in Gauteng province, South Africa. Glob Health Action. 2017;10(1):1305765.doi:https://doi.org.10.1080/16549716.2017.1305765
220. Baker P, Shand T. Men's health: time for a new approach to policy and practice? Journal of global health. 2017;7(1).doi:https://doi.org.10.7189/jogh.07.010306
221. Bechis SK, Carroll PR, Cooperberg MR. Impact of Age at Diagnosis on Prostate Cancer Treatment and Survival. J Clin Oncol. 2011;29(2):235-41.doi:https://doi.org.10.1200/jco.2010.30.2075
222. Alibhai SMH, Krahn MD, Fleshner NE, Cohen MM, Tomlinson GA, Naglie G. The association between patient age and prostate cancer stage and grade at diagnosis. BJU International. 2004;94(3):303-6.doi:https://doi.org.https://doi.org/10.1200/jco.2010.30.2075
223. Muhammad A, Abdulwahab-Ahmed A, Agwu N, Yunusa G, Abdullahi K, Mungadi I. Use of transrectal ultrasound in the management of urologic diseases in a tertiary hospital in North-Western Nigeria. 2018;25(3):172-6.doi:https://doi.org.10.4103/npmj.npmj_131_18
224. Yeboah ED, Hsing AW, Mante S, Mensah JE, Kyei MY, Yarney J, et al. Management of Prostate Cancer in Accra, Ghana. J West Afr Coll Surg. 2016;6(4):31-65.
225. Cassim N, Mapundu M, Olago V, Ahmad A, George JA, Glencross DK. Using text mining techniques to extract prostate cancer predictive information (Gleason score) from semi-structured narrative laboratory reports in the Gauteng province, South Africa. [Article]. In press 2021.
226. Spackman K. SNOMED® CT Unlocks the Power of Clinical Data for Pathologists. Laboratory Medicine. 2002;33(4):263-6.doi:https://doi.org.10.1309/f9c2-uj61-mth4-38mh

227. Dahl A, Özkan A, Dalianis H, editors. Pathology text mining - on Norwegian prostate cancer reports. 2016 IEEE 32nd International Conference on Data Engineering Workshops (ICDEW); 2016 16-20 May 2016.
228. Heyns CF, Mathee S, Isaacs A, Kharwa A, De Beer PM, Pretorius MA. Problems with prostate specific antigen screening for prostate cancer in the primary healthcare setting in South Africa. BJU Int. 2003;91(9):785-8.doi:https://doi.org.10.1046/j.1464-410x.2003.04241.x
229. Droz JP, Balducci L, Bolla M, Emberton M, Fitzpatrick JM, Joniau S, et al. Management of prostate cancer in older men: recommendations of a working group of the International Society of Geriatric Oncology. BJU Int. 2010;106(4):462-9.doi:https://doi.org.10.1111/j.1464-410X.2010.09334.x
230. Statistics South Africa (STATS SA). Mid-year population estimates Pretoria, South Africa: Statistics South Africa (STATS SA). 2016 [17]. Available from: http://www.statssa.gov.za/?page_id=1854&PPN=P0302] (Accessed Date: P0302).
231. Dabrowski JM. Development of pesticide use maps for South Africa. South African Journal of Science. 2015;111:07-.doi:https://doi.org.10.17159/sajs.2015/20140091
232. Cadet J-M, Bencherif H, Portafaix T, Lamy K, Ncongwane K, Coetzee GJR, et al. Comparison of Ground-Based and Satellite-Derived Solar UV Index Levels at Six South African Sites. International journal of environmental research and public health. 2017;14(11):1384.doi:https://doi.org.10.3390/ijerph14111384
233. Moore AL, Dimitropoulou P, Lane A, Powell PH, Greenberg DC, Brown CH, et al. Population-based prostate-specific antigen testing in the UK leads to a stage migration of prostate cancer. BJU Int. 2009;104(11):1592-8.doi:https://doi.org.10.1111/j.1464-410X.2009.08652.x
234. Carmona S, Bor J, Nattey C, Maughan-Brown B, Maskew M, Fox MP, et al. Persistent High Burden of Advanced HIV Disease Among Patients Seeking Care in South Africa's National HIV Program: Data From a Nationwide Laboratory Cohort. Clinical Infectious Diseases. 2018;66(suppl_2):S111-S7.doi:https://doi.org.10.1093/cid/ciy045
235. Glencross DK, Cassim N, Coetzee LM. Documented higher burden of advanced and very advanced HIV disease among patients, especially men, accessing healthcare in a rapidly growing economic and industrial hub in South Africa: A call to action. South African Medical Journal. 2020;110(6):9.doi:https://doi.org.10.7196/SAMJ.2020.v110i6.14352
236. Leichliter JS, Paz-Bailey G, Friedman AL, Habel MA, Vezi A, Sello M, et al. 'Clinics aren't meant for men': Sexual health care access and seeking behaviours among men in Gauteng province, South Africa. SAHARA-J: Journal of Social Aspects of HIV/AIDS. 2011;8(2):82-8.doi:https://doi.org.10.1080/17290376.2011.9724989
237. Treves-Kagan S, El Ayadi AM, Pettifor A, MacPhail C, Twine R, Maman S, et al. Gender, HIV Testing and Stigma: The Association of HIV Testing Behaviors and Community-Level and Individual-Level Stigma in Rural South Africa Differ for Men and Women. AIDS and Behavior. 2017;21(9):2579-88.doi:https://doi.org.10.1007/s10461-016-1671-8
238. Fleming PJ, DiClemente RJ, Barrington C. Masculinity and HIV: Dimensions of Masculine Norms that Contribute to Men's HIV-Related Sexual Behaviors. AIDS and Behavior. 2016;20(4):788-98.doi:https://doi.org.10.1007/s10461-015-1264-y
239. National Department of Health (NDOH). National consolidated Guidelines for the prevention of mother-to-child transmission of HIV (PMTCT) and the management of HIV in children, adolescents and adults Pretoria, South Africa: National Department of Health (NDOH). 2015 [128]. Available from: http://www.sahivsoc.org/Files/ART%20Guidelines%2015052015.pdf] (Accessed Date: 12 April 2017).
240. Gandaglia G, Albers P, Abrahamsson P-A, Briganti A, Catto JWF, Chapple CR, et al. Structured Population-based Prostate-specific Antigen Screening for Prostate Cancer: The

European Association of Urology Position in 2019. European Urology. 2019;76(2):142-50.doi:https://doi.org.https://doi.org/10.1016/j.eururo.2019.04.033
241. Business Tech. The astonishing number of South Africans who do not have medical aid Lyttelton, Pretoria, South Africa: Business Tech. 2016 [updated 5 July 2016. Available from: https://businesstech.co.za/news/lifestyle/129166/the-shocking-number-of-south-africans-who-do-not-have-medical-aid/] (Accessed Date: 24 August 2018).
242. Snell R. Cape Slave Naming Patterns: South African History Online. 2011 [Available from: https://www.sahistory.org.za/archive/cape-slave-naming-patterns-robert-shell] (Accessed Date: 12 March 2018).
243. Singh E, Underwood JM, Nattey C, Babb C, Sengayi M, Kellett P. South African National Cancer Registry: Effect of withheld data from private health systems on cancer incidence estimates. South African Medical Journal. 2015;105(2):107-9.doi:https://doi.org.https://doi.org/10.7196/samj.8858
244. Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Medical Informatics and Decision Making. 2018;18(3):74.doi:https://doi.org.10.1186/s12911-018-0654-2
245. Bhardwaj S, Barron P, Pillay Y, Treger-Slavin L, Robinson P, Goga A, et al. Elimination of mother-to-child transmission of HIV in South Africa: rapid scale-up using quality improvement. S Afr Med J. 2014;104(3 Suppl 1):239-43.doi:https://doi.org.10.7196/samj.7605
246. Gous NM, Berrie L, Dabula P, Stevens W. South Africa's experience with provision of quality HIV diagnostic services. Afr J Lab Med. 2016;5(2):436.doi:https://doi.org.10.4102/ajlm.v5i2.436
247. Sidibé M, Loures L, Samb B. The UNAIDS 90-90-90 target: a clear choice for ending AIDS and for sustainable health and development. J Int AIDS Soc. 2016;19(1):21133-.doi:https://doi.org.10.7448/IAS.19.1.21133

APPENDIX A – PUBLICATION

Using Systematized Nomenclature of Medicine (SNOMED) clinical terms (CT) codes to assign histological findings for prostate biopsies in the Gauteng Province, South Africa: Lessons learnt

Cassim N, Ahmad A, Wadee R, George JA and Glencross DK.

Afr J Lab. 2020; 9(1):1-9 https://doi.org/10.4102/ajlm.v9i1.909

African Journal of Laboratory Medicine
ISSN: (Online) 2225-2010, (Print) 2225-2002

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 Ahmad[3]
Reubina Wadee[4]
Jaya A. George[5]
Deborah K. Glencross[1,2]

Affiliations:
[1]National Health Laboratory Service(NHLS), National Priority Programme, Johannesburg, South Africa

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

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

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

[5]Department 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 timeous 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.

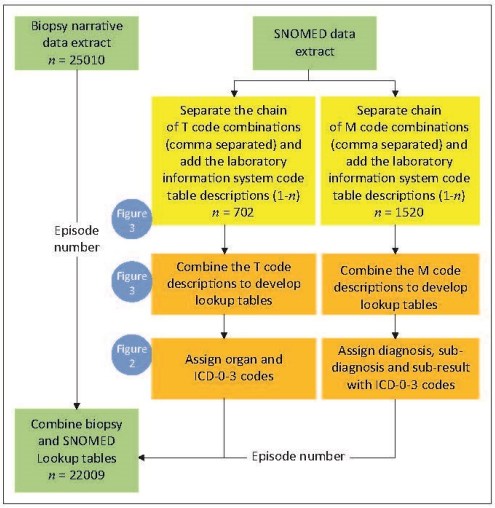
Data extraction and preparation

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 algorithm[24]), (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 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).



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.

