On the Diagnosis and Characterisation of Prostate Cancer in Pathology Reports in Spanish

Rosa M. Montañés-Salas[1], Sergio Gracia-Borobia[1], María de la Vega Rodrigálvarez-Chamarro[1], Ángel Borque-Fernando[2], Patricia A. Guerrero-Ochoa[3], Alejandro Camón-Fernández[3], Jorge Alfaro-Torres[4], Isabel Marquina-Ibáñez[4], Sofia Hakim-Alonso[4], Luis M. Esteban[5] and Rafael del-Hoyo-Alonso[1]

[1]Aragon Institute of Technology (ITA), María de Luna, 7–8, 50018 Zaragoza, Spain
[2]Department of Urology, Miguel Servet University Hospital (GIIS071-uro-servet), 50009 Zaragoza, Spain
[3]Health Research Institute of Aragon Foundation (GIIS071-uro-servet), 50009 Zaragoza, Spain
[4]Department of Pathology, Miguel Servet University Hospital (GIIS071-uro-servet), 50009 Zaragoza, Spain
[5]Department of Applied Mathematics, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, 50100 Zaragoza, Spain

Abstract

Prostate cancer is a prevalent disease worldwide, with early diagnosis enabling better prognosis. Natural language processing (NLP) techniques show promise in extracting information from electronic health records to support clinical decision-making. This paper presents an NLP approach to detect and characterise prostate cancer (PCa) diagnoses from Spanish pathology reports. A combination of lexical-morphological analysis, rule-based techniques and transformer models is used to identify PCa, Gleason scores, procedures, organs and other markers. The system achieves near 96% agreement in detecting cancer diagnoses compared to expert annotation.

Keywords

Medical Natural Language Processing, Prostate Cancer, Pathology Reports, Information Extraction

1. Introduction

Prostate cancer (PCa) was the fourth most diagnosed cancer worldwide in 2022. In Spain, in 2023, there have been reported between 33,000-34,000 new cases diagnosed, with a 5-year prevalence of more than 140,000 cases, making it the leading cancer in terms of incidence among the male population, as reported by the Spanish Cancer Association in 2023 [1]. Early detection of PCa enables treatment at initial stage, resulting in higher cure rates and reduced side effects of aggressive treatments, as well as lower healthcare costs. In this context, the use of advanced Artificial Intelligence (AI) techniques presents itself as a promising tool to support clinical decision-making systems and predictive model research [1]. Specifically, Natural Language Processing can play a decisive role by facilitating information retrieval from non-structured sources and opening the range of processing techniques [2].

The acquisition of large volumes of patient data for research purposes has become feasible today, thanks to the digitization of healthcare systems and systematic recording of medical procedures. However, there are still limitations stemming from the non-uniformity of information systems, the diverse repositories for clinical analyses, radiological or pathological reports, or other Electronic Health Records (EHRs), and the assessment and follow-up of patients conducted by different healthcare professionals. Therefore, the sources and data are significantly heterogeneous, including a substantial amount of textual information containing valuable clinical knowledge provided by experts in the field, which allows for precise and accurate diagnoses. Moreover, privacy concerns have to be taken into account when dealing with patient sensitive data [3].

The research presented here is part of the AI4HealthyAging project, whose mission is to leverage distributed AI technologies for early diagnosis and treatment of diseases that are highly prevalent in the ageing population. Within this project, several work packages are organized regarding various diseases such as Parkinson, sarcopenia, deafness or cancer. The use case related to the diagnostic management of prevalent cancers in the elderly is focused on prostate and colon cancers. Particularly, the overarching goal for prostate cancer is to develop decision support and risk interpretation tools based on the biological footprint present in patient EHRs, histological preparations, and radiological images, thus enhancing diagnosis through the use of hybrid data.

