Using Text Mining Techniques to Extract Prostate Cancer Predictive Information (Gleason Score) from Semi-structured Narrative Laboratory Reports in the Gauteng Province, South Africa

Naseem Cassim (naseem.cassim@wits.ac.za)
Department of Molecular Medicine and Haematology, Faculty of Health Sciences, University of Witwatersrand and National Health Laboratory Service (NHLS), 7 York Road, Parktown, Johannesburg

Michael Mapundu
School of Public Health, Faculty of Health Sciences, University of Witwatersrand, 7 York Road, Parktown, Johannesburg

Victor Olago
National Health Laboratory Service (NHLS), National Cancer Registry (NCR), 1 Modderfontein Road, Sandringham, Johannesburg

Turgay Celik
School of Electrical & Information Engineering and Wits Institute of Data Science, University of Witwatersrand, 1 Jan Smuts Avenue, Braamfontein, Johannesburg

Jaya Anna George
Department of Chemical Pathology, Faculty of Health Sciences, University of Witwatersrand and National Health Laboratory Service (NHLS), 7 York Road, Parktown, Johannesburg

Deborah Kim Glencross
Department of Molecular Medicine and Haematology, Faculty of Health Sciences, University of Witwatersrand and National Health Laboratory Service (NHLS), 7 York Road, Parktown, Johannesburg

Research Article

Keywords: Prostate cancer, Gleason score, late presentation, text mining, algorithm, public health.

Posted Date: August 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-778832/v1

Version of Record: A version of this preprint was published at BMC Medical Informatics and Decision Making on November 25th, 2021. See the published version at https://doi.org/10.1186/s12911-021-01697-2.

Abstract

Background:

Prostate cancer (PCa) is the leading male neoplasm in South Africa with an age-standardised incidence rate of 68.0 per 100,000 population in 2018. The Gleason score (GS) is the strongest predictive factor for PCa treatment and is embedded within semi-structured prostate biopsy narrative reports. The manual extraction of the GS is labour-intensive. The objective of our study was to explore the use of text mining techniques to automate the extraction of the GS from irregularly reported text-intensive patient reports.

Methods:

We used the associated Systematized Nomenclature of Medicine clinical terms morphology and topography codes to identify prostate biopsies with a PCa diagnosis for men aged >30 years between 2006 and 2016 in the Gauteng Province, South Africa. We developed a text mining algorithm to extract the GS from 1,000 biopsy reports with a PCa diagnosis from the National Health Laboratory Service database and validated the algorithm using 1,000 biopsies from the private sector. The logical steps for the algorithm were data acquisition, pre-processing, feature extraction, feature value representation, feature selection, information extraction, classification, and discovered knowledge. We evaluated the algorithm using precision, recall and F-score. The GS was manually coded by two experts for both datasets. The top five GS were reported, with the remaining scores categorised as “Other” for both datasets. The percentage of biopsies with a high-risk GS (≥8) was also reported.

Results:

The first output reported an F-score of 0.99 that improved to 1.00 after the algorithm was amended (the GS reported in clinical history was ignored). For the validation dataset, an F-score of 0.99 was reported. The 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% compared to 17.4% for the validation dataset.

Conclusions:

We demonstrated reliable extraction of information about GS from narrative text-based patient reports using an in-house developed text mining algorithm. A secondary outcome was that late presentation could be assessed.

Background

Globally, prostate cancer (PCa) is an important non-communicable disease (NCD) due to both population growth and a concomitant increase in life expectancy [1, 2]. It is the leading male neoplasm in South Africa with an age-standardised incidence rate (ASIR) of 68.0 per 100,000 population in 2018 [3].

Local treatment guidelines indicate that men with PCa are assigned to risk categories using the prostate specific antigen (PSA) result, Gleason score (GS) and clinical stage [4, 5]. The GS is based on the predominant histological pattern noted across all prostate biopsy samples submitted for anatomical pathology (AP) review, with a score of 1 reflecting the presence of normal cells and incremental mutational (grade) malignant change reflected in a score of 2 to 5. Within the scoring system, the first GS reflects the predominant cell pattern whereas the second Gleason grading is determined by the second most predominant pattern. 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 $\leq$ 6: low-risk, (ii) GS = 7: intermediate-risk and (iii) GS $\geq$ 8: high-risk [4, 5]. Patients with a high-risk GS have a poorer prognosis with an increased risk of metastatic progression and death [6]. For these patients, the PCa mortality risk is 60 to 87% compared to between 42 and 70% for an intermediate-risk GS [6].