Combining Systemized Nomenclature of Medicine descriptions for the lookup table

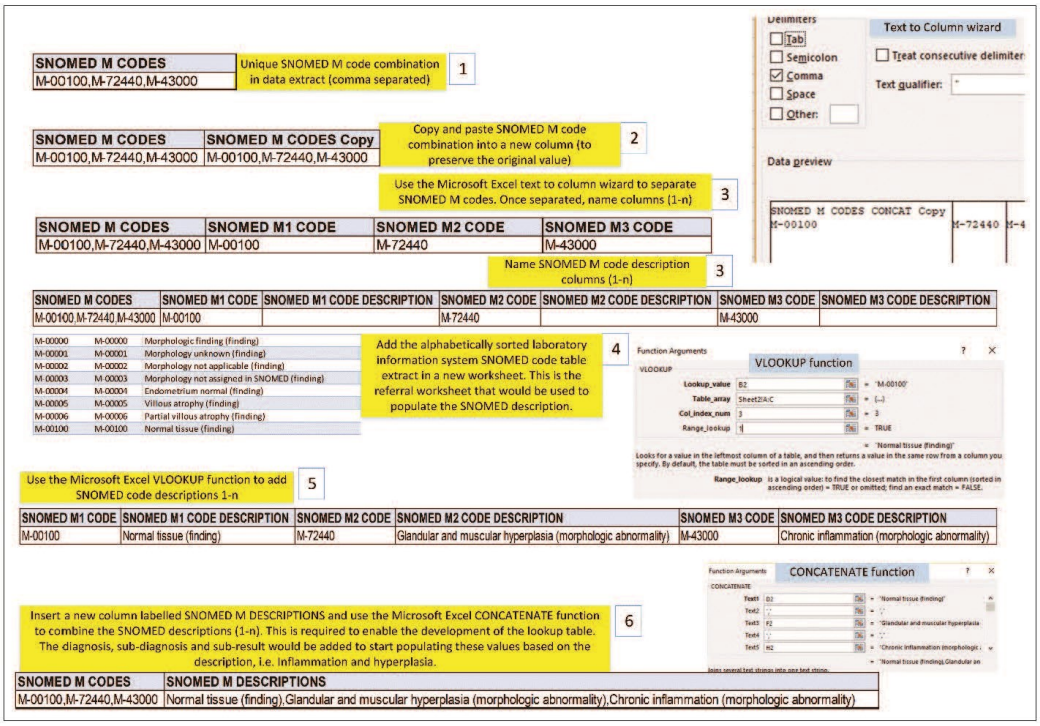
The chained SNOMED codes were provided in a format that could not readily be analysed and had to be separated into individual columns with the applicable descriptions added from the LIS code table. Data were prepared using Microsoft Excel (Microsoft Corporation, Redmond, Washington, United States)[25]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



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

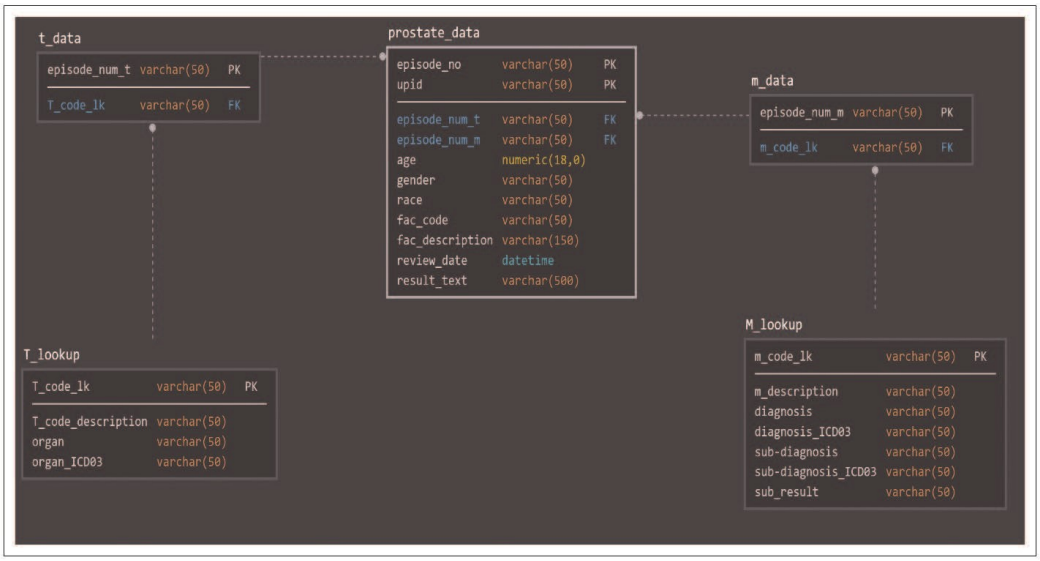


FIGURE 3: Relational database diagram describes how the various tables were joined (left outer). For each table, the primary and foreign keys are provided. The lookup tables contain only the unique Systemized Nomenclature of Medicine code combinations. The lines indicate a relationship join between tables. 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

1. 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
2. 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
3. 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
4. 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.
5. 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
6. 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
7. 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
8. 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/
9. Centers for Disease Control and Prevention (CDC) NCfCDPaHP, 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
10. 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
11. 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/
12. 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
13. 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/
14. National Cancer Registry. Summary statistics of cancer diagnosed histologically in South Africa in 2014. Johannesburg: National Institute for Occupational Health (NIOH); 2014.
15. 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
16. 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
17. 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
18. 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
19. 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
20. 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-
21. 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?preview=/28742871/47677485/doc_StarterGuide_Current-en-US_INT_20170728.pdf
22. 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
23. 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.aginity.com/documentation/WB/NZ/Default.htm
24. 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
25. 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

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. Nanoo 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

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.oa.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

APPENDIX B – PUBLICATION

Using text mining techniques to extract prostate cancer prognostic information (Gleason score) from semi-structured narrative laboratory reports in the Gauteng province, South Africa

Naseem Cassim, Michael Mapundu, Victor Olago, Turgay Celik, Jaya Anna George and Deborah Kim Glencross

Proof that manuscript has been submitted

Naseem Cassim

From: BMC Medical Informatics and Decision Making <bmcmedicalinformaticsanddecisionmaking@biomedcentral.com>
Sent: Wednesday, 04 August 2021 05:49
To: Naseem Cassim
Subject: BMC Medical Informatics and Decision Making - Receipt of Manuscript 'Using text mining...'

Ref: Submission ID 76d76599-585f-4603-a2c1-f2222a866663

Dear Dr Cassim,

Thank you for submitting your manuscript to BMC Medical Informatics and Decision Making.

Your manuscript is now at our initial Quality Check stage, where we look for adherence to the journal's submission guidelines, including any relevant editorial and publishing policies. If there are any points that need to be addressed prior to progressing we will send you a detailed email. Otherwise, your manuscript will proceed into peer review.

You can check on the status of your submission at any time by using the link below and logging in with the account you created for this submission:

https://researcher.nature.com/your-submissions?utm_source=submissions&utm_medium=email&utm_campaign=confirmation-email&journal_id=12911

To help with your article processing charge at acceptance, and to see if you're eligible for any waivers or BMC membership discounts, please click here:

https://payment.springernature.com/quote/identify/submission/76d76599-585f-4603-a2c1-f2222a866663

Kind regards,

Peer Review Advisors
BMC Medical Informatics and Decision Making

Springer Nature offers an open access support service to make it easier for our authors to discover and apply for APC funding. For further information please visit http://www.springernature.com/gp/open-research/funding

Our flexible approach during the COVID-19 pandemic

If you need more time at any stage of the peer-review process, please do let us know. While our systems will continue to remind you of the original timelines, we aim to be as flexible as possible during the current pandemic.

APPENDIX C – PUBLICATION

Trends in first-ever prostate specific antigen (PSA) testing for primary health care facilities in the Gauteng province, South Africa, between 2006 and 2016

N Cassim, Timothy R Rebbeck, Deborah K Glencross and Jaya A George

Proof that manuscript has been submitted

Naseem Cassim

From: BMJ Open <onbehalfof@manuscriptcentral.com>
Sent: Wednesday, 29 July 2020 16:46
To: Naseem Cassim; timothy_rebbeck@dfci.harvard.edu; debbie.glencross@nhls.ac.za; Jaya George
Subject: BMJ Open - Manuscript ID bmjopen-2020-043145

COVID-19: A message from BMJ: https://authors.bmj.com/policies/covid-19

29-Jul-2020

Dear Mr. Cassim:

Your manuscript entitled "Trends in first-ever prostate specific antigen (PSA) testing for primary health care facilities in the Gauteng province, South Africa, between 2006 and 2016" has been successfully submitted online and is presently being given full consideration for publication in BMJ Open.

Your manuscript ID is bmjopen-2020-043145.

Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your street address or e-mail address, please log in to ScholarOne Manuscripts at https://mc.manuscriptcentral.com/bmjopen and edit your user information as appropriate.

Please check that all author names are correctly entered as this will be the name displayed in any PubMed search.

You can also view the status of your manuscript at any time by checking your Author Center after logging in to https://mc.manuscriptcentral.com/bmjopen.

Authors should be aware that this journal is an open access publication and levies an Article Processing Charge (APC) that will be payable should the manuscript be accepted. Waivers and discounts are offered in certain circumstances. View the journal's author guidelines for further information.

Please note, a number of institutions have taken out open access memberships with BMJ, which either covers the cost of open access publishing for authors at participating institutes, or allows authors to receive a discount on the Article Processing Charge. Please visit our open access page to see a full list of participating institutions, find out if you are eligible and how to obtain your discount code - http://journals.bmj.com/site/authors/openaccess.xhtml#open-access-institutional-memberships.

We are constantly trying to find ways of improving the peer review system and continually monitor processes and methods by including article submissions and reviews in our research. If you do not wish your paper or review entered into our peer review research programme, please let us know by emailing editorial.bmjopen@bmjgroup.com as soon as possible.

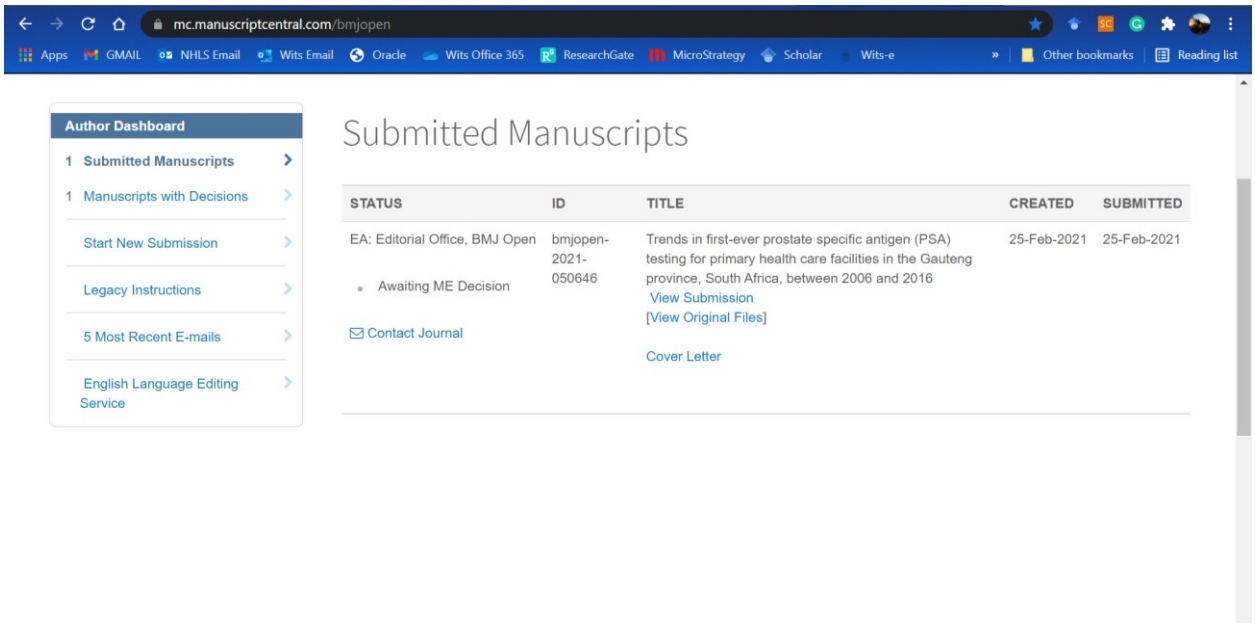
Any individuals listed as co-authors on this manuscript are copied into this submission confirmation email. If you believe that you have received this email in error, please contact the Editorial Office.