In this article, we present our experience in analysing, extracting and structuring information using Natural Language Processing techniques applied to pathology reports of prostate cancer in Spanish. The main challenge presented is the unavailability of a truly reliable and medically consistent labeled dataset in Spanish from which to incorporate new clinical features for the development of advanced hybrid predictive models. Therefore, the first stages of this project consisted of the development of a working methodology to retrieve relevant data and implement an NLP-based system that would enable to efficiently detect and characterise cancer diagnoses based on pathologists' reports. The proposed approach has facilitated the compilation of a comprehensive

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.
rmontanes@ita.es (R. M. Montañés-Salas); sgracia@ita.es (S. Gracia-Borobia); vrodrigalvarez@ita.es (M. d. l. V. Rodrigálvarez-Chamarro); aborque@salud.aragon.es (Á. Borque-Fernando); pguerrero@iisaragon.es (P. A. Guerrero-Ochoa); acamon@iisaragon.es (A. Camón-Fernández); jalfaro@salud.aragon.es (J. Alfaro-Torres); imarquina@salud.aragon.es (I. Marquina-Ibáñez); shakim@salud.aragon.es (S. Hakim-Alonso); lmeste@unizar.es (L. M. Esteban); rdelhoyo@ita.es (R. del-Hoyo-Alonso)
0000-0003-4636-5868 (R. M. Montañés-Salas); 0009-0005-4863-8550 (S. Gracia-Borobia); 0000-0003-1393-8260 (M. d. l. V. Rodrigálvarez-Chamarro); 0000-0003-0178-4567 (Á. Borque-Fernando); 0000-0002-1657-4792 (P. A. Guerrero-Ochoa); 0000-0002-3007-302X (L. M. Esteban); 0000-0003-2755-5500 (R. del-Hoyo-Alonso)

[1]Observatory of the Spanish Cancer Association (AECC) (2023). Dynamic report: Prostate cancer | AECC Observatory. Available at https://observatorio.contraelcancer.es/informes/informe-dinamico-cancer-de-prostata. Last accessed: 2024-04-03

and validated dataset, providing both, a final/clear diagnosis and several features of interest, for the development of explainable risk prediction and prognosis machine learning models fueled by multiple information sources.

The paper is organised as follows: a contextual introduction that delineates the research setting has been outlined, followed by a related work exploration. Section 3 encompasses the materials and methods integral to underpinning the system's development. Subsequently, the attained results are discussed upon finishing with the conclusions of the work and prospective areas for future development and research.

2. Related work

Natural Language Processing applied to the biomedical domain has witnessed significant growth and innovation in recent years, driven by the increasing availability of large-scale healthcare data and advancements in NLP techniques. Electronic health records have emerged as a significant source of information for the detection and diagnosis of several diseases, including cancer. With the incorporation of rich textual data, natural language processing has been extensively applied to these records. [4] and [5] systematically reviewed the applications of NLP in sifting through EHRs, highlighting its potential in detecting chronic diseases and signs of various cancer types, respectively, emphasizing the challenges of data heterogeneity and the significance of domain-specific annotations. This is also shown in [6] and [7], where information was extracted from free-text pathology reports related to breast and lung cancer and colorectal cancer, respectively, both using expert annotated textual data.

In the realm of prostate cancer, a niche yet rapidly growing area of research focuses on leveraging NLP techniques for diagnostic purposes. DiBello et al. demonstrated that NLP can accurately identify metastatic PCa by searching unstructured text in medical records such as pathology, radiology and clinic notes. Thomas et al. validated an NLP program to accurately identify patients with prostate cancer and extract relevant information from pathology reports. Some approaches also underscore the critical role of domain-specific knowledge in curating and understanding the specialized terminology present in such reports [10]. Additionally, the extraction of such domain-specific knowledge helps improve unimodal models within PCa diagnosis: Morote et al. assessed the ability of microscopic findings in prostate biopsies to improve the prediction of clinically significant prostate cancer using numerical-only data, and Khosravi et al. developed an AI based model for PCa diagnosis using magnetic resonance images labelled with manually-assigned histopathology information.

A great variety of text-related machine learning models have seen application in this domain: Breischneider et al. developed an unsupervised rule-based ontology system for feature extraction in free-form text obtained from clinical reports. Yoon et al. dived deeper and demonstrates how graph neural networks can be trained with in-context textual data for multitask cancer labeling. Also, transformer-based models, renowned for their prowess in NLP tasks, have been introduced into the biomedical field. ClinicalBioBERT [15] and OncoBERT [16] showcased the utility of BERT and its variants in comprehending medical narratives, aiming at identifying signs of cancers. Their findings illuminate the potential of fine-tuning such models with domain-specific data to boost diagnostic performance.

Despite these advancements, challenges persist in achieving optimal performance for the detection and diagnosis of diseases, particularly due to the inherent noise and variability in EHRs and other medical reports. Moreover, the multilingualism challenge is still an open issue, although multiple efforts are being conducted to develop annotation standards and AI systems in Spanish (Seda et al.; Miranda-Escalada et al.; Solarte-Pabón et al.), the scope of research in the oncology field is still limited.