Across the National Health Laboratory Service (NHLS), a laboratory information system (LIS) is used to record, manage, and store patient laboratory reports and related demographic health data [7, 8]. This LIS documents all processes within the laboratory workflow including sample registration, test order generation, tracking orders and reporting results [7, 8]. For AP reporting, the assigned pathologist voice-records the biopsy narrative report for electronic capture by data typists. These narrative AP reports are not standardised and are pathologist dependent in terms of patient history, pathological tumour/biopsy description and language used. As a result, these are irregularly reported text-intensive patient reports. Table 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 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.
PSA: prostate specific antigen MM: millimetre P63: Protein 63 CK5/6: Cytokeratin 5/6	

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 may be captured 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 [9]. 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 [9]. These vocabularies and ontologies work well with coding systems such as International Classification of Diseases for Oncology (ICD-O-3), Systemized Nomenclature of Medicine (SNOMED) Clinical Terms (CT) and International Classification of Diseases Tenth Revision (ICD-10) for example [9]. Such vocabularies and ontologies do not exist for the GS. As a result, the manual coding of the GS is time-consuming resulting in a paucity of local data describing late presentation in South Africa.

Globally, artificial intelligence (AI) has been used to automate decision making through mimicking human cognitive function by using mathematical, statistical, logical, and computer programming approaches [10, 11]. The AI model can be trained using existing data and applied to new data to automate decisions [11]. AI can also be applied to semi-structured healthcare data using techniques such as natural language processing (NLP) [10]. This is achieved by employing computational techniques to extract semantic meaning from text [12, 13]. In essence, these NLP procedures convert text to machine-readable structured data [10]. This includes computational approaches such as tokenisation that help to identify words and punctuations within a sentence [13]. In summary, NLP can be used to extract clinical information from unstructured data to supplement and enrich structured medical data [10].

There is a need to develop automated algorithms that can extract the GS from narrative prostate biopsy reports. The objective of our study was to explore the use of text mining techniques to extract the predictive GS from narrative prostate biopsy reports.

Methods

All methods were carried out in accordance with relevant guidelines and regulations of the Human Research Ethics Committee (Medical) at the University of the Witwatersrand (Faculty of Health Sciences).

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 [14, 15]. 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) os, (ii) pandas, (iii) time, (iv) matplotlib, (v) seaborn, (vi) WordCloud and, (vii) Natural Language Toolkit (NLTK). We followed the text mining pipeline as depicted in the flowchart below (Fig. 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.

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. The narrative prostate biopsy reports are captured as free-text in the LIS and stored in the national laboratory data repository. The Systematised Nomenclature of Medicine (SNOMED) clinical terms (CT) dataset was used to develop lookup tables to identify biopsies with an adenocarcinoma histological finding (n = 8,201) [16]. Once the biopsies with PCa were identified (adenocarcinoma histological findings with a reported GS), we extracted a random sample of 1,000 biopsies using Microsoft Excel (Redmond, Washington, USA) [17]. We chose a random sample as we did not want to select biopsies that were reported in a similar fashion from one laboratory.

To evaluate the text mining algorithm, we also randomly extracted 1,000 prostate biopsy narrative reports with a PCa diagnosis that were submitted from private sector laboratories to the National Cancer Registry (NCR) (referred to as the validation dataset). These narrative reports are generated by various private sector pathology practices and could be used to validate the algorithm. We received only the narrative biopsy reports.

For both datasets, the GS were manually coded by two experts. Manual coding was required as the GS is not extracted by the NCR and is embedded within the narrative report. Following this, a random sample of 369 biopsies were independently verified to validate the manual coding.

Pre-processing

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 a corpus). A corpus 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 [14, 18, 19]. Next, the data cleaning process involved using the NLP tokenization, stopwords removal and stemming techniques [15, 19]. We used tokenization to condense the streams of text into smaller meaningful elements (called tokens) that comprised of words, phrases and symbols. For example, the words ‘do not stop’ would result in 3 tokens (do-not-stop). 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, e.g. the, is, at, etc. This removed all possible English stopwords. We also converted text to lowercase for standardisation.