Thank you for submitting your manuscript to BMJ Open.

Sincerely,
BMJ Open Editorial Office

P.s. What did you think of the article submission process?

Proof that Review 1 has been completed



APPENDIX D – PUBLICATION

Prostate Cancer age-standardised incidence increase between 2006 and 2016 in the Gauteng Province, South Africa: a laboratory data-based analysis

Naseem Cassim, Ahsan Ahmad, Reubina Wadee, Timothy R Rebbeck, Deborah K Glencross, and Jaya A George

South African Medical Journal, [S.l.], v. 111, n. 1, p. 26-32, doi: 10.7196/SAMJ.2020.v111i1.14850.

This open-access article is distributed under Creative Commons licence CC-BY-NC 4.0.

Prostate cancer age-standardised incidence increase between 2006 and 2016 in Gauteng Province, South Africa: A laboratory data-based analysis

N Cassim,[1,2] MPH; **A Ahmad,**[3] MB ChB, MMed (Urology); **R Wadee,**[2,4] MB BCh, FC Path (SA) Anat, MMed (Anatomical Pathology), PhD; **T R Rebbeck,**[5] PhD; **D K Glencross,**[1,2] MB ChB, MMed (Haematology); **J A George,**[6] MB BCh, MSc, FC Path (Clinical Pathology), PhD

[1] *Department of Molecular Medicine and Haematology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa*
[2] *National Health Laboratory Service, Johannesburg, South Africa*
[3] *Department of Urology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa*
[4] *Department of Anatomical Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa*
[5] *Harvard T H Chan School of Public Health and Dana Farber Cancer Institute, Boston, USA*
[6] *Department of Chemical Pathology, Faculty of Health Sciences, University of the Witwatersrand, and National Health Laboratory Service, Johannesburg, South Africa*

Corresponding author: *N Cassim (naseem.cassim@wits.ac.za)*

Background. Prostate cancer (PCa) is the leading male neoplasm in South Africa (SA) and is the second most frequently diagnosed cancer among men globally. Age-specific incidence rates (ASIRs) vary by up to 189-fold globally, with an ASIR of 68.0 per 100 000 in 2018 in SA.
Objectives. To describe PCa among men undergoing prostate biopsy in Gauteng Province, SA.
Methods. We undertook a retrospective descriptive study using prostate biopsy data collected from the National Health Laboratory Service (NHLS) database between 2006 and 2016. We extracted the Systematized Nomenclature of Medicine (SNOMED) clinical terms morphology and topography codes to assign histological findings using the International Classification of Diseases for Oncology. PCa was defined as adenocarcinoma with a reported Gleason Score (GS). The new grade group (GG) based on the GS is defined as follows; (*i*) GG1 for a GS ≤6; (*ii*) GG2 for a GS of 3 + 4 = 7 ; (*iii*) GG3 for a GS of 4 + 3 = 7; (*iv*) GG4 for a GS of 8; and (*v*) GG5 for a GS ≥9. Higher-grade disease was defined as GG4 and GG5 (GS ≥8), in line with local guidelines. We reported associations of PCa with a GS ≥7 with age and race and used provincial and world standard population data to determine annual ASIRs.
Results. We identified 22 937 biopsies referred to the NHLS between 2006 and 2016. Of the 6 448 biopsies (39%) with a PCa finding for black Africans, 46% were diagnosed with high-risk PCa compared with 36 - 40% for other race groups ($p<0.0001$). Black Africans were more likely than whites to have GG4 or GG5 PCa (odds ratio 1.45; 95% confidence interval 1.27 - 1.67). The ASIR increased from 44.9 per 100 000 in 2006 to 57.3 per 100 000 in 2016.
Conclusions. Black African men were significantly more likely to present with PCa with a GS ≥8 (GG4 and GG5) compared with the other racial groups in Gauteng. The ASIR increased dramatically during the study period, perhaps as a result of increased screening and awareness. There is a need for additional research to better understand why black African men present with higher-grade disease.

S Afr Med J 2021;111(1):26-32. https://doi.org/10.7196/SAMJ.2021.v111i1.14850

Prostate cancer (PCa) is becoming a major public health concern as life expectancy increases globally.[1,2] In 2018, the global PCa incidence was the second highest after lung cancer.[3] PCa incidence varied by up to 189-fold across 186 countries, with age-standardised incidence rate (ASIR) ranges varying widely from 1.0 per 100 000 in Bhutan to as high as 189.1 in Guadeloupe in 2018.[4] Comparatively higher ASIRs were reported for some less-developed regions such as Barbados (129.3 per 100 000), Bahamas (85.8), South Africa (SA) (68.0) and Reunion (63.7).[4] The PCa ASIR is highest among men of African descent, reflecting racial and genetic predisposition in these countries.[5] The varying PCa ASIR could also be due to varying PCa screening and testing patterns that affect incidence.[6]

Limited SA data reported that black Africans present with more aggressive PCa, with a Gleason Score (GS) ≥7 reported in 36%, compared with 17% for African Americans.[7] The GS is the strongest prognostic factor for PCa treatment.[8] Local guidelines categorise PCa risk using the GS as follows: (*i*) 2 - 6 low; (*ii*) 7 intermediate; and (*iii*) 8 - 10 high.[9,10] The new grade group (GG) is based on the GS as follows: (*i*) GG1 for a GS ≤6; (*ii*) GG2 for a GS of 3 + 4 = 7; (*iii*) GG3 for a GS of 4 + 3 = 7; (*iv*) GG4 for a GS of 8; and (*v*) GG5 for a GS ≥9.[9,11] Limited PCa biopsy data exist for black African men, however, with most local studies reporting data for selected health facilities.[7,12,13] Currently, aside from local cancer registry data, which reported national 2014 data by 2018, there are no comprehensive public sector PCa data for Gauteng Province.[14]

In SA, PCa is the leading neoplasm in men.[3,4] Risk factors include age, race (black African) and family history.[1,9,10,15] The National Health Laboratory Service (NHLS) provides diagnostic services to 80% of the population.[16] Routinely accumulated laboratory data through analysis of diagnostic samples offer a unique opportunity to describe PCa histological characteristics with the related demographic detail in local populations. The Systematized Nomenclature of Medicine (SNOMED) clinical terms (CT) topography (T) and morphology (M) codes captured routinely based on the histological findings could be used to represent relevant clinical information consistently and comprehensively as an essential part of producing

electronic health information.[17] The T code identifies anatomical terms for each organ, and the M code describes the microscopic changes observed.[17]

We chose Gauteng for our study because it is home to 24.1% of the national population[18] and represents a typical dense urban area. It is the economic hub of the country with the highest numbers of both international and domestic migrants, the latter originating from provinces such as Limpopo, KwaZulu-Natal and Eastern Cape.[18,19] These domestic migrants are aged between 15 and 64 years, with males outnumbering females by 111 to 100.[20] Over 70% of migrants are black Africans.[20]

Objectives

To describe PCa among men undergoing prostate biopsies in Gauteng Province, SA, using data fields directly extracted from patient laboratory records.

Methods

Study design

We extracted all prostate biopsies done on men aged >30 years between 1 January 2006 and 31 December 2016 and referred to the NHLS for pathology evaluation in Gauteng. Two data sets were extracted from the national laboratory data repository that houses laboratory information system (LIS) collated patient laboratory reports.[16] These data sets were: (*i*) narrative prostate biopsy data; and (*ii*) SNOMED CT code(s). Ethical clearance was obtained from the University of the Witwatersrand, Johannesburg (ref. no. M170419).

Biopsy data preparation

The narrative biopsy extract included the following variables: (*i*) episode number; (*ii*) unique patient identifier; (*iii*) age; (*iv*) gender; (*v*) racial group; (*vi*) facility name; (*vii*) reviewed date; and (*viii*) biopsy narrative report. The episode number identifies samples for all laboratory processes. The unique patient identifier was generated using the organisation probabilistic matching algorithm that includes fuzzy logic matching, minimising the duplication of PCa case reports.[21] For duplicate records, only the first record was used to calculate the ASIR. A sequence number was generated for chronologically sorted patient data. The GS was manually coded, with a random sample independently verified to validate manual coding (n=369).

The SNOMED CT data extract included the episode number and the M and T codes. The M code describes the microscopic changes noted in cells, tissues, and organs.[17] The T code describes the anatomical site of origin (prostate).[17] We used the SNOMED CT codes to develop lookup tables to report: (*i*) histological findings (PCa/not PCa); (*ii*) biopsy findings; (*iii*) tumour type; and (*iv*) organ, with the assistance of an experienced anatomical pathologist and urologist (Appendix A, http://samj.org.za/public/sup/14850.png). We used the International Classification of Diseases for Oncology to assign PCa findings, i.e. adenocarcinoma tumour type with the associated tumour grading (GS).[22] For missing M codes, we manually coded biopsies using the narrative result text.

Racial group was identified as an important PCa risk factor.[7,12,13] As racial group was poorly recorded (94.5% missing), we used a locally developed hot-deck imputation method to assign missing data,[23] described in an unpublished study (Chen W, Kellett P, Greyling M, Sengayi M, 'The use of surnames to impute missing ethnicity data in the National Cancer Registry database of South Africa', 2019). Well-populated cancer registry data with patient-reported race grouping were used to construct the imputation reference panel. The racial groups reported by Statistics South Africa in the 2011 census[24] are Indian/Asian, black African, coloured (mixed ancestry from European, Indian subcontinent and local indigenous African peoples[25]) and white.

We used a relational database to combine the respective tables using a left outer join with the narrative biopsy data extract as the master table (Appendix A, database diagram showing how data from the SNOMED CT lookup tables and the race imputation data sets were used to create the prostate cancer data for this study, http://samj.org.za/public/sup/14850.png). Data were prepared using Microsoft Office Professional Plus 2013 (Microsoft, USA) and analysed using SAS 9.4 (SAS Institute, USA) and Stata 12 (StataCorp, USA).

Descriptive statistics by racial group were: (*i*) mean age (with range); (*ii*) biopsy numbers; (*iii*) biopsy findings; and (*iv*) number of biopsies reporting a PCa histological finding. We compared the mean age across racial groups by one-way analysis of variance (ANOVA). We used the χ^2 test to identify whether there was a statistical difference for age category and biopsy findings for the four race groups. An alpha level of 0.05 was used for all association tests.

We used a pyramid chart to report the number of biopsies with a PCa and non-PCa finding for age categories (excluding unknown age). Age was reported on the x-axis and percentage of biopsies on the y-axis. Annual biopsy numbers were reported as a bar chart.

A multiple logistic regression method was used to determine the factors associated with PCa with a GG ≥4, controlling for age and racial group. This is in line with local guidelines for a GS ≥8 classified as high risk.[9,10] We reported odds ratios (ORs), p-values and confidence intervals (CIs).

Biopsies with a PCa finding were analysed by GS and GG risk group for the four racial groups, reported as a stacked bar chart.[9,10] We used a stacked bar chart to report the percentage of biopsies per year for the GS and GG risk categories, with GG4 and GG5 reported together.

The ASIR per 100 000 and average annual percentage change (AAPC) were calculated for defined age categories for each calendar year using the Joinpoint trend analysis software. Data variables for the model were: (*i*) gender; (*ii*) year of diagnosis; (*iii*) age categories/groups; (*iv*) number of incident cases (count variable); (*v*) population estimates; and (*vi*) Segi world standard population.[26-28] The last was used to enable comparison with cancer registry reports.

Results

There were 25 010 biopsies extracted from the LIS. The SNOMED CT codes were used to assign 90% of histological findings, with 10% manually coded. We excluded 285 biopsies that did not meet the inclusion criteria, and excluded a further 2 073 non-prostatic biopsies (8.3%). The total sample size was therefore 22 937. It was not possible to determine a GS for 455/8 390 biopsies (5.4%) with a PCa finding. We randomly reviewed GSs for accuracy and noted no misclassification errors.

Prostate biopsy outcomes by age and racial group

The mean age by racial group was 66.3 (range 31 - 92) years for Indians/Asians, 67.4 (30 - 99) years for black Africans, and 67.7 (30 - 99) years for whites (Table 1). A statistically significant difference for mean age for the four racial groups was determined by one-way ANOVA (p<0.001). There was a significant difference in biopsy frequency by age category across the four racial groups (p<0.0001). A malignant biopsy finding was reported for 46% of black Africans compared with 39%, 37% and 30% of coloureds, whites and Indians/Asians, respectively. For black Africans, 39% of biopsies reported a PCa finding, compared with 23%, 29% and 31% for Indians/Asians, whites and coloureds, respectively (p<0.0001).

Twenty-three percent of biopsies from the 65 - 69-year age category reported a PCa finding, as opposed to 19% and 20% for the 60 - 64-

Table 1. Prostate biopsy descriptive statistics by racial group, Gauteng Province, 2006 - 2016

	Indian/Asian	Black African	Coloured	White
Biopsy samples, *n*	487	16 491	1 129	3 422
Age (years), mean (range)	66.3 (31 - 92)	67.4 (30 - 99)	67.4 (35 - 99)	67.7 (30 - 99)
p-value*				<0.001
Age category (years), *n* (%)				
30 - 39	4 (1)	49 (0)	2 (0)	7 (0)
40 - 49	13 (3)	298 (2)	22 (2)	44 (1)
50 - 59	76 (16)	2 418 (15)	160 (14)	436 (13)
≥60	380 (78)	12 502 (76)	890 (79)	2 830 (83)
Unknown	14 (3)	1 224 (7)	55 (5)	105 (3)
p-value†				<0.0001
Biopsy findings, *n* (%)				
Atypia/dysplasia	21 (4)	575 (3)	43 (4)	129 (4)
Benign/negative for malignancy	317 (65)	8 321 (50)	641 (57)	2 031 (59)
Neoplasm, malignant	147 (30)	7 561 (46)	442 (39)	1 255 (37)
Uncertain whether benign or malignant	2 (0.5)	34 (0)	3 (0)	7 (0)
p-value†				<0.0001
Prostate cancer, *n* (%)‡	111 (23)	6 448 (39)	354 (31)	1 009 (29)

*One-way analysis of variance.
†χ^2 test.
‡Adenocarcinoma histological finding with Gleason score reported.

and 70 - 74-year age categories, respectively (Fig. 1). The percentage of non-PCa biopsies was lowest in older ages, with 12% reported in the 75 - 79-year and 8% in the ≥80-year age categories. Eleven percent and 5% of biopsies reported a PCa finding for the 50 - 54-year and 55 - 59-year age categories, respectively. A non-PCa finding was reported for 22% of biopsies in the 60 - 64-year age category, decreasing to 19% and 11% for the 65 - 69-year and 70 - 74-year age categories, respectively.

Black Africans were more likely to have PCa with a GG ≥4 than whites (OR 1.45; 95% CI 1.27 - 1.67) (Table 2). Similarly, coloureds and Indians/Asians were also more likely to have PCa with a GG ≥4 than whites, with ORs of 1.14 (95% CI 0.89 - 1.46) and 1.06 (95% CI 0.71 - 1.59), respectively.

Men aged <70 years were determined to have a reduced probability of being diagnosed with PCa with a GG ≥4 compared with those aged ≥70 years, with an OR of 0.99 (95% CI 0.69 - 1.42) for ages 30 - 49 years, 0.90 (95% CI 0.79 - 1.03) for 50 - 59 years, and 0.87 (CI 0.79 - 0.96) for 60 - 69 years. Controlling for racial group, males aged ≥70 years were at a significantly greater risk of having PCa with a GG ≥4.

Between 2006 and 2016, the number of biopsies increased from 1 630 to 2 761 (Fig. 2). We noted lower volumes in 2008 (*n*=1 168), decreasing from 1 610 in 2007 (27% decrease). A 25% increase in biopsies was reported between 2010 (*n*=1 868) and 2011 (*n*=2 335), compared with 19% between 2014 (*n*=2 357) and 2015 (*n*=2 805).

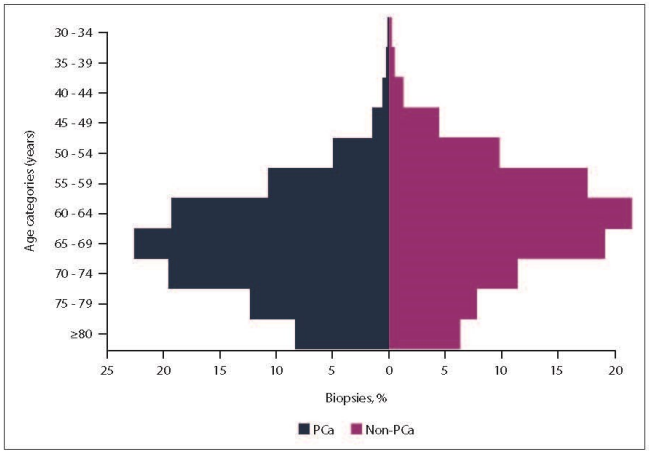


Fig. 1. Age distribution for PCa findings. Biopsy numbers reported by age category for both PCa and non-PCa histological findings in Gauteng Province between 2006 and 2016, as a population pyramid. (PCa = prostate cancer.)

Analysis of low-, intermediate- and high-risk GSs

A high-risk GS (≥8) was reported for 46% of biopsies (*n*=6 448) with a PCa finding for black Africans compared with 40%, 38% and 36% for coloureds, Indians/Asians and whites, respectively (Fig. 3). The percentage of biopsies with a GG2 ranged from 17% for whites to 24% for coloureds. Similarly, the percentage of biopsies with a GG3 ranged from 10% for coloureds to 15% for Indians/Asians. Of biopsies for black Africans, GG2 and GG3 were reported in 18% and 13%, respectively.

Between 2006 and 2016, the percentage of biopsies with a GG1 decreased from 27% to 18% (Fig. 4). Conversely, the percentage of biopsies with a GG ≥4 increased from 42% in 2006 to 48% by 2016. Biopsies with a GG3 increased from 12% in 2006 to 27% by 2016. Conversely, the percentage of biopsies with a GG2 decreased from 20% to 18% over the same time period.

ASIRs for prostate cancer

The ASIR increased from 44.92 per 100 000 in 2006 to 57.31 by 2016 (Fig. 5). Between 2006 and 2016, we reported an AAPC of 2.7%

(95% CI −1.2 - 6.7), which was significant ($p<0.05$). A lower ASIR was reported for 2008 (27.0 per 100 000). Between 2014 and 2015, the ASIR increased from 40.8 to 52.6 per 100 000, respectively (largest year-on-year increase).

Between 2006 and 2016, an annual percentage change (APC) and average annual percentage change (AAPC) of 2.7 (95% CI −1.2 - 6.7) was reported (Table 3). The p-value was <0.0001 for both the APC and AAPC, indicating a statistical difference.

Discussion

Our study reported an increase in biopsy numbers between 2006 and 2016, with PCa diagnosed predominantly in older men. The analysis also reveals that the ASIR in Gauteng Province has increased substantially from 44.9 (2006) to 57.3 (2016) per 100 000. We also found that black Africans were statistically significantly more likely to have advanced disease with a GG ≥4 compared with other racial groups.

The large increase in prostate biopsies reported in our study may indirectly reflect improved access to care. This would include more patient referrals to a urologist following an elevated prostate-specific antigen (PSA) level and/or abnormal findings on digital rectal examination. Urology services that were previously limited to academic centres have been introduced in several provincial hospitals,[26] so patients can be managed locally without the need to travel to tertiary hospitals.

Our findings of an increasing ASIR are in keeping with those of Babb *et al.*,[27] who reported a rise in SA from 17 per 100 000 in 1986 to 27 per 100 000 by 2006. Similarly, between 2007 and 2013, the National Cancer Registry (NCR) reported that the ASIR increased from 29 to 44 per 100 000.[28,29] We have shown a slightly higher increase compared with Babb *et al.*[27] and the NCR-reported national values.[30] Between 2007 and 2013, the PCa ASIR in Gauteng was substantially higher than national estimates at 41.1 v. 29 per 100 000 and 48.4 v. 44 per 100 000, respectively.[28,29] By reporting provincial trends, we are able to highlight areas with an ASIR increasing faster than nationally reported data. Some of the reasons for the reported increase in incidence include improved urological guidelines, better access to screening, enhanced patient awareness, and improved medical aid coverage for PCa diagnosis.[9,10]

The 2013 PCa urology treatment guidelines recommend screening from the age of 40 for black Africans and 45 for all males.[10] Studies have shown that PSA screening is associated with increased diagnosis and decreased mortality.[8] Our own unpublished data (Cassim N, Ahmad A, Glencross DK, George JA, 'Trends in prostate specific antigen (PSA) testing for primary health care facilities in the Gauteng province, South Africa, between 2006 and 2016', 2020) indicate that primary healthcare (PHC) PSA testing increased substantially in Gauteng between 2014 and 2016. It is reasonable to assume that improved access to screening in Gauteng, specifically at PHC facilities, may have resulted in the increased PCa diagnosis. Screening and education both play important roles in identifying men at risk of developing PCa.[8-10,31]

It is also possible that PCa awareness is a cause for the increased ASIR noted.

Table 2. Multiple logistic regression analysis to assess the association between prostate cancer outcomes, GS, racial group and age category*

Characteristic	OR	95% CI	p-value
Racial group			
White	1	-	-
Indian/Asian	1.06	0.71 - 1.59	0.764
Black African	1.45	1.27 - 1.67	<0.0001
Coloured	1.14	0.89 - 1.46	0.311
Age category (years)			
30 - 49	0.99	0.69 - 1.42	0.948
50 - 59	0.90	0.79 - 1.03	0.131
60 - 69	0.87	0.79 - 0.96	0.005
≥70	1	-	-

GS = Gleason Score; OR = odds ratio; CI = confidence interval; GG = grade group.
*For the dependent variable, a GG of 4 or 5 was coded as 1 (GS ≥8) and GG1 - 3 as 0 for a prostate cancer histological finding with a GS reported. For the logistic regression analysis, the 95% CI, p-value and age- (for analysis of racial group) or race-adjusted (for analysis of age group) ORs are reported.

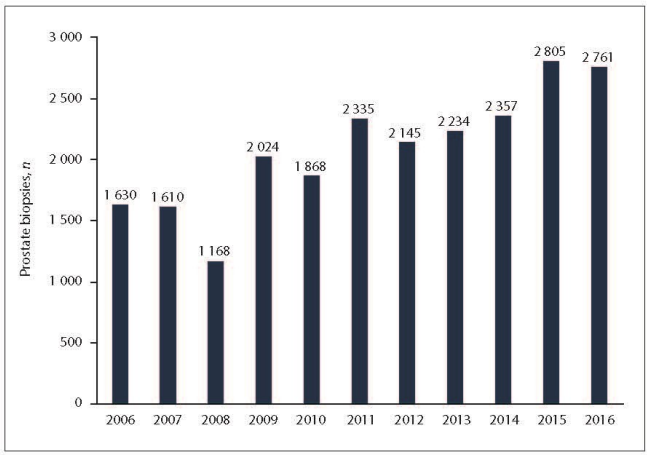


Fig. 2. Prostate biopsy numbers for Gauteng Province between 2006 and 2016.

Table 3. APC and AAPC output from the Joinpoint model for age-standardised prostate cancer incidence rates between 2006 and 2016

	Segment	Lower endpoint	Upper endpoint	APC	Lower CI	Upper CI	p-value
APC	1	2006	2016	2.7	−1.2	6.7	<0.0001
AAPC	1	2006	2016	2.7	−1.2	6.7	<0.0001

APC = annual percentage change; AAPC = average annual percentage change; CI = confidence interval.

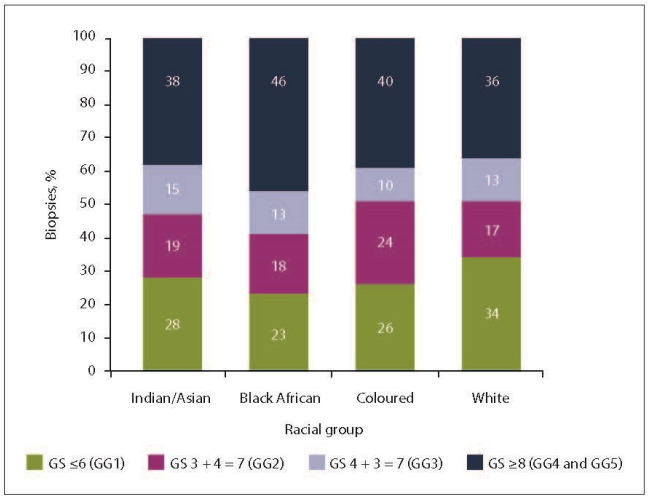


Fig. 3. Percentage of prostate biopsies with a low- (GS ≤6), intermediate- (GS 7: 3 + 4 = 7 and 4 + 3 = 7) and high- (GS ≥8) risk GS by racial group. The GG is indicated in brackets, with 4 and 5 combined to report a GS ≥8. (GS = Gleason Score; GG = grade group.)

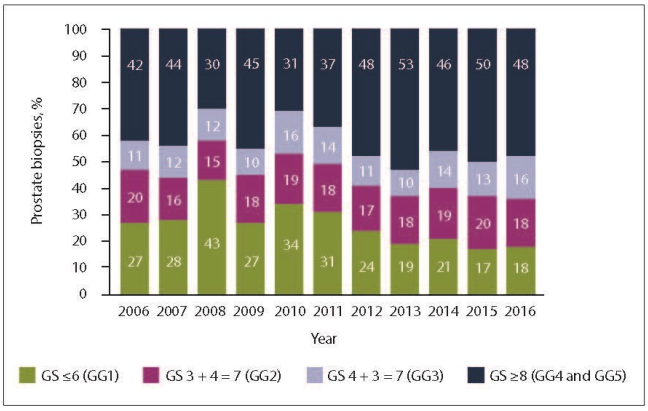


Fig. 4. Percentage of prostate biopsies with a low- (GS ≤6), intermediate- (GS 7: 3 + 4 = 7 and 4 + 3 = 7) and high- (GS ≥8) risk GS by calendar year. The GG is indicated in brackets, with 4 and 5 combined to report a GS ≥8. (GS = Gleason Score; GG = grade group.)

Public awareness of female cancers started to improve in 2009, promoting early detection of breast cancer with mobile mammography units.[32] In contrast, local awareness campaigns for PCa started much later. with initiatives such as 'Movember', 'Daredevil Run' and 'Suit Up September' beginning in 2015 and operating predominantly in metropolitan areas.[33-35]

PCa has the highest reported heritability of any major cancer, and several genetic susceptibility loci have been identified in European and Asian men.[5,36] Another reason for the increased ASIR may be a combination of genetic and environmental risk factors that are not known at present.

We noted an increase in the percentage of biopsies with a high-risk GS between 2006 and 2016. From 2012, a high-risk GS was reported for ~50% of biopsies, highlighting late presentation in Gauteng. A system review by Ilic *et al.*[37] reported that PSA testing increased the detection of early-stage disease and slightly decreased the detection of advanced PCa. Similarly, other studies have shown that access to PSA testing has increased PCa detection rates, combined with a migration to lower-stage disease, throughout the USA, Canada and Europe.[38] Unpublished data from our province (Chen *et al.*) showed that PHC PSA testing increased dramatically in 2013. We would therefore have expected to see an increase in early-stage disease (GS ≤6). It is possible that it is too early to realise the benefit of increased PSA testing.

We noted that black African men were statistically significantly more likely than whites and coloureds to present with a GG ≥4. While the reasons for this finding are uncertain, racial differences in healthcare-seeking behaviour have been reported, with white and coloured respondents more likely to seek care than black Africans.[39] In SA, black males have been shown to be less likely to access healthcare than females, and sought care first from traditional healers.[39] Male underutilisation of healthcare services has been identified in multiple countries.[40] One of the challenges men face is the working hours of health services.[40] Although increasing numbers of men presented for diagnosis in our study, it may be that lack of knowledge is one of multiple factors for late presentation.[33-35]

Our finding that older age was a risk factor for high-grade PCa is in keeping with some studies from Canada and the USA,[41,42] while local studies have produced conflicting results.[7,12,13] Although local guidelines recommend PCa screening from 40 years, we found that most patients were diagnosed in their mid-60s.[9,10] This is similar to other local findings and published data from Ghana and Nigeria.[7,12,13,43,44] Our local unpublished data (Cassim *et al.*) indicate that the mean age for PSA testing at PHC facilities was 55.8 years, highlighting possible delayed diagnosis.[30] Maphayi *et al.*[45] showed that only 7% of black African men who had a PSA ≥4 µg/L had a follow-up biopsy.

Our study builds on previous work providing evidence that laboratory data are an invaluable repository of health data. Specifically, this work has extended PCa reporting using laboratory data. Applying similar methods of reporting, such as the use of SNOMED CT lookup tables, could provide national and provincial PCa statistics across SA to complement cancer registry reporting. These approaches can be complemented and supplemented with the application of text mining and predictive analytics that have demonstrated the reliable extraction of prognostic GS data from narrative biopsy reports (Cassim N, Mapundu M, Olago V,

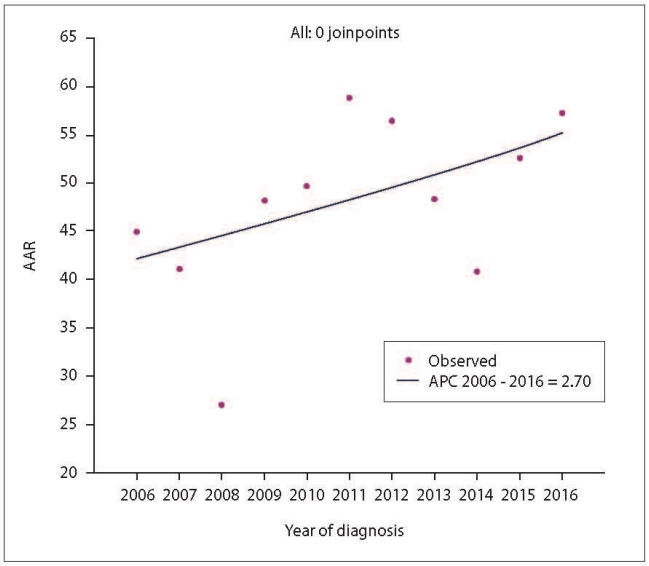


Fig. 5. Age-standardised prostate cancer incidence rates between 2006 and 2016 reported as mauve dots, with the AAPC reported as a blue line. (AAPC = average annual percentage change; APC = annual percentage change; AAR = age-adjusted rate.)

Ahmad A, George JA, Glencross DK, 'Using big data techniques to extract Gleason scores to improve prostate cancer reporting in the Gauteng province, South Africa', 2019, unpublished). These tools have the potential to generate automated PCa reporting, leading to the development of PCa epidemiological reporting using laboratory data.

Study limitations

This study did not investigate underlying contributing factors in under-performing districts as this was beyond the scope of the study. The incidence data reported for our study were reliant on the probabilistic matching algorithm for the de-duplication of patient records in the absence of a national unique patient identifier.

A limitation of our study is the absence of private sector laboratory data, affecting generalisability. There are planned discussions to receive private sector PCa data routinely. However, a local study reported that withheld private sector cancer data only accounted for 4% of cancer reporting.[46]

The accuracy of the hot-deck race imputation was also a limitation. In their unpublished study, Chen *et al.* used a hold-out test to compare imputed versus cancer registry patient-reported racial group for 406 642 unique surname-race group pairings. The historically well-populated cancer registry data were split into patient-reported racial group values and imputed values. Chen *et al.* compared the imputed and original values. For the imputed data set, a sensitivity of ≥90% was reported for black Africans, whites and Indians/Asians. However, a lower sensitivity of 63.79% was reported for coloureds. This finding indicates that the race imputation may be less accurate for coloureds, which may result in some misclassification. A challenge in SA is that common surnames are used interchangeably by coloureds and whites.[47]

It was not possible to control for confounders such as socioeconomic status and environmental exposure, as these data were not available.

Conclusions

There is a need to improve our understanding of the contribution of genetic, environmental and lifestyle factors to PCa in black African men. There is also a need to improve our understanding of late presentation in black African men. Late presentation is associated with limited treatment options and an increased mortality rate. Broader public health initiatives are required to identify men earlier by improving awareness and health-seeking behaviour as well as providing better access to urology services that meet their needs. Here, the provision of tailor-made male-only healthcare services could play an important role in reducing PCa incidence.

Declaration. The research for this study was done in partial fulfilment of the requirements for NC's PhD degree, done in the School of Pathology at the University of the Witwatersrand.

Acknowledgements. The authors thank Prof. Martin Hale, Prof. Mohammed Haffejee, Dr Innocent Maposa, Dr Elvira Singh, Dr Mazvita Sengayi and Ms Patricia Kellett for their assistance.

Author contributions. NC: executed the research, data analysis, prepared the first draft and final draft submitted. AA and RW: data analysis assistance. TRR: data analysis and editorial assistance. DKG and JAG: co-developed and executed research, editorial input and project leader/budget owner.

Funding. This work was supported in part by grants from the US National Cancer Institute to TRR (U01-CA184374) and from the NHLS Research Trust to NC (development grant).

Conflicts of interest. None.

1. Cooperberg MR, Chan JM. Epidemiology of prostate cancer. World J Urol 2017;35(6):849. https://doi.org/10.1007/s00345-017-2038-0
2. Neupane S, Bray F, Auvinen A. National economic and development indicators and international variation in prostate cancer incidence and mortality: An ecological analysis. World J Urol 2017;35(6):851-858. https://doi.org/10.1007/s00345-016-1953-9
3. Ferlay J, Soerjomataram I, Ervik M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136(5):E359-E386. https://doi.org/10.1002/ijc.29210
4. 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
5. Rebbeck TR, Devesa SS, Chang BL, et al. Global patterns of prostate cancer incidence, aggressiveness, and mortality in men of African descent. Prostate Cancer 2013;2013:560857. https://doi.org/10.1155/2013/560857
6. Center MM, Jemal A, Lortet-Tieulent J, et al. International variation in prostate cancer incidence and mortality rates. Eur Urol 2012;61(6):1079-1092. https://doi.org/10.1016/j.eururo.2012.02.054
7. Tindall EA, Monare LR, Petersen DC, et al. Clinical presentation of prostate cancer in black South Africans. Prostate 2014;74(8):880-891. https://doi.org/10.1002/pros.22806
8. European Association of Urology. Guidelines on Prostate Cancer. Aarnheim, Netherlands: EAU, 2016 https://uroweb.org/wp-content/uploads/EAU-Guidelines-Prostate-Cancer-2016.pdf (accessed 19 February 2018).
9. Anderson D, Barnes R, Bida M, et al. South African Prostate Cancer Guidelines – draft version 2017. South African Urological Association, South African Society of Medical Oncology, South African Society for Clinical and Radiation Oncologists, South African Society of Nuclear Medicine, South African Oncology Consortium and Prostate Cancer Foundation of South Africa, 2017. https://prostate-ca.co.za/wp-content/uploads/2017ProstateGuidelinesDraftVersion2016.pdf (accessed 12 March 2017).
10. Segone AM, Haffejee M, Wentzel S, et al. Prostate Cancer Diagnostic and Treatment Guidelines: The Prostate Cancer Foundation of South Africa. June 2013. http://prostate.acitravel.co.za/cake/app/webroot/uploads/files/Prostate_Cancer_Guidelines_2013.pdf (accessed 12 January 2017).
11. Epstein JI, Egevad L, Amin MB, et al. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of grading patterns and proposal for a new grading system. Am J Surg Pathol 2016;40(2):244-252. https://doi.org/10.1097/PAS.0000000000000530
12. Heyns CF, Fisher M, Lecuona A, van der Merwe A. Prostate cancer among different racial groups in the Western Cape: Presenting features and management. S Afr Med J 2011;101(9):267-270.
13. 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.

14. National Cancer Registry. Summary statistics of cancer diagnosed histologically in South Africa in 2014. Johannesburg: National Institute for Occupational Health, 2014. https://www.nicd.ac.za/wp-content/uploads/2019/12/2014-NCR-tables.pdf (accessed 12 May 2018).
15. National Cancer Institute at the National Institutes of Health. Understanding prostate changes: A health guide for men. 2018. https://www.cancer.gov/types/prostate/understanding-prostate-changes (accessed 19 February 2018).
16. National Health Laboratory Service. 2017/18 Annual Report. 2018. http://www.nhls.ac.za/assets/files/an_report/NHLS_AR_2018.pdf (accessed 18 July 2019).
17. US National Library of Medicine. Overview of SNOMED CT. 2016. https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html (accessed 20 February 2018).
18. Statistics South Africa. Mid-year population estimates, 2016. Statistical release P0302. Pretoria: Stats SA, 2016. https://www.statssa.gov.za/publications/P0302/P03022016.pdf (accessed 17 June 2018).
19. Statistics South Africa. Mid-year population estimates, 2020. Statistical release P0302. Pretoria: Stats SA, 2020. http://www.statssa.gov.za/publications/P0302/P03022020.pdf (accessed 23 July 2018).
20. Oosthuizen M, Naidoo P. Internal migration to the Gauteng Province. Development Policy Research Unit, University of Cape Town, 1 December 2004. https://www.africaportal.org/publications/internal-migration-to-the-gauteng-province/ (accessed 7 February 2019).
21. Ngobeni T. Corporate Data Warehouse patient linking technical document. Johannesburg: National Health Laboratory Service, 2017.
22. International Agency for Research on Cancer, World Health Organisation. International Classification of Diseases for Oncology. 3rd ed. Geneva: WHO, 2013. https://apps.who.int/iris/bitstream/handle/10665/96612/9789241548496_eng.pdf (accessed 25 April 2018).
23. Singh E, Greyling M. A review of hot deck imputation for survey non responses: The example of population group in the national cancer registry database. Johannesburg: National Institute for Communicable Diseases, 2015.
24. Statistics South Africa. Census 2011: Key results. 2011. http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Key_results.pdf (accessed 23 January 2019).
25. FamilyTreeDNA. South African Cape Coloured. https://www.familytreedna.com/groups/cape-coloured-dna-project/surnames (accessed 31 July 2018).
26. Gauteng Department of Health. Annual Report: 2013/14. Johannesburg: GDOH, 2017. https://cmbinary.gauteng.gov.za/Media?path=cphealth/Documents/Annual%20Reports/Department%20of%20Health%20Annual%20Report%202014-2015.pdf&Item=3318&Type=Documents&Location=/cphealth (accessed 11 March 2019).
27. Babb C, Urban M, Kielkowski D, Kellett P. Prostate cancer in South Africa: Pathology based national cancer registry data (1986 - 2006) and mortality rates (1997 - 2009). Prostate Cancer 2014;2014:419801. https://doi.org/10.1155/2014/419801
28. National Cancer Registry. Summary Statistics of Cancer Diagnosed Histologically in South Africa in 2007. Johannesburg: National Institute for Occupational Health, 2007. https://www.nicd.ac.za/wp-content/uploads/2019/12/2007-NCR-2007-results.pdf (accessed 12 May 2018).
29. National Cancer Registry. Summary statistics of cancer diagnosed histologically in South Africa in 2013. Johannesburg: National Institute for Occupational Health, 2013. https://www.nicd.ac.za/wp-content/uploads/2019/12/2013-NCR.pdf (accessed 12 May 2018).
30. National Cancer Registry. Summary statistics of cancer diagnosed histologically in South Africa in 2012. Johannesburg: National Institute for Occupational Health, 2012. https://www.nicd.ac.za/wp-content/uploads/2019/12/2012-NCR-2012-results.pdf (accessed 26 February 2018).
31. Mndzebele SL, Mogale JW. Prostate cancer awareness and screening behaviour among men in Gauteng Province, South Africa: Utilization of primary health care services. Afr J Phys Health Educ Recreation Dance 2015;21(Suppl 2):296-306.
32. Pink Drive. Pink Drive: TLC Ladies MTB and Trail Running Weekend. 2016. Available https://www.pinkdrive.co.za/event/tlc-ladies-mtb-and-trail-running-weekend/http://www.pinkdrive.co.za/ (accessed 30 August 2016).
33. The Men's Foundation. Movember. 2018. https://za.movember.com/story/11810/highlights-of-movember-2018 (accessed 28 August 2018).
34. Hollard. Hollard Daredevil Run. 2018. https://www.daredevilrun.com/ (accessed 28 August 2018).
35. Prostate Cancer Foundation. Suit Up September – Prostate Cancer Awareness Campaign. 2017. http://www.prostate-ca.co.za/cake/index.php/article/76 (accessed 28 August 2018).
36. Rebbeck TR. Prostate cancer genetics: Variation by race, ethnicity, and geography. Semin Radiat Oncol 2017;27(1):3-10. https://doi.org/10.1016/j.semradonc.2016.08.002
37. Ilic D, Djulbegovic M, Jung JH, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: A systematic review and meta-analysis. BMJ 2018;362:k3519. https://doi.org/10.1136/bmj.k3519
38. Thompson IM, Ankerst DP. Prostate-specific antigen in the early detection of prostate cancer. Can Med Assoc J 2007;176(13):1853-1858. https://doi.org/10.1503/cmaj.060955
39. Abera Abaerei A, Ncayiyana J, Levin J. Health-care utilization and associated factors in Gauteng province, South Africa. Glob Health Action 2017;10(1):1305765. https://doi.org/10.1080/16549716.2017.1305765
40. Baker P, Shand T. Men's health: Time for a new approach to policy and practice? J Glob Health 2017;7(1). https://doi.org/10.7189/jogh.07.010306
41. Bechis SK, Carroll PR, Cooperberg MR. Impact of age at diagnosis on prostate cancer treatment and survival. J Clin Oncol 2011;29(2):235-241. https://doi.org/10.1200/Jco.2010.30.2075
42. Alibhai SMH, Krahn MD, Fleshner NE, Cohen MM, Tomlinson GA, Naglie G. The association between patient age and prostate cancer stage and grade at diagnosis. EJU Int 2004;94(3):303-306. https://doi.org/10.1111/j.1464-410X.2004.04883.x
43. Muhammad A, Abdulwahab-Ahmed A, Agwu N, Yunusa G, Abdullahi K, Mungadi I. Use of transrectal ultrasound in the management of urologic diseases in a tertiary hospital in north-western Nigeria. 2018;25(3):172-176. https://doi.org/10.4103/npmj.npmj_131_18
44. Yeboah ED, Hsing AW, Mante S, et al. Management of prostate cancer in Accra, Ghana. J West Afr Coll Surg 2016;6(4):31-45.
45. Maphayi M, Cassim N, Bello B, George JA. Mining laboratory data to describe prostate specific antigen testing and prostate cancer in Johannesburg, South Africa. Pan Afr Med J 2020;35:61. https://doi.org/10.11604/pamj.2020.35.61.21331
46. Singh B, Underwood JM, Nattey C, Babb C, Sengayi M, Kellett P. South African National Cancer Registry: Effect of withheld data from private health systems on cancer incidence estimates. S Afr Med J 2015;105(2):107-109. https://doi.org/10.7196/SAMJ.8858
47. Snell R. Cape slave naming patterns. South African History Online, 2 August 2016. https://www.sahistory.org.za/archive/cape-slave-naming-patterns-robert-shell (accessed 12 March 2018).

Accepted 19 June 2020.

APPENDIX E – ETHICS AND OTHER APPROVAL LETTERS

NATIONAL HEALTH LABORATORY SERVICE

Academic Affairs, Research & Quality Assurance (AARQA)
1 Modderfontein Road, Sandringham, 2031
Tel: +27 (0)11 885 5388 Fax: +27 (0)11 386 66296
Reference:naseemab

29 March 2017

Naseem Cassim
Faculty of Health Sciences
Department of Molecular Medicine and Haematology
Room 3B10
7 York Road
Parktown

Naseem Cassim_Approval to conduct PhD study

This letter serves to confirm that you have been granted permission to conduct the following study: *'Analysis of retrospective laboratory data, in the Gauteng province, between 2006 and 2016, to identify trends in prostate cancer screening and diagnosis by level of care'*, for your PhD degree.

The protocol has been reviewed and can confirm that you may access the applicable prostate cancer data required from the NHLS Corporate Data Warehouse (CDW).

Please submit the CDW application / request forms attached to the Academic Affairs & Research (AAR) Department addressed to Dr Babatyi Malope-Kgokong for final approval and the release of the required data.

Wishing you all the best with your studies.

Kind regards

Dr Sergio Carmona
Acting Executive Manager: AARQA

Dr Babatyi Malope-Kgokong
National Manager: AAR

R14/49 Mr Naseem Cassim et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M170419

NAME: (Principal Investigator)	Mr Naseem Cassim et al
DEPARTMENT:	Molecular Medicine and Haematology National Health Laboratory Service
PROJECT TITLE:	A Retrospective Analysis of Laboratory Data to Identify Clinical Practice Trends in Prostate Cancer Screening and Diagnosis by Level of Care in the Gauteng Province between 2006 and 2016
DATE CONSIDERED:	05/05/2017
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Prof J. George and Prof D.K. Glencross
APPROVED BY:	Professor P Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	10/05/2017

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

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 301, Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized 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 the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in April and will therefore be due in the month of April each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 12/6/2017

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
OFFICE OF THE DEPUTY VICE CHANCELLOR RESEARCH
12 JUN 2017
RESEARCH OFFICE HREC (Medical)

APPENDIX F - DECLARATION AND CO-AUTHOR SIGNATURE

APPENDIX D - DECLARATION AND CO-AUTHOR SIGNATURE

I, Naseem Cassim, student number 8130536, declare that this Thesis/Dissertation/Research Report is my own work and that I contributed adequately towards research findings published in the article(s) stated below which are included in my Thesis/Dissertation/Research Report .

Signature of Student **Date** 05/05/2020

Name of Supervisors

Prof Jaya A George **Signature** **Date** 05/05/2020

Prof Deborah K Glencross **Signature** **Date** 05/05/2020

Agreement by co-authors: By signing this declaration, the co-authors listed below agree to the use of the article(s) by the student as part of his/her Thesis/Dissertation/Research Report. In cases where the student is not the 1st author of a published article, the primary supervisor must explain (under comments) why the student is entitled to use the paper for his/her degree purposes.

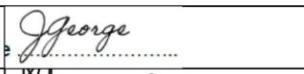
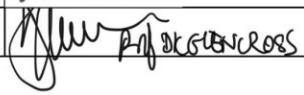
Article 1:

Title: Using Systematized Nomenclature of Medicine (SNOMED) clinical terms (CT) codes to assign histological findings for prostate biopsies in the Gauteng province, South Africa: Lessons Learnt

Afr J Lab. 2020;9(1):1-9

Authors: Naseem Cassim, Ahsan Ahmad, Reubina Wadee, Jaya A George and Deborah K Glencross

Authors	Name	Signature	Date
1st author	Naseem Cassim		05/10/2020
2nd author	Ahsan Ahmad	Agreement provided (email)	05/10/2020
3rd author	Reubina Wadee		05/10/2020

4th author	Jaya A George		05/10/2020
5th author	Deborah K Glencross		05/10/2020

Comments by supervisors:

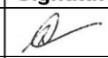
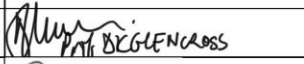
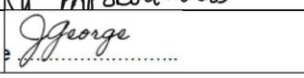
This published work forms part of a series of papers submitted by candidate, N. Cassim, in fulfilment of his PhD degree.

Article 2:

Title: Prostate Cancer age-standardised incidence increase between 2006 and 2016 in the Gauteng province, South Africa: a laboratory data-based analysis

S Afr Med J. 2020;():-

Authors: Naseem Cassim, Ahsan Ahmad, Reubina Wadee, Timothy R Rebbeck, Deborah K Glencross and Jaya A George

Authors	Name	Signature	Date
1st author	Naseem Cassim		05/10/2020
2nd author	Ahsan Ahmad	Agreement provided (email)	05/10/2020
3rd author	Reubina Wadee		05/10/2020
4th author	Timothy Rebbeck	Agreement provided (email)	05/10/2020
5th author	Deborah K Glencross		05/10/2020
6th author	Jaya A George		05/10/2020

Comments by supervisors:

Again, this paper is one of a series of papers submitted by the candidate, N. Cassim, for fulfilment of the requirements of his PhD.

APPENDIX G - SIMILARITY REPORT

a0042716:Prostate_Biopsy_Paper_21_August_2019.pdf

by Naseem Cassim

Submission date: 21-Aug-2019 10:25AM (UTC+0200)
Submission ID: 1161969515
File name: -4aa2-b142-b0877b61c2ed_Prostate_Biopsy_Paper_21_August_2019.pdf (469.52K)
Word count: 7649
Character count: 39958

Introduction

Prostate cancer (PCa) is a major public health concern as life expectancy is increasing globally [1, 2]. In 2018, global PCa incidence was the second highest after lung cancer [3]. PCa incidence varied by up to 189-fold across 186 countries, with age-standardised incidence rates (ASIR) ranges recorded varying widely from 1.0 in Bhutan to as high as 189.1 per 100,000 in Guadeloupe in 2018 [4]. Comparatively higher ASIR were reported for some less developed regions such as Barbados (129.3), Bahamas (85.8), South Africa (68.0), and Reunion (63.7) per 100,000 respectively [4]. PCa ASIR is highest amongst men of African descent, reflecting racial and genetic predisposition in these countries [5]. This could also be due to varying PCa screening and testing patterns that affect incidence [6].

Limited local data reported that Black Africans present with more aggressive PCa with 36% reporting a histopathological review-derived Gleason score ≥7 compared to 17% for African Americans [7]. The GS is the strongest prognostic factor for PCa treatment [8]. Local guidelines categorise PCa risk using the GS as follows; (i) 2-6 low, (ii) 7 intermediate and (iii) 8-10 high [9, 10]. Limited PCa biopsy data, however, exist for Black African men, with most local studies reporting data for selected health facilities [7, 11, 12]. Currently, aside from local cancer registry data, which reported national 2014 data by 2018, there is no comprehensive public sector PCa data for the Gauteng province [13].

In South Africa, PCa is the leading neoplasm in men [3, 4]. Risk factors include age, race (Black African) and family history [1, 9, 10, 14]. The National Health Laboratory Service (NHLS) provides diagnostic services to 80% of the South African public [15]. Routinely accumulated data through analysis of diagnostic samples offers a unique opportunity to describe PCa histological characteristics with the related demographic detail in local populations. The Systematized Nomenclature of Medicine (SNOMED) clinical terms (CT) topography (T) and morphology (M) codes captured routinely based on the histological findings could be used to represent relevant clinical information consistently and

a0042716:Prostate_Biopsy_Paper_21_August_2019.pdf

ORIGINALITY REPORT

12%	9%	7%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.science.gov Internet Source	1%
2	journals.plos.org Internet Source	1%
3	Submitted to University of Greenwich Student Paper	1%
4	Submitted to Saint Joseph's College of Maine Student Paper	1%
5	academic.oup.com Internet Source	1%
6	onlinelibrary.wiley.com Internet Source	<1%
7	www.sajsm.org.za Internet Source	<1%
8	www.tandfonline.com Internet Source	<1%
9	epub.uni-regensburg.de Internet Source	<1%

10	www.thieme-connect.de Internet Source	<1%
11	Timothy R. Rebbeck. "Prostate Cancer Genetics: Variation by Race, Ethnicity, and Geography", Seminars in Radiation Oncology, 2017 Publication	<1%
12	open.uct.ac.za Internet Source	<1%
13	Submitted to University of Sunderland Student Paper	<1%
14	Submitted to University of Cape Town Student Paper	<1%
15	cebp.aacrjournals.org Internet Source	<1%
16	dash.harvard.edu Internet Source	<1%
17	Submitted to University of Stellenbosch, South Africa Student Paper	<1%
18	Submitted to University of Limpopo Student Paper	<1%
19	www.ncbi.nlm.nih.gov Internet Source	<1%

APPENDIX H - DETAILED INFORMATION ON THE HIV M&E DASHBOARD

The table below describes the section heading, report heading, indicator and information reported in the HIV M&E dashboard.

Section Heading	Report heading	Indicator	Information reported
Indicators past 12 months	Detectable vs Undetectable VL Tests	% Detectable HIV viral load (>1000 copies/ml)	Pie chart indicated the percentage of patients with a detectable and undetectable HIV viral load
	Viral Load Tests (copies/ml)	HIV viral load test ranges: (i) <400, 400 to ≤1000 and >1000 copies/ml	Pie chart indicated the percentage of patients with a HIV viral load in the three test ranges
	Distribution of VL tests by sex	Number of patients with a sex of male, female or unknown	Pie chart indicated the sex distribution
	Distribution of VL tests by age groups	Number of patients by age category: (i) <15, (ii) 15 -19, (iii) 20 - 44 and (iv) ≥45.	Pie chart indicated the age category distribution
Indicators for VL Tests by Sex Past 12 Months	Detectable vs Undetectable VL Tests	% Detectable HIV viral load (>1000 copies/ml) for males and females	Bar chart indicated the percentage of males and females with a detectable and undetectable HIV viral load
	Viral Load Tests (copies/ml)	HIV viral load test ranges reported for males and females: (i) <400, 400 to ≤1000 and >1000 copies/ml	Bar chart indicated the percentage of males and females with a HIV viral load in the three test ranges
	Age distribution	Number of males and females by detailed age category: (i) <15, (ii) 15 - 19, (iii) 20 - 24, (iv) 25 - 29, (v) 30 - 34, (vi) 35 - 39, (vii) 40 - 44, (viii) 45 - 49, (ix) 50 - 54, (x) 55 - 59 and (xi) ≥60.	Bar chart indicated the percentage of males and females reported for the detailed age category
Monthly Indicators for Females Past 12 Months	Detectable vs Undetectable VL Tests	% Detectable HIV viral load (>1000 copies/ml)	Bar chart indicated the percentage of patients with a detectable and undetectable HIV viral load for the last twelve months
	Viral Load Tests (copies/ml)	HIV viral load test ranges reported for females: (i) <400, 400 to ≤1000 and >1000 copies/ml	Bar chart indicated the percentage females with a HIV viral load in the three test ranges

Section Heading	Report heading	Indicator	Information reported
	Age distribution	Number of females by detailed age category: (i) <15, (ii) 15 - 19, (iii) 20 - 24, (iv) 25 - 29, (v) 30 - 34, (vi) 35 - 39, (vii) 40 - 44, (viii) 45 - 49, (ix) 50 - 54, (x) 55 - 59 and (xi) ≥60.	Bar chart indicated the percentage of females reported for the detailed age category

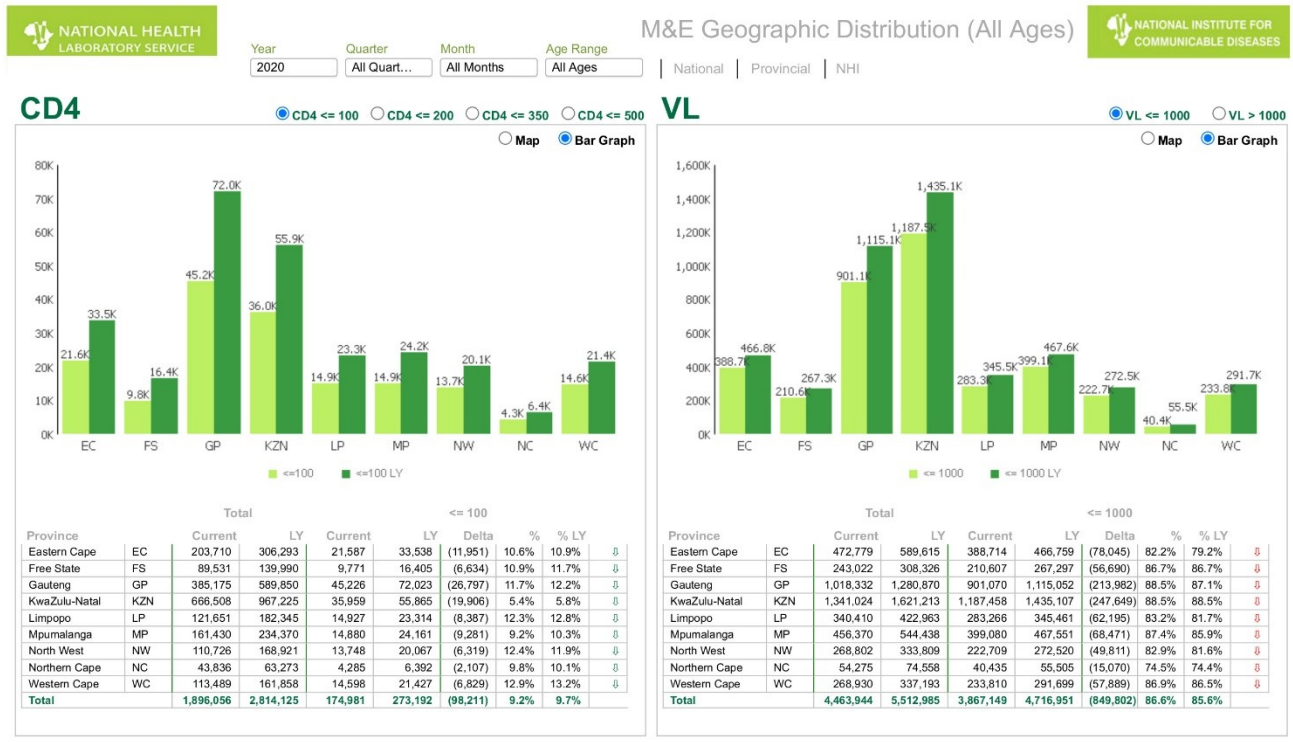
APPENDIX I – OTHER HIV M&E DASHBOARDS

Dashboard title: M&E Geographic Distribution (All Ages)

Information reported:

For both CD4 and HIV viral load testing, the dashboard depicted below reports the percentage of people with a CD4 for defined test ranges. For example, a CD4 ≤200 cells/µl indicates advanced disease. Similarly, a CD4 ≤100 cells/µl is indicative of very advanced disease. Data is reported for two twelve-month periods to indicate trends over time. The dashboard user is able to toggle the ranges, drill from national to a provincial data as well customise which ages to report. The data is reported as both a bar chart and a table. The bar chart can also be displayed as a choropleth map.

Screenshot



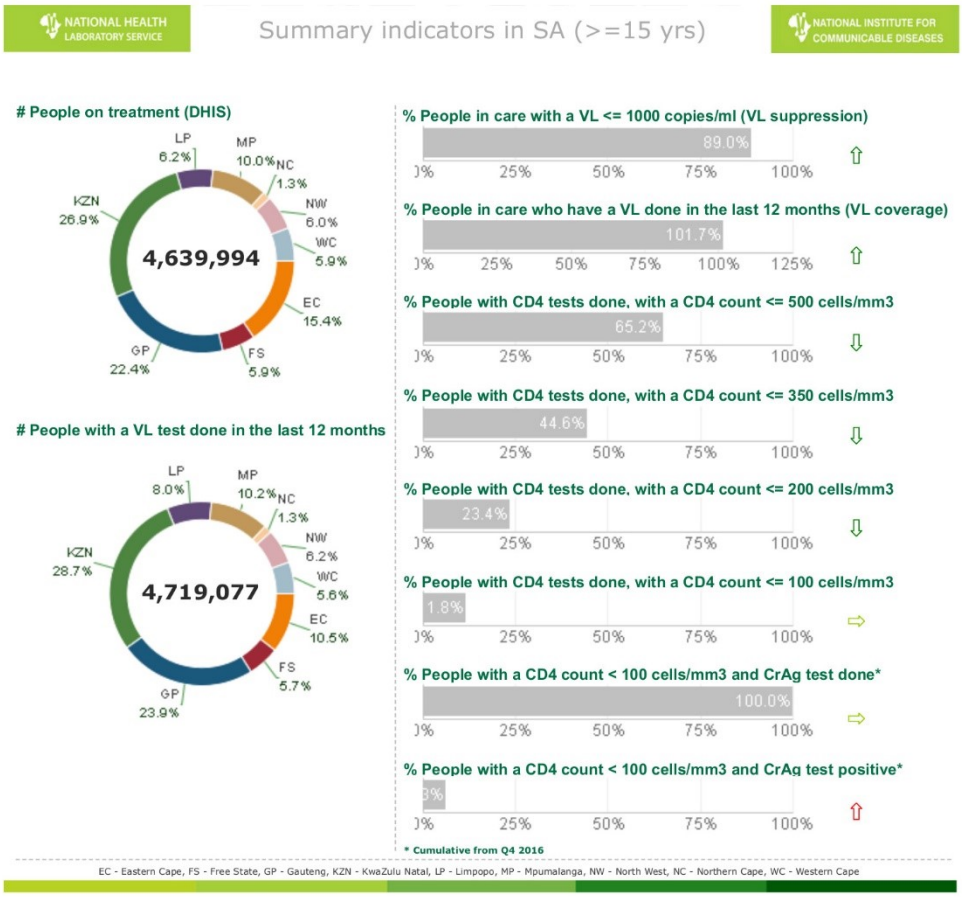
(National Institute for Communicable Disease (NICD) – M&E Geographic Distribution (All Ages) – August 2020)

Dashboard title: Summary indicators in SA (>=15 yrs)

Information reported:

This dashboard integrates the number of people receiving ART that is obtained from aggregate data collate through the District Health Information System (DHIS). Laboratory data is then de-duplicated to create an HIV cohort to report on progress in South Africa for the last 90 target [247]. The indicators reported include the number of people that have received a HIV viral load in the last 12 months and the percentage of people in care with a HIV viral load ≤1000 copies/ml (virological suppression). In addition, there are various CD4 and cryptococcal antigenemia (CrAg) testing indicators that are related to advanced HIV disease, e.g. number of people with a CD4 <100 cells/µl. This dashboard provides important insights on the progress in South Africa to meet the target of 90% of people on ART being virologically suppressed.

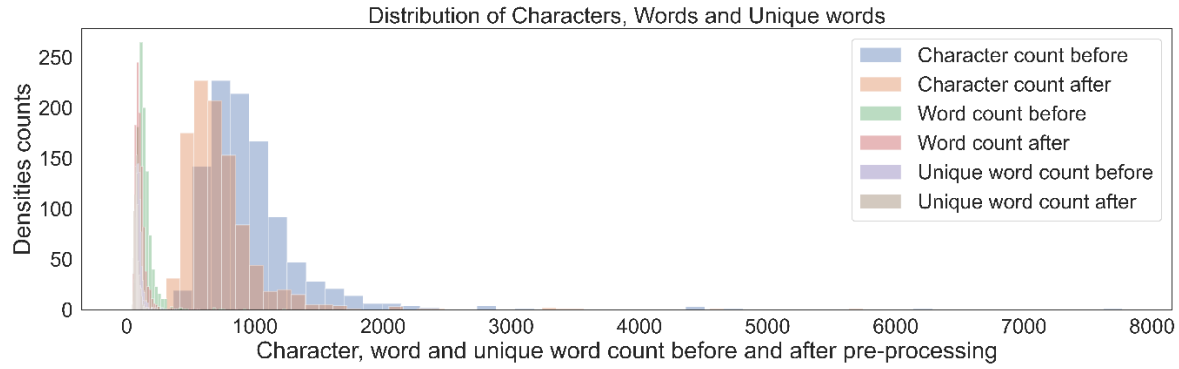
Screenshot



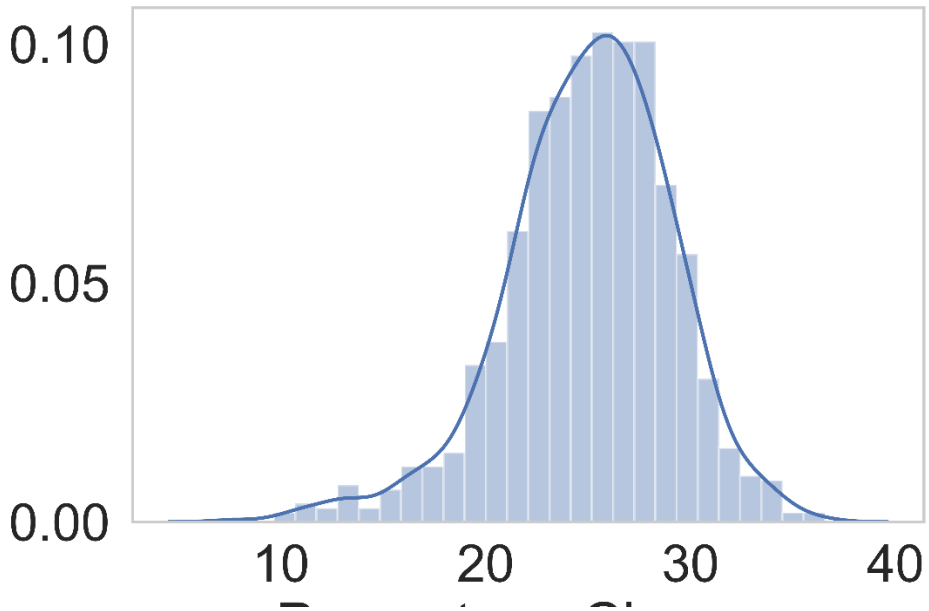
(National Institute for Communicable Disease (NICD) – Summary indicators in SA (>=15 yrs) – August 2020)

APPENDIX J – ADDITIONAL VISUALISATIONS FOR THE TEXT MINING ALGORITHM

Distribution of characters, words and unique words before and after pre-processing of the narrative prostate biopsy reports



Pre-processing percentage change



Pairplot depicting the relationship between characters, words and unique words before and after pre-processing of the narrative prostate biopsy reports.

