MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190472

Using Big Data Techniques to Improve Prostate Cancer Reporting in the Gauteng Province, South Africa

Cassim N[a,b], Mapundu M[c], Olago V[c,d], George JA[b,e] and Glencross DK[a,b]

[a] *Department of Molecular Medicine and Haematology, University of the Witwatersrand, Johannesburg, South Africa*
[b] *National Health Laboratory Service (NHLS), Johannesburg, South Africa*
[c] *School of Public Health, Faculty of Health Sciences, University of Witwatersrand, Johannesburg, South Africa*
[d] *National Health Laboratory Service (NHLS), National Cancer Registry (NCR), Johannesburg, South Africa*
[e] *Department of Chemical Pathology, Faculty of Health Sciences, University of Witwatersrand, Johannesburg, South Africa*

Abstract

Prostate cancer (PCa) data is of public health importance in South Africa. Biopsy data is recorded as semi-structured narrative text that is not easily analysed. Our study reports a pilot study that applied predictive analytics and text mining techniques to extract prognostic information that guides patient management. In particular, the Gleason score (GS) reported in a number of formats were extracted successfully. Our study reports that predominantly older men were diagnosed with PCa reporting a high-risk GS (8-10). Where cell differentiation was reported, 64% of biopsies reported poor differentiation. The approaches demonstrated in our study should be extended to a larger dataset to assess whether it has the potential to scale up to the national level.

Keywords:

Prostate cancer, Gleason score, Risk, Cell differentiation

Introduction

Prostate cancer (PCa) is an important non-communicable disease (NCD) in South Africa with a reported age-standardised incidence rate of (ASIR) of 67.9 per 100 000 in 2012 [1]. Local studies have reported that African men present with advanced and aggressive PCa reducing the opportunity for remission [2]. Local guidelines indicate that patient management is directly linked to Gleason score (GS) grading, percentage and numbers of cores as well as ancillary information such as seminal vesicle involvement and peri-neural invasion [3]. Prostate biopsy results are stored as semi-structured narrative text that cannot be easily analyzed. The GS is captured as follows; (i) major score, e.g. 4 and (ii) minor score, e.g. 3. This equates to a 4+3=7 GS. This data is captured in multiple formats depending on reporting practices at each laboratory. For example, a GS of 3+4=7 could be reported as follows; (i) major 3 and minor 4 = 7, (ii) 3+4=7, (iii) (major pattern 3 + minor pattern 4) = 7, (iv) 3+4=7 and (v) (3+4)=7. The number of cores, GS and cell differentiation provides valuable insights to assess late presentation and poor prognosis [3]. The objective of our study was to investigate the use of big data analytics (text mining) to extract meaningful data from narrative prostate biopsy results.

Methods

The retrospective descriptive study design was used to analyse laboratory data between 2006 and 2016 for men 30 years and older. The sample population of 1000 cases was randomly selected from prostate biopsies with an adenocarcinoma histological finding determined using Systematized Nomenclature of Medicine (SNOMED) morphology codes [4]. Convenience sampling was used. Local guidelines were used to categorise PCa risk using the GS as follows; (i) GS 2-6: low-risk disease (LRPCa), (ii) GS 7: intermediate-risk disease (IRPCa) and (iii) GS 8-10: high-risk disease (HRPCa) [3]. An example of a fictitious biopsy narrative report is provided:

> EPISODE NUMBER: ABC1234 SPECIMEN DETAILS: PROSTATE BIOPSIES. CLINICAL DETAILS: THE PATIENT IS 66-YEAR-OLD MALE WITH AN ENLARGED PROSTATE AND A PSA LEVEL OF 20.8. PROSTATE CORE BIOPSIES WERE SUBMITTED TO EXCLUDE CARCINOMA OF THE PROSTATE. MACROSCOPY: SEVEN CORE BIOPSIES ARE RECEIVED, RANGING IN LENGTH FROM 25MM TO 10MM. PATHOLOGICAL DIAGNOSIS: PROSTATE CORE BIOPSIES, WITH REPRESENTATION OF THE SEMINAL VESICLE: INVASIVE MODERATELY DIFFERENTIATED PROSTATIC ADENOCARCINOMA. GLEASON SCORE = MAJOR 3 + MINOR 4 = 7 THERE IS PERINEURAL INVASION IDENTIFIED. NO LYMPHOVASCULAR SPACE INVASION IS PRESENT IN THE SECTIONS EXAMINED. THE TUMOUR IS PRESENT IN 2 OUT OF 7 CORE BIOPSIES WITH AN APPROXIMATE BULK OF 20%.

Figure 1: Fictitious prostate biopsy narrative report

This study applied text mining techniques and specifically made use of regular expressions to address the complex relationships among the data [5], including structured, unstructured and semi-structured data [6]. Prostate biopsy reports data were extracted and loaded to the Spyder Integrated Development Environment (IDE) which uses the Python programming language [7]. Spyder IDE offers features such as advanced editing, debugging, profiling, data exploration and interactive execution [8-10]. The data was first pre-processed and cleaned using tokenisation and stop word removal. Stemming was used to derive the root word. Using feature generation, the bag of words was developed and n-grams generated. Feature selection was used to develop a vector space by selecting a subset of features from the biopsy reports. Regular expressions representative of the GS target feature such

as "gleason", "Gleason", "GLEASON", "Gleeson", etc. were used to identify the target feature. Lastly, the required data was then extracted for each biopsy and produced as an output extract that was analysed and visualised. The output variables included; (i) episode number (ABC1234), (ii) age (66), (iii) cell differentiation type (moderate), (iii) number of cores (7), (iv) type on cellular invasion (Perineural) and (v) GS (3+4=7). The ages extracted from the biopsy report were correlated against the manually captured values. The data reported includes; (i) GS, (ii) percentage of biopsies by cell differentiation and (iii) distribution of the number of cores submitted. Age were categorised as follows; (i) <50, (ii) 50-59, (iii) 60-69, (iv) 70+ and (v) "Age not stated".

Results

The episode number was extracted for all biopsies (100%). The age was extracted for 893 biopsies (89%). An age was captured for 877/1000 (88%) of biopsies compared to 956/1000 (91%) in the LIS. There were 17 biopsies where different ages were recorded (2%); 3/17 likely transcription errors, e.g. 6 recorded instead of 60. The majority of biopsies were requested for men ≥60 years (86%).

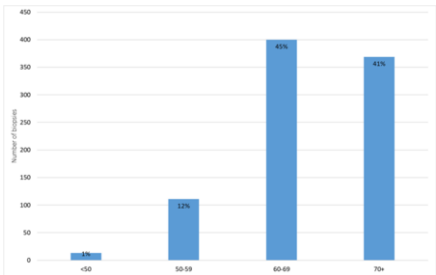

Figure 2: Distribution of the extracted age

Of 1000 biopsies, 286 reported cell differentiation characteristics (29%). Poorly differentiated adenocarcinoma was reported for 64% of biopsies followed by 29% for moderate differentiation.

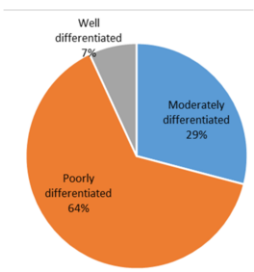

Figure 3: Cell differentiation distribution

The GS was extracted for all 1000 biopsies. A GS of 3+3=6 and 5+4=9 was reported for 18% of biopsies each (total of 26%). This was followed by the 4+3=7 GS (16%). Half of the patients reported a high-risk GS (8-10: 47%).

Conclusion

The use of predictive analytics holds tremendous potential especially in a resource-poor setting where documented clinical data is poor or absent. The prognostic information could be used to categorise patients into risk groups. The accrued laboratory

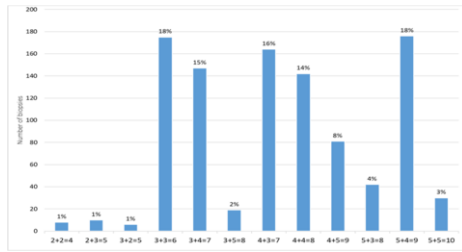

Figure 4: Distribution of Gleason score risk category

data could enable an understanding of PCa differentiation and GS by unlocking clinical information captured in the narrative reports. This could enable real-time cancer data analysis (as opposed to cancer registry reporting that is delayed). The approach reported in our study should be extended to a larger dataset to assess the ability to scale up to the national level.

Acknowledgements

The authors would like to thank the Academic Affairs, Research and Quality Assurance (AARQA) department for providing access to the data.

References

[1] International Agency for Research on Cancer, *Cancer Fact Sheets: Prostate Cancer*, International Agency for Research on Cancer, Geneva, Switzerland, 2012.

[2] H.A. Le Roux, R.J. Urry, B. Sartorius, and C. Aldous, Prostate cancer at a regional hospital in South Africa: we are only seeing the tip of the iceberg, *S Afr J Surg* **53** (2015), 57-62.

[3] D. Anderson *et al.*, *South African Prostate Cancer Guidelines*, South African Urological Association, Pretoria, South Africa, 2017.

[4] U.S. National Library of Medicine, *SNOMED CT International Edition*, U.S. National Library of Medicine, Bethesda, MD, USA, 2016.

[5] X. Wu, X. Zhu, G. Wu, and D. Wei, Data Mining with Big Data, *IEEE Trans Knowl Data Eng* **26** (2014), 97-107.

[6] A. Oussous, F.Z. Benjelloun, A.A. Lahcen, and S. Belfkih, Big Data technologies: a survey, *J King Saud Univ Comp Inf Sci* **30** (2017), 431-448.

[7] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly, Cambridge, MA, USA, 2009.

[8] V. Korde and C.N. Mahender, Text classification and classifiers: a survey, *Int J Artif Intel Appl* **3** (2012), 85-99.

[9] M. Hofmann and A. Chisholm, *Text Mining and Visualization: Case Studies Using Open-Source Tools*, CRC Press, Boca Raton, FL, USA, 2016.

[10] C.D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

Address for correspondence

Naseem Cassim
Room 3B10, Faculty of Health Sciences
University of the Witwatersrand
7 York Road, Parktown, 2193, South Africa
Email: naseem.cassim@wits.ac.za