Feature extraction

We extracted 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* [9, 20]. Next, we used N-grams as our

feature extraction strategy to extract the major and minor Gleason 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, i.e. sequence of n words. N-grams are a set of co-occurring terms that were reported in a sentence or paragraph in the corpus [21, 22]. For example, when n = 1 (unigram) this represents single words in a sentence [22]. 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 [22]. From the N-grams generated, we extracted the GS feature for each biopsy. The N-gram feature extraction output is provided for a sample of biopsies (Table 2)

Table 2
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']	

Feature value representation

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. We reported the twenty most frequently

occurring unigrams, bigrams, trigrams and quadgrams as horizontal bar graphs (Fig. 2).

Feature selection

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 chose relevant features and reduced the feature space (without using dimensionality reduction). Reducing the number of features selected would improve the model performance. Even though the feature space was reduced, there was no loss of information [23].

Information extraction

Information extraction is used to select specific entities and relationships of interest [9]. 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.

Classification

We classified biopsies into the three risk categories: (i) low (≤ 6), (ii) intermediate (7) and (iii) high-risk (≥ 8) based on local guidelines [5]. The classification process was automated using a rule-based approach and implemented within the algorithm.

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 1, the following discovered knowledge was reported: (i) ABC1234, (ii) 4, (iii) 3, (iv) 7, (v) 4 + 3 = 7 and (vi) intermediate.

Text Mining Algorithm Evaluation

A confusion matrix (also known as a sensitivity/specificity analysis) was used to compare the text mining algorithm extracted against the manually coded values [24]. The confusion matrix consists of four values: (i) True Positives (TP): correctly extracting the GS, (ii) True Negatives (TN): correctly extracting a biopsy without a GS, (iii) False Positive (FP): falsely extracting a GS and (iv) False Negative (FN): falsely extracting the manually coded GS [24]. The precision and recall are calculated using these four values as follows: (i) $\frac{TP}{TP+FP}$ and (ii) $\frac{TP}{TP+FN}$ respectively. Precision and recall are similar to positive predictive value (PPV) and sensitivity respectively. The F-score is the harmonic mean of precision and recall and is calculated using the formula $\frac{2*(Recall*Precision)}{(Recall+Precision)}$. As the manually coded values were assumed to be the gold standard without any incorrect values, there was no need to report the zero values. Therefore, we removed the 'Actual: No' column. However, the zero values were still used for the F-score calculation.

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 categorised 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) [25].

Results

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

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 = 1,754). 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.

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). On manual inspection of the N-grams (Table 2), we identified that two different GS were reported in both the clinical history and pathological diagnosis for 16 biopsies (example '3 + 2', '3 + 3', '3 + 3' in Table 2). The algorithm was updated 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
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		

Gleason score formats reported

We identified ten different GS reporting formats (Table 4). The variations in reporting included: (i) use of both the equal sign as well as the word equals, (ii) use of brackets, (iii) spelling of major and minor (for

example using the word major and pattern), (iv) use of both the words and symbols (plus versus +) and (v) use of colons and commas to separate major and minor scores.

Table 4
Different Gleason score formats reported for the study. The clean extracted score reported, and the original value reported in the prostate biopsy report is indicated for the study dataset.

#	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

Gleason score frequency analysis

The most commonly reported GS were 5 + 4 = 9, 3 + 3 = 6 for 17.6% (n = 176) and 17.5% (n = 175) of biopsies respectively (Table 5). 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. 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 5

The table reported the frequency for the top five reported Gleason scores with the remaining values grouped and reported as “Others”. Data is reported for this study as well as for the separate dataset.

Study Dataset				Validation Dataset		
No.	**Gleason score**	**n=**	**%**	**Gleason score**	**n=**	**%**
1	5 + 4 = 9	176	17.6%	3 + 3 = 6	377	37.7%
2	3 + 3 = 6	175	17.5%	3 + 4 = 7	194	19.4%
3	4 + 3 = 7	164	16.4%	4 + 3 = 7	149	14.9%
4	3 + 4 = 7	147	14.7%	4 + 4 = 8	100	10.0%
5	4 + 4 = 8	142	14.2%	4 + 5 = 9	74	7.4%
6	Others	196	19.6%	Others	106	10.6%
Total		**1000**	**100%**	**Total**	**1000**	**100%**
High- Risk GS ≥ 8		**318**	**31.8%**	**High-Risk GS ≥ 8**	**174**	**17.4%**
GS: Gleason score						

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 6). 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 6
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 (≤ 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		
GS: Gleason score			

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 16/1,000 biopsies GS (1.6%) was 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 (F-score of 0.99). The F-score reported for both datasets is similar to a Perl routine that also used regular expressions to extract the GS [20]. Similar approaches using regular expressions have been reported by two other studies [9, 20].

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 can 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. This could be explained by the racial variation in medical aid coverage [26]. A limitation of this study is the small sample sizes.

As we reported data for a multi-class problem and compared the predicted and manually coded values categorised as low, intermediate and high-risk [5]. The analysis revealed an acceptable macro-averaged F-score indicating that the text mining algorithm was able to accurately classify the GS risk category.

Our findings indicate that the text mining algorithm could be used to reliably extract the GS from laboratory data in similar settings. 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 API [27]. 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 [28]. 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) [28]. Therefore, the techniques employed in our study could be applied to other narrative laboratory data such as immunophenotyping, fluorescence in situ hybridization 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 machine learning (ML) to extract an adenocarcinoma histological finding in an automated fashion [9]. This has the potential to offer near real-time cancer registry type information removing the need for manual coding [29]. 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. Deep learning is composed of multiple processing layers that learn representations of data with multiple levels of abstraction [30]. This approach has dramatically improved AI approaches for visual object recognition and object detection [30]. Deep learning models are able to extract information from large datasets and will continue to improve the knowledge discovery as more data is generated [31]. This enables deep learning to outperform classical ML approaches [31]. One of the benefits is that deep learning can 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 [30].

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 it 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 [32].

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.

Conclusion

Our study has shown that a text mining algorithm can be used to extract the predictive GS from narrative biopsy reports. This could also 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. This approach is applicable to other cancers of public health interest. Furthermore, ML and deep learning approaches should be investigated to replicate results shown for the SNOMED CT lookup tables to address data gaps. These could be used to reduce the delays in the publication of cancer registry data. These algorithms could be moved to a cloud service to extend automated PCa surveillance data generation across Africa.

Abbreviations

AI: Artificial Intelligence; AP: Anatomical Pathology; ASIR: Age-standardised incidence rate; CSV: Comma separated value; CT: Clinical terms; FN: False negative; FP: False positive; GS: Gleason score; ICD: International Classification of Diseases; ICD-0-3: International Classification of Diseases for Oncology – 3rd Edition; ICD-10: International Classification of Diseases –Tenth revision; IDE: Integrated development environment; LIS: Laboratory Information System; ML: Machine Learning; NCD: Non-Communicable Diseases; NCR: National Cancer Registry; NHLS: National Health Laboratory Service; NLP: Natural Language Processing; NLTK: Natural Language Toolkit; PCa: Prostate Cancer; PPV: Positive predictive value; PSA: Prostate specific antigen; SNOMED: Systemized Nomenclature of Medicine; TN: True negative; TP: True positive;

Declarations

Ethics approval and consent to participate

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

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available as the authors do not have permission to share them. Consent to access the datasets should be directed to naseem.cassim@wits.ac.za.

Competing Interests

The authors do not have any competing interests

Funding

No funding was received for this study.

Authors' contributions

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.

Acknowledgement

The authors would like to acknowledge the anatomical pathologists that generated the narrative reports.

Consent for publication

Not applicable.

References

1. 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
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):851-8.doi:https://doi.org.10.1007/s00345-016-1953-9
3. 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
4. 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).
5. 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).

6. 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).
7. 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
8. 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
9. 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
10. 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
11. 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
12. Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015;349(6245):261-6.doi:https://doi.org.10.1126/science.aaa8685
13. 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
14. Seaborn. seaborn: statistical data visualization 2018 [Available from: https://seaborn.pydata.org/] (Accessed Date: 17 December 2018).
15. 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).
16. 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
17. 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).
18. 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).

19. The Matplotlib development team. Matplotlib: The Matplotlib development team. 2018 [Available from: https://matplotlib.org/] (Accessed Date: 17 December 2018).
20. 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
21. 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
22. 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
23. 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).
24. 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.
25. 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).
26. 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).
27. Spyder Project Contributors. Spyder Python Integrated Development Environment (IDE) 2018 [Available from: https://www.spyder-ide.org/] (Accessed Date: 31 October 2018).
28. 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
29. 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
30. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.doi:https://doi.org.10.1038/nature14539
31. 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
32. 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

Figures

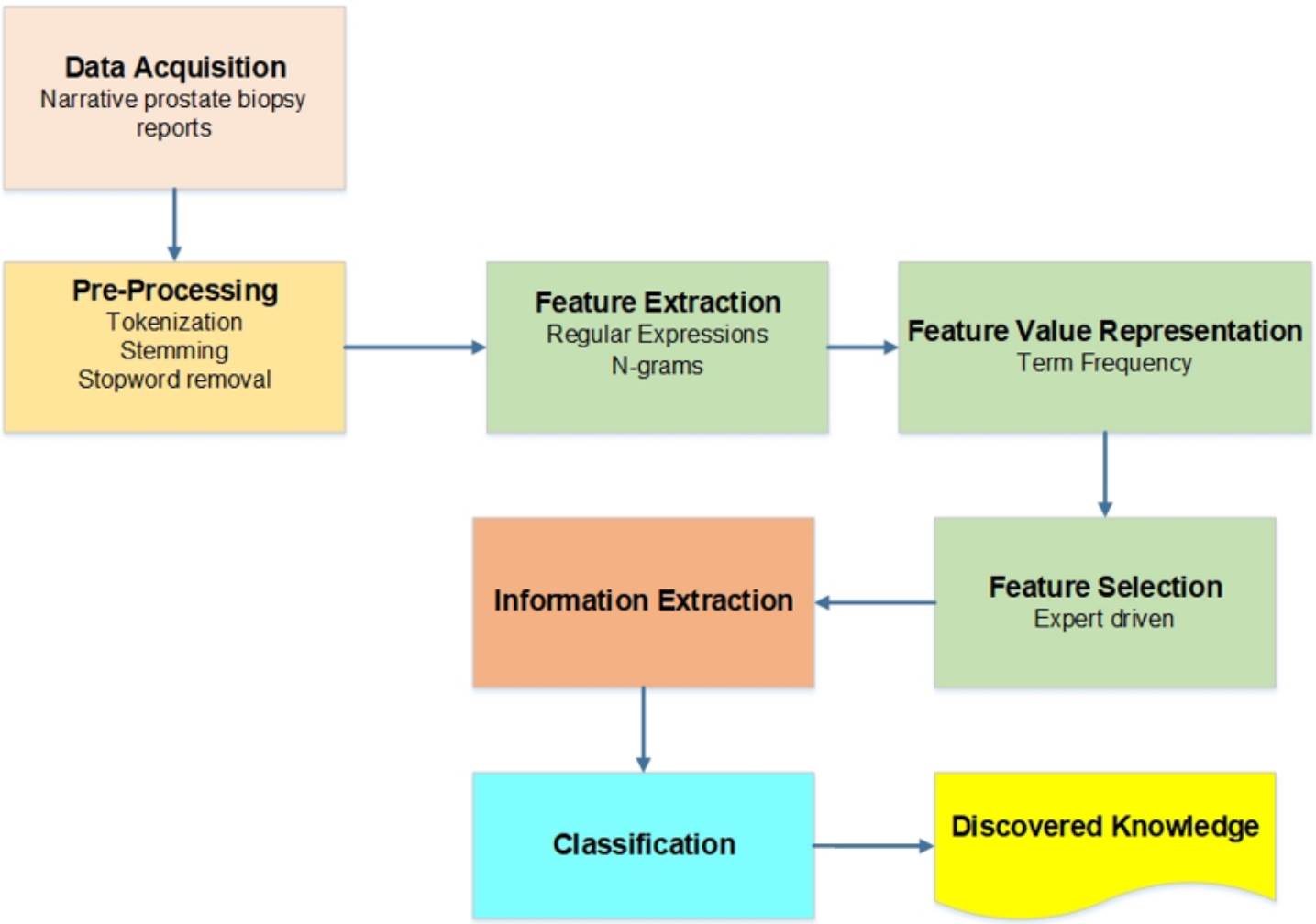

Figure 1

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.