3. Proposed approach

This section outlines the proposed approach for the development of the clinical report analysis system. The system aims to extract relevant information from medical documents and classify them according to their diagnosis or the requirements proposed. It poses a working methodology customizable to different types of medical text reports, in which, starting from basic resources in the form of a textual corpus and elementary terminology, different natural language processing strategies are applied semi-automatically to characterise the data set and obtain implicit information useful for feeding other learning systems.

The materials and methods presented here have been developed by three independent teams in order to ensure data privacy[2] and broad applicability of the system. First, the team responsible for accessing the healthcare databases retrieves two sets of data. Thereafter, the text set is analysed by the natural language processing team and its results are combined with the medical data modelling team's set for validation purposes.

3.1. Materials

The initial study population in this work consists of all males affiliated with Zaragoza II healthcare sector (392,177 individuals), who have been identified with at least one Prostate-Specific Antigen (PSA) determination in the last 5 years, i.e., in the interval between 2017 and 2022. With these characteristics, a total of 92,171 patients were identified. From this initial population, those with at least one biopsy performed, and consequently, with at least one pathology report available in the system, were selected, resulting in a total of 10,563 patients of interest. All available data in the hospital systems accountable for performing these procedures were retrospectively collected since 1999, ultimately retrieving a total of 24981 textual pathology reports.

3.1.1. Data preparation

The team responsible for accessing healthcare databases and retrieving reports work independently of the other teams to ensure the confidentiality and protection of sensitive data. The former team has conducted a triple pseudonymization process upon the pathology reports, compiling a document database for natural language processing with two content sections: the macroscopic description and the diagnosis section. The macroscopic description provides concrete

[2]Following the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (GDPR).

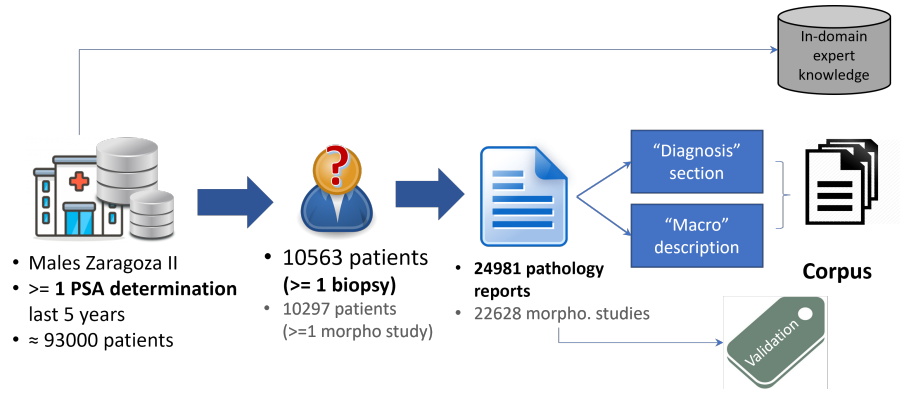

Figure 1: Data sources schema.

details regarding the tissue removal procedure, while in the diagnosis section, pathologists determine the findings in the extracted tissue samples and provide detailed information they deem relevant for patient monitoring. Both, "diagnosis" and "macro" sections are considered for the detection and characterisation of PCa (referred to as "corpus" in figure 1).

Along with this base corpus, a lightweight thesaurus was built from the exploration of PCa-related concepts in standard ontologies and classification schemas such as SNOMED [20] and ICD10-codes [21] in their Spanish versions. This resource has been constructed by the NLP team with the guidance of the medical staff at the Zaragoza II healthcare sector in charge of reporting prostate cancer, with the aim of simplifying the external knowledge integration, implementing an easily adaptable tool and adjusting the linguistic analysis to real-world usage by reproducing the communicative registry employed in those healthcare reports, according to the American College of Pathologist-CAP guidelines actualized every 6 months by the specialist pathologist staff of the Departament of Pathology. This dictionary is designed as a simple hierarchy in which each concept or characteristic to be extracted is paired with a carefully refined list of expressions drawn from the standards mentioned above, comprising the expert domain knowledge base of the system.

3.1.2. Validation set

To assess the results of the analysis system, an independent dataset from morphology study repositories has been retrieved under the same assumptions as the textual data corresponding to the pathology reports, corresponding to 10297 patients with 22628 cases. These studies contain numerical and categorical data, particularly a summary label assigned by clinical staff during patient exams, following the SNOMED nomenclature. A preliminary analysis of these labels revealed a high dimensionality and variability of the assigned classes. Additionally, excessively specific categories were used, making it difficult to study PCa diagnosis adequately. This diversity can be attributed to the complexity of the standard used on a day-to-day basis and the subjectivity of human criteria for assignment.

An overview of the aforementioned resources is depicted in the following figure 1.

3.2. Methods

In the design of the clinical record information extraction and labelling system depicted in figure 2, we aimed to follow an iterative yet simple methodology: based on the set of language resources outlined in the previous section (see 3.1) and the application of different biomedical natural language processing techniques, a comprehensive structured dataset is built and refined in conjunction with the in-domain knowledge. Both, dataset information and expert knowledge are easily customized according to the specific needs of subsequent machine learning models or expert requirements. In our use case, a structured dataset of prostate cancer diagnosis features is built and then validated by an independent team, performing a limited number of feedback iterations to improve the overall performance and guarantee the generalization and customizable capabilities of the system.

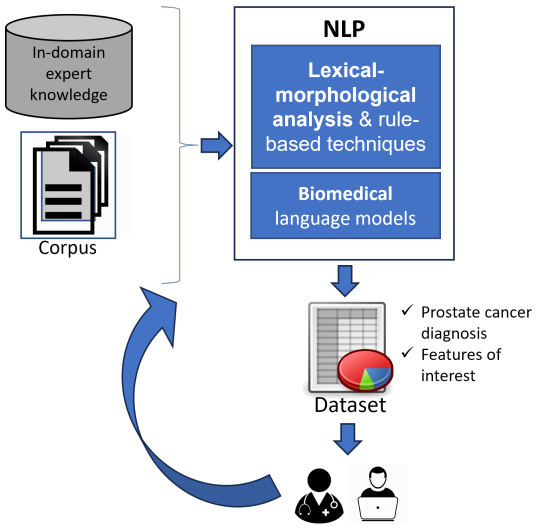

Figure 2: System schema.

3.2.1. Analysis and tagging

The preliminary analysis phase consists of building automatically a glossary of categorized terms and expressions that are morphologically and semantically very close to the base in-domain knowledge provided by the medical staff. Therefore, a lightweight customized thesaurus is constructed,

through similarity searches over the corpus along with a set of approximate regular morphological data patterns usually found on cancer reports. These terms and expressions facilitate the further identification of information patterns and data in the textual records, aiding in the determination of diagnoses and their context.

From this, the available pathology reports are processed and both the prostate cancer diagnosis, in terms of positive or negative presence, and a number of additional features of interest are inferred, building a rich dataset that can feed other AI multimodal systems. To accomplish the detection of diagnoses and the extraction of significant variables employing the compiled lexicon and patterns, various strategies are integrated into a three-step approach:

Firstly, a content-based filtering is applied, discarding empty or non-informative reports (i.e. many documents consist of only "VER B" or similar texts). The second step is based on a mixture of lexical-morphological analysis supported by fuzzy matching and the integration of a pretrained language model based on transformers. In this case, we utilize biomedical language models in the Spanish language, specifically the RoBERTa-base biomedical model that has been already finetuned for the Named Entity Recognition (NER) task on the Cantemist dataset for tumour morphology extraction by Carrino et al.. The joint objective is to extract cancer-related terms and expressions from the text reports, which represent the relevant features pursued. The third step consists of applying rule-based techniques designed for the extraction of objective features. On the one hand, a simple set of grammatical rules has been applied, i.e. occurrence of certain constituents in the sentences, negation detection and comparison of lexical-morphological analysis with NER output. On the other hand, domain-specific rules have been designed: a priority system between labels has been established for the PCa diagnosis distinguishing from positive cases to different gravity levels and not PCa as less critical; specific rules to compute the final Gleason degrees and groups when different values are retrieved; and warning rules to analyse whether values extracted are incoherent.

The detailed characteristics retrieved through the described NLP techniques on the pathology reports of prostate cancer are the following:

- Prostate Cancer diagnosis: a binary output (*PCa+*, *PCa-*, for positive or negative Prostate Cancer diagnosis, respectively) