Journal of Pathology Informatics 13 (2022) 100154

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

An accessible, efficient, and accurate natural language processing method for extracting diagnostic data from pathology reports

Hansen Lam, Freddy Nguyen, Xintong Wang, Aryeh Stock, Volha Lenskaya, Maryam Kooshesh, Peizi Li, Mohammad Qazi, Shenyu Wang, Mitra Dehghan, Xia Qian, Qiusheng Si, Alexandros D. Polydorides *

Department of Pathology, Molecular and Cell-Based Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA

ARTICLE INFO

Keywords:
Unstructured
Free-text
Narrative
Extraction
XML
Carcinoma
Algorithm
Python

ABSTRACT

Context: Analysis of diagnostic information in pathology reports for the purposes of clinical or translational research and quality assessment/control often requires manual data extraction, which can be laborious, time-consuming, and subject to mistakes.

Objective: We sought to develop, employ, and evaluate a simple, dictionary- and rule-based natural language processing (NLP) algorithm for generating searchable information on various types of parameters from diverse surgical pathology reports.

Design: Data were exported from the pathology laboratory information system (LIS) into extensible markup language (XML) documents, which were parsed by NLP-based Python code into desired data points and delivered to Excel spreadsheets. Accuracy and efficiency were compared to a manual data extraction method with concordance measured by Cohen's κ coefficient and corresponding *P* values.

Results: The automated method was highly concordant (90%–100%, $P<.001$) with excellent inter-observer reliability (Cohen's κ: 0.86–1.0) compared to the manual method in 3 clinicopathological research scenarios, including squamous dysplasia presence and grade in anal biopsies, epithelial dysplasia grade and location in colonoscopic surveillance biopsies, and adenocarcinoma grade and amount in prostate core biopsies. Significantly, the automated method was 24–39 times faster and inherently contained links for each diagnosis to additional variables such as patient age, location, etc., which would require additional manual processing time.

Conclusions: A simple, flexible, and scaleable NLP-based platform can be used to correctly, safely, and quickly extract and deliver linked data from pathology reports into searchable spreadsheets for clinical and research purposes.

Introduction

Cancer, the second leading cause of mortality in the United States, is diagnosed and characterized within surgical pathology reports, which provide prognostic information and form the basis for patient treatment.[1,2] Recent recommendations and guidelines advocate for the use of synoptic or structured reporting, which improves completeness and communication of clinically relevant pathology data.[3,4] However, narrative reports, comprised of free-form or semi-structured text formats, remain universally employed in most pathology practice settings.[5] Both structured and unstructured formats present obstacles to large-scale data analysis for the purposes of clinical research, including applications in quality assessment and control, given that most data must be manually extracted, subjecting its collection to limitations of personnel resources, time, and human error.[6,7] Parsing elements of the free text-based narrative report into discrete, searchable data points allows for the compilation and analysis of pathologic characteristics of interest.[8] Thus, there is great value in and need for simple, secure, and scalable methods to achieve this data extraction in a timely and efficient manner.

Natural language processing (NLP) is computational processing of human-generated free text with conversion into more uniform formats that facilitate investigation and analysis.[9,10] Several NLP tools exist, from modifiable, open-source programming language packages to commercial platforms.[11,12] However, most of these tools are built around general-use language; the few that are medicine-specific only incorporate terms in specific medical coding systems, such as the International Classification of Diseases for Oncology or have limited pathology-related dictionaries.[13–15] The complexity of these NLP-based applications varies from simple rule-based extraction (applying pre-defined logic operations to annotated text in order to identify keywords and word relationships) to machine learning

* Corresponding author at: Department of Pathology, Molecular and Cell Based Medicine, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1194, New York, NY 10029, USA.

E-mail address: alexandros.polydorides@mountsinai.org (A.D. Polydorides).

http://dx.doi.org/10.1016/j.jpi.2022.100154
Received 29 August 2022; Received in revised form 9 October 2022; Accepted 2 November 2022
Available online 8 November 2022

algorithms (applying different sets of data to statistical models for fine tuning of model parameters).[15–17] Pathology-specific NLP systems have been so far developed and tailored for narrow sets of cases, extracting select information from specific sources, such as semi-structured melanoma or prostate biopsy reports.[7–10] Additionally, many of these systems require additional components and coding language to interface with laboratory information systems (LIS) in order to obtain information, or rely on raw data provided by non-pathology groups such as national registries or clinician-curated data warehouses.[6,11,15,18] Rule-based methods have been limited in obtaining more granular information such as tumor characteristics or only report results for a single specimen type or within a single organ system.[7,8,11,12,18,19] More complex machine learning algorithms require tedious manual annotation by expert reviewers as part of a training dataset, and require tuning of statistical models to identify different parts of the report.[6,16,20] Similarly to rule-based methods, the output from these machine learning algorithms may be in the form of annotated mark-up files which may not be immediately useful for pathology research.[6,17,21–23] In addition, availability of resources such as computing power and specialty-trained personnel may limit their feasibility and usefulness.[6,17,19,23,24]

We introduce herein a simple, dictionary- and rule-based NLP pipeline for extracting diagnosis information and data on various parameters from a diverse array of surgical pathology report types and present data on the efficiency and accuracy of this platform in a number of relevant scenarios in clinical practice and translational research.

Materials and methods

The study was approved by our Institutional Review Board (#18-00479). Basic searches for specimen type, date accessioned, and diagnoses were implemented on our Department's LIS (Powerpath; Sunquest Information Systems, Tucson, AZ) using a built-in, Boolean word search function and results were exported in extensible markup language (XML) format documents. XML is semi-structured and text is labeled with tags corresponding to pre-defined or automatically generated sub-sections (i.e., case numbers, specimen labels, final diagnosis, etc.). Test cases included anal biopsies, endoscopic biopsies with dysplasia from IBD patients, and prostatic core biopsies with adenocarcinoma, collected through a simple word search for specimens collected between 2018 and 2020.

The automated method for collecting data on variables of interest consisted of the NLP-based programming algorithm (described in more detail below), which delivered output into an Excel spreadsheet (Microsoft Corporation, Redmond, WA). This was compared to a manual method for extracting the same information, which consisted of having pathology trainees physically go through each case one at a time and record the required information directly from the LIS into a reporting Excel dataset. The 2 datasets (derived from manual and automated methods) were compared with measured outcomes including time to task completion and frequency of concordant and discordant cases for each variable assessed. Discordant cases were independently and manually verified by the first author for final determination of the correct value. Overall concordance was measured as percent raw agreement and also evaluated with a calculated Cohen's kappa coefficient and corresponding P value, with <.05 considered significant. All analysis was performed using Statistical Package for the Social Sciences software (SPSS; Build 1.0.0.1327; copyright 2019; IBM, Armonk, NY, USA).

Results

The programming algorithm established and employed in this study is described in detail in the Supplementary Text file and available at https://github.com/hansenlam/public_PathReporter. Briefly, an object-oriented Python program accessed XML files from the LIS as input and produced an Excel-compatible file as output with each data row corresponding to a diagnostic summary of each case-specimen. XML files were parsed into collections of case numbers, patient information, specimen labels, part labels, and diagnosis text and a simple loop algorithm scanned and saved connections among them (Fig. 1). Meta-data such as position in the diagnosis and matching words or phrases in the pathology term dictionary, allowed for a term map to be created with machine-readable and searchable text (Fig. 2). Collecting data on variables of interest was accomplished by using functions tailored for the specific topic at hand and stacking the resulting Python lists with extracted features in an Excel table (Fig. 3).

This NLP-based algorithm was tested on 3 different scenarios involving clinicopathological research questions by comparing its resulting output to that obtained by a manual method, the latter consisting of tasking Pathology residents with physically searching the LIS for the data required. The first scenario concerned the identification of dysplasia and, if present, the determination of its grade among 72 anal biopsy specimens (Table 1). There was significant concordance (90.3%; $P < .001$) and excellent inter-observer reliability (Cohen's κ: 0.86) in the results obtained between the manual and automated (i.e., NLP) methods. There were discordant designations in 7 cases (9.7%), which, upon further review, were found to have been correctly assigned by the NLP algorithm in 6 of 7 (85.7%) instances. One case of low-grade dysplasia was incorrectly labelled as negative by the automated method. The manual method required a total of 2.27 person*hours, i.e., an average of approximately 2 min per case for data recording, whereas the automated method took 6 min to input all parameters and run the program, with an average of 5 s per case, or approximately 24-times faster. Importantly, the NLP-derived data table already included a link of each diagnosis to other variables of interest for each specimen (e.g., patient age, location/site of biopsy, etc.) which would take significant additional time and effort in order to be retrieved by the manual method.

Next, we evaluated the NLP-based algorithm in its ability to correctly identify dysplasia grade and location among 306 colorectal biopsies obtained during endoscopic surveillance of patients with inflammatory bowel disease (Table 2). In terms of the highest grade and location of dysplasia, there was again significant concordance (99.0% and 97.1%, respectively; both $P < .001$) and excellent inter-observer reliability (Cohen's κ: 0.97 and 0.96, respectively) when comparing the results of the manual and automated (NLP) methods. There were discordant designations in terms of dysplasia grade in 3 cases (1.0%): 2 incorrectly assigned by the manual method (a case of low-grade dysplasia was called negative and a negative case was called indefinite) and 1 by the automated method: a case of high-grade dysplasia/intramucosal adenocarcinoma where the algorithm placed diagnostic priority on the term "adenocarcinoma" over high-grade dysplasia. In terms of the anatomic location of dysplastic lesions in the colon, there were 9 discordant cases (2.9%), 4 erroneously assigned by the manual method and 5 by the automated algorithm. All 5 mistakes by the automated method consisted of designating the location as "other" and would have been correctly assigned when explored further. The manual method required 6.5 person*hours (an average of approximately 1.3 min per biopsy) for data recording. In contrast, the automated method needed only 10 min for the entire process (approximately 2 s per case), being 39-times faster. Furthermore, by virtue of its design, the automated method already included in its results output a connection between location and dysplasia grade which would have necessitated additional time investment from the manual method to ascertain.

Finally, the automated algorithm was tested on extracting diagnostic information from 300 prostate core biopsies, in terms of the presence of adenocarcinoma and its primary, secondary, and total Gleason grade/score as well as the percentage of tissue with carcinoma, percentage with Gleason grade 4 and/or 5, and amount of tissue with carcinoma in mm (Table 3). There was outstanding concordance (99.1% or 100% for all parameters; $P < .001$) and inter-observer reliability (Cohen's κ: 0.99 or 1.0, across the board) between the 2 methods. One case had its recorded primary and secondary grade switched with the manual method, resulting in the 1 discrepancy in each variable. Another case had had its % of tissue with Gleason 4/5 erroneously designated by the manual method again. The automated method was quicker in this scenario as well, needing only 10 min (2 s per case) or 30-times faster than the manual method, which required 5 person*hours (average of 1 min per biopsy).

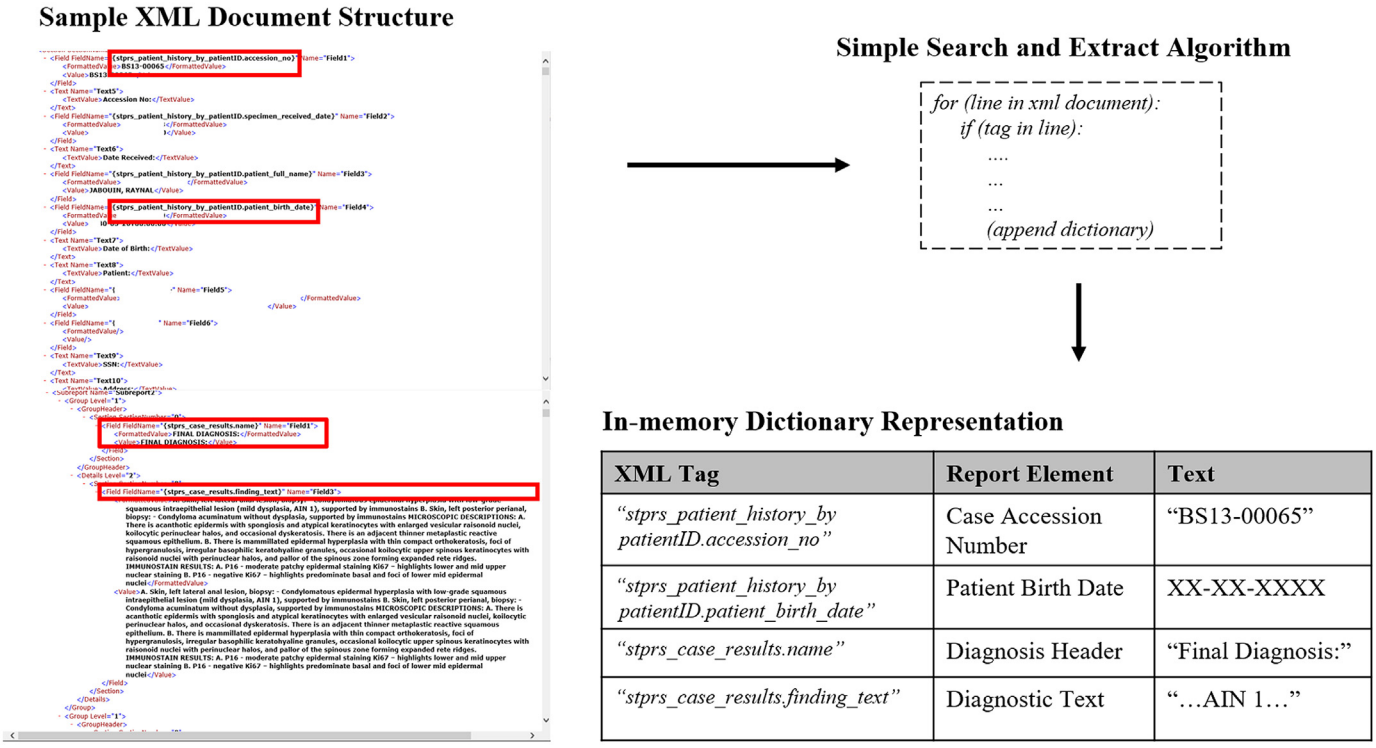

Fig. 1. Schematic representation of the process by which the Python program employed a search and extract algorithm to scan the XML file, obtained directly from the LIS, for specific data points ("XML tags") as instructed ("*for… if…*") and subsequently extract the text corresponding to the desired report elements.

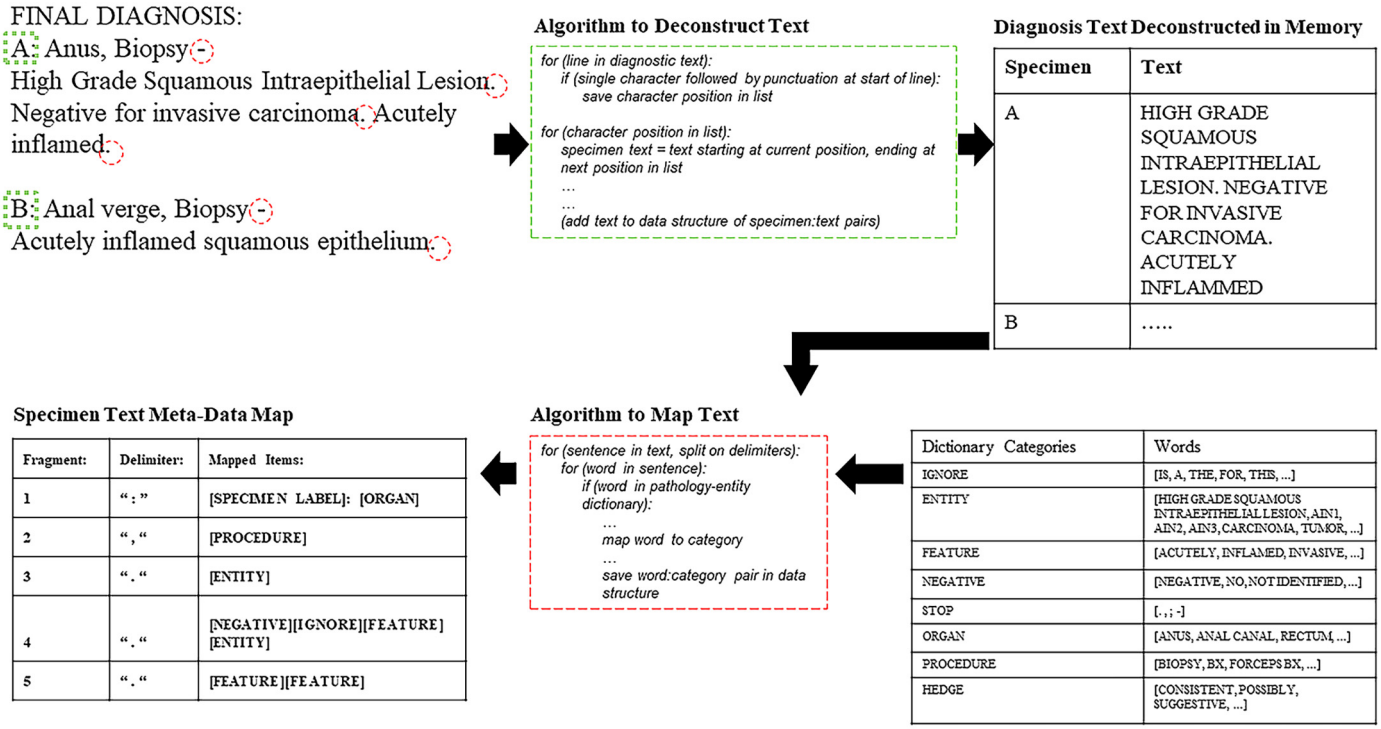

Fig. 2. Diagram depicting the iterative flowchart outlining how the program deconstructed pathologic diagnostic text into discrete ("parsed") dictionary terms and subsequently mapped these terms into meta-data to produce specific machine-readable and searchable text.

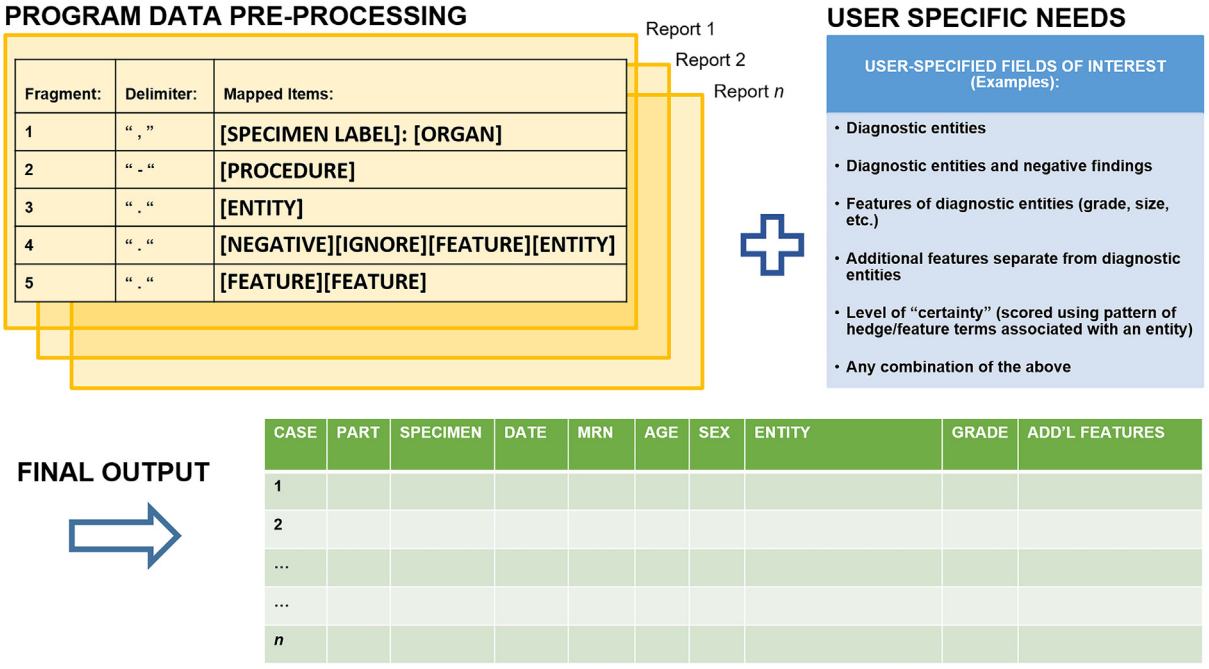

Fig. 3. Final data collection consisted of layering information obtained from all pathology reports, tailored for different pathologic search parameters ("user specific needs") and outputting the resulting Python stacks into organized spreadsheets.

Discussion

This study presents an accessible, efficient, and accurate NLP-based platform to extract diagnostic data from pathology LIS reports in a way that is amenable and useful for both clinical applications and research projects. This flexible homebrew tool, developed using an open-source programming language library, is independent of commercial platforms. It obtains direct LIS input by leveraging XML file structures to parse information into standard elements using object-oriented functional programming and subsequently applies rule-based methods to accurately extract data-rich features of interest across multiple different pathology report types, resulting in readily accessible and searchable spreadsheets. Finally, it is secure, scalable without complex coding or computational requirements, and significantly faster compared to a manual method of data extraction in a number of different scenarios related to clinical practice and translational research.

This method allows for an efficient workflow with relatively simple coding requirements to parse pathology report diagnostic text into discrete sets of case information, before the application of increasingly complex algorithms, as needed, for more granular feature extraction. Additionally, our design directly accesses LIS data without the need for additional interfaces for raw data extraction, and allows pathologists without programming experience to obtain and feed raw data into the program. Our pipeline does not require manual annotation of reports to classify subsections and requires minimal annotation of diagnostic text in the form of the pathology term dictionary. Rule-based approaches work well for pathology reports, providing an alternative to developing complete machine-learning methods. However, more complex algorithms can easily interface with this NLP-based method for further studies due to its modular design nature.

The use of Python objects and data flow was designed to leverage the XML file tags to prioritize parsing the diagnosis text correctly into pairs of specimen letters, and specimen diagnosis text, which was associated in memory with the corresponding case numbers and other accessioning information (patient demographics and specimen labels). The raw diagnosis text for each specimen could be easily accessed through the XMLDocument Object, and in our case, used in rule-based NLP algorithms. Different functions tailored for different report types and styles can be written independently of the main parsing algorithm and easily applied to the diagnostic text of interest, allowing for code organization and versatility. Additionally, other methods such as machine learning pipelines could be developed

Table 1
Comparison of manual and automated data extraction methods for dysplasia grade in surgical pathology reports of anal biopsies.

	Data extraction method		Statistics
	Manual	Automated (NLP)	
Specimen diagnosis			
Negative for dysplasia	31 (43.1%)	27 (37.5%)	Concordance: 90.3% Cohen's κ: 0.86 *P* value: <.001
Low-grade dysplasia (LSIL)	20 (27.8%)	24 (33.3%)	
High-grade dysplasia (HSIL)	11 (15.3%)	10 (13.9%)	
Squamous cell carcinoma	10 (13.9%)	11 (15.3%)	
Time invested			
Person*hours	2.27	0.1	

Table 2
Comparison of manual and automated data extraction methods for grade and location of dysplasia in colorectal surveillance biopsies among patients with inflammatory bowel disease.

	Data extraction method		Statistics
	Manual	Automated	
Dysplasia grade			
Negative for dysplasia	249 (81.4%)	249 (81.4%)	
Indefinite for dysplasia	0 (0.3%)	1 (0.3%)	Concordance: 99.0%
Low-grade dysplasia	50 (16.3%)	49 (16.0%)	Cohen's κ: 0.97
High-grade dysplasia	5 (1.6%)	4 (1.3%)	*P* value: <.001
Adenocarcinoma	2 (0.7%)	3 (1.0%)	
Dysplasia location			
Rectum/sigmoid	79 (25.8%)	75 (24.5%)	Concordance: 97.1% Cohen's κ: 0.96 *P* value: <.001
Descending colon/SF	67 (21.9%)	64 (20.9%)	
Transverse colon	35 (11.4%)	34 (11.1%)	
Ascending colon/HF	63 (20.6%)	64 (20.9%)	
Cecum/Ileocecal valve	26 (8.5%)	27 (8.8%)	
Other	36 (11.8%)	42 (13.7%)	
Time invested			
Person*hours	6.5	0.17	

Table 3
Comparison of manual and automated data extraction methods for evaluating histopathological features of adenocarcinoma in prostate core biopsies.

	Data extraction method		Statistics
	Manual	Automated	
Adenocarcinoma			
Present (positive)	108 (36.0%)	108 (36.0%)	Concordance: 100%
Absent (negative)	192 (64.0%)	192 (64.0%)	Cohen's κ: 1.0
			P value: <.001
Primary Gleason Grade			
Not indicated	1 (0.9%)	1 (0.9%)	
3	76 (70.4%)	75 (69.4%)	Concordance: 99.1%
4	27 (25.0%)	28 (25.9%)	Cohen's κ: 0.98
5	4 (3.7%)	4 (3.7%)	*P* value: <.001
Secondary Gleason Grade			
Not indicated	1 (0.9%)	1 (0.9%)	
3	43 (39.8%)	44 (40.7%)	Concordance: 99.1%
4	48 (44.4%)	47 (43.5%)	Cohen's κ: 0.99
5	16 (14.8%)	16 (14.8%)	*P* value: <.001
Total Gleason Score			
Not indicated	1 (0.9%)	1 (0.9%)	
6	32 (29.6%)	32 (29.6%)	Concordance: 100%
7	53 (49.1%)	53 (49.1%)	Cohen's κ: 1.0
8	4 (3.7%)	4 (3.7%)	*P* value: <.001
9	18 (16.7%)	18 (16.7%)	
Tissue with carcinoma (%)			
Not indicated	1 (0.9%)	1 (0.9%)	
0%–25%	44 (40.7%)	45 (41.7%)	Concordance: 99.1%
26%–50%	24 (22.2%)	23 (21.3%)	Cohen's κ: 0.99
51%–75%	17 (15.7%)	17 (15.7%)	*P* value: <.001
76%–100%	22 (20.4%)	22 (20.4%)	
Amount of Gleason 4/5 (%)			
Not indicated	76 (70.4%)	76 (70.4%)	
0%–20%	18 (16.7%)	18 (16.7%)	Concordance: 100%
21%–40%	9 (8.3%)	9 (8.3%)	Cohen's κ: 1.0
41%–60%	5 (4.6%)	5 (4.6%)	*P* value: <.001
Amount of carcinoma (mm)			
Not indicated	50 (46.3%)	50 (46.3%)	
0–5 mm	25 (23.1%)	25 (23.1%)	Concordance: 100%
6–10 mm	13 (12.0%)	13 (12.0%)	Cohen's κ: 1.0
11–20 mm	20 (18.5%)	20 (18.5%)	*P* value: <.001
Time invested			
Person*hours	5	0.17	

separately and easily applied to diagnosis text, while maintaining association with correct case numbers, patient information, and specimen type, if so desired by users. Finally, the actual report text of interest associated with each tag was saved in variables within the XMLDocument Object and can be easily accessible and retrievable.

Reports reviewed from each biopsy type were signed out by multiple pathologists across different lab locations within our institution's health system, reflecting the variety of reporting styles and formatting present across pathology reports. The program and its generated data was restricted to the local workstation on which it was run and where the XML files were saved. There was no web or internet dependencies required and no network connections made while the program ran. The automated extraction results were exported as a discrete file type that can be safely stored and handled like other electronic documents containing protected health information (PHI). Finally, extracted data were obtained from pathology reports in cases (biopsies) that do not usually contain synoptic reports, therefore allowing parsed data in situations where data extraction might otherwise be more difficult and laborious.

Specifically, in anal biopsy reports, we show that recording and summarizing the grade of dyplasia while accounting for the different ways of reporting anal intraepithelial neoplasia can be achieved by a rule-based NLP pipeline. In the case of IBD-related biopsies, we show that the normally tedious extraction of site, presence, and grade of dysplasia among biopsies performed for surveillance in IBD patients can be done at scale accurately using the same pipeline. Finally, in the case of prostate biopsies, we show that the characteristics of prostatic adenocarcinoma can be extracted in similar fashion with a variation on the same pipeline. We show that this tool can be robustly applied to biopsy reports while requiring only a low to moderate level of technical expertise, and still preserves the ability to easily incorporate more complex algorithms such as machine learning through the modular nature of the design. This represents a flexible alternative to tools with highly technical requirements, and can be used without extensive, large-scale computational resources. However, a potential limitation to this method is its reliance on the XML document structure that our LIS produces for each report. The initial parsing of each report into patient information, and diagnostic text utilizes XML document tags to correctly identify each section. Thus, the methods described here are not generalizable to institutions that do not use XML formatting to store reports in their LIS.

There are several approaches to applying NLP to pathology reports, coinciding with the varying levels of complexity inherent to reporting different specimen types (Table 4). Choosing which approach to use when designing an NLP tool requires careful consideration of desired output, resources, and maintenance requirements. For instance, an NLP tool using a simple search for specific keyword requires less technical expertise, less complex verification, and less maintenance, however, may not capture the granularity of information obtained using more complex algorithms. Such designs may work for well for extracting items such as special stains, or in situations where reporting has been made highly uniform across an organization. On the other hand, for larger report items, tools designed using machine learning algorithms may achieve the most accurate results, at the cost of requiring continuous training using carefully selected datasets. These algorithms typically require manual annotation by expert reviewers to label entire documents with part-of-speech tags, pathology-specific tags, and context tags. The labeled document must then be passed through a data pipeline that applies any combination of machine learning

Table 4
Summary of advantages and disadvantages with different methods of data extraction.

Method	Advantages	Disadvantages
Simple search: Create a list of entity and feature names on which a program can cross reference as it scans each line of text and extract any matches	• Relatively easy to code • Does not require high--performance computing hardware	• Relationships between extracted words are lost (when multiple entities/features extracted, cannot determine which features associated with which entities) • **Difficulty dealing with negative statements (Not identified, negative, etc.)**
Rule-based term map: Create dictionary that a program can use, along with punctuation, to associate text with a map of standard terms (Entity, Feature, Negative, etc.), and extract terms of interest and/or ***term patterns of interest***	• Basic relationship of terms maintained • Does not involve choosing between different algorithms to determine most accurate way to extract meaning	• Moderate level of coding experience required • **Dictionary must be built from the ground up**
Statistical methods (machine learning): Use readily available code libraries that assign speech parts (noun, verb, etc.) to text; statistical methods identify the most likely subject of each sentence	• Relationship of every term is defined relative to other terms • Extracted meaning can be inferred and evaluated numerically	• Requires high level of coding and statistical knowledge • Requires high--performance computer hardware • **Available libraries are not designed with medical vocabulary**

models.[7,8,12,13] Such algorithms require careful consideration of training datasets and test datasets, and optimization of model parameters to achieve high accuracies.[14–16] We chose a balanced approach, employing rule-based and object relational model methods, that partitions the reports into structures which can be utilized by more complex algorithms, should the need arise.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2022.100154.

References

1. Nakhleh RE. Quality in surgical pathology communication and reporting. Arch Pathol Lab Med Nov 2011;135(11):1394–1397. https://doi.org/10.5858/arpa.2011-0192-RA.
2. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2006. CA Cancer J Clin Mar-Apr 2006;56(2):106–130. https://doi.org/10.3322/canjclin.56.2.106.
3. Sluijter CE, van Lonkhuijzen LR, van Slooten HJ, Nagtegaal ID, Overbeek LI. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. Virchows Arch Jun 2016;468(6):639–649. https://doi.org/10.1007/s00428-016-1935-8.
4. Ellis DW, Srigley J. Does standardised structured reporting contribute to quality in diagnostic pathology? The importance of evidence-based datasets. Virchows Arch Jan 2016;468(1):51–59. https://doi.org/10.1007/s00428-015-1834-4.
5. Srigley JR, McGowan T, Maclean A, et al. Standardized synoptic cancer pathology reporting: a population-based approach. J Surg Oncol Jun 15, 2009;99(8):517–524. https://doi.org/10.1002/jso.21282.
6. Kim Y, Lee JH, Choi S, et al. Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. Sci Rep Nov 20, 2020;10(1):20265. https://doi.org/10.1038/s41598-020-77258-w.
7. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J Am Med Inform Assoc May-Jun 2010;17(3):253–264. https://doi.org/10.1136/jamia.2009.002295.
8. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. JCO Clin Cancer Inform Dec 2018;2:1–8. https://doi.org/10.1200/CCI.17.00128.
9. Buckley JM, Coopey SB, Sharko J, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 2012;3:23. https://doi.org/10.4103/2153-3539.97788.
10. Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Stud Health Technol Inform 2004;107(Pt 1):565–572.
11. Malke JC, Jin S, Camp SP, et al. Enhancing case capture, quality, and completeness of primary melanoma pathology records via natural language processing. JCO Clin Cancer Inform Aug 2019;3:1-11. https://doi.org/10.1200/CCI.19.00006.
12. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. J Am Med Inform Assoc Mar-Apr 2013;20(2):349–355. https://doi.org/10.1136/amiajnl-2012-000928.
13. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics Feb 15, 2020;36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682.
14. Zhu Q, Li X, Conesa A, Pereira C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. Bioinformatics May 1, 2018;34(9):1547–1554. https://doi.org/10.1093/bioinformatics/btx815.
15. Hammami L, Paglialonga A, Pruneri G, et al. Automated classification of cancer morphology from Italian pathology reports using Natural Language Processing techniques: a rule-based approach. J Biomed Inform Apr 2021;116, 103712. https://doi.org/10.1016/j.jbi.2021.103712.
16. Oliwa T, Maron SB, Chase LM, et al. Obtaining knowledge in pathology reports through a natural language processing approach with classification, named-entity recognition, and relation-extraction heuristics. JCO Clin Cancer Inform Aug 2019;3:1–8. https://doi.org/10.1200/CCI.19.00008.
17. Altieri N, Park B, Olson M, DeNero J, Odisho AY, Yu B. Supervised line attention for tumor attribute classification from pathology reports: Higher performance with less data. J Biomed Inform Oct 2021;122, 103872. https://doi.org/10.1016/j.jbi.2021.103872.
18. Odisho AY, Bridge M, Webb M, et al. Automating the capture of structured pathology data for prostate cancer clinical care and research. JCO Clin Cancer Inform Jul 2019;3:1–8. https://doi.org/10.1200/CCI.18.00084.
19. Oliveira CR, Niccolai P, Ortiz AM, et al. Natural language processing for surveillance of cervical and anal cancer and precancer: algorithm development and split-validation study. JMIR Med Inform Nov 3, 2020;8(11), e20826. https://doi.org/10.2196/20826.
20. Giannaris PS, Al-Taie Z, Kovalenko M, et al. Artificial intelligence-driven structurization of diagnostic information in free-text pathology reports. J Pathol Inform 2020;11:4. https://doi.org/10.4103/jpi.jpi_30_19.
21. Lee J, Song HJ, Yoon E, et al. Automated extraction of biomarker information from pathology reports. BMC Med Inform Decis Mak May 21, 2018;18(1):29. https://doi.org/10.1186/s12911-018-0609-7.
22. Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc Sep-Oct 2014;21(5):824–832. https://doi.org/10.1136/amiajnl-2013-002443.
23. Odisho AY, Park B, Altieri N, et al. Natural language processing systems for pathology parsing in limited data environments with uncertainty estimation. JAMIA OPen Oct 2020;3(3):431–438. https://doi.org/10.1093/jamiaopen/ooaa029.
24. Alawad M, Gao S, Qiu JX, et al. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. J Am Med Inform Assoc Jan 1, 2020;27(1):89–98. https://doi.org/10.1093/jamia/ocz153.