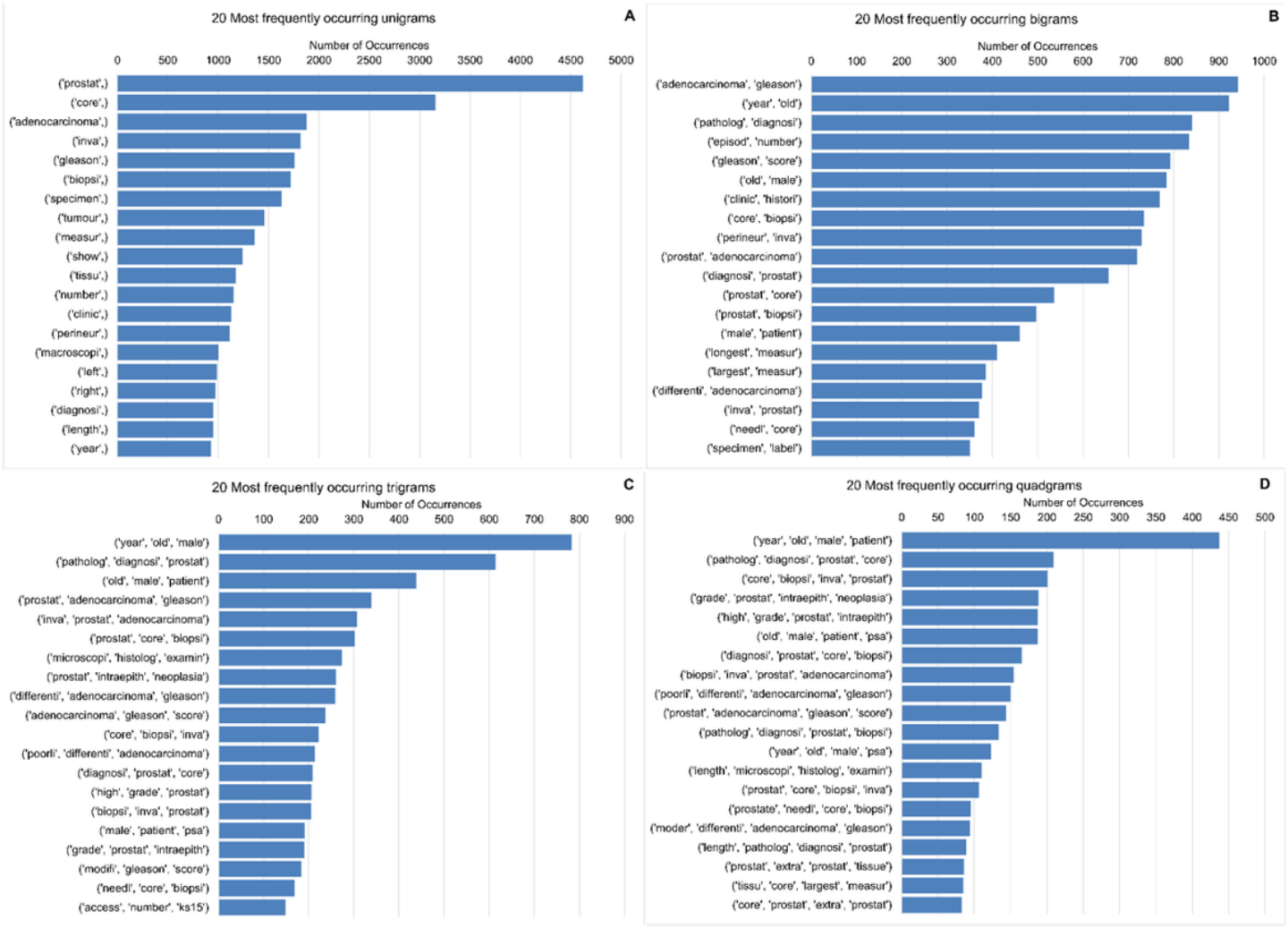

Figure 2

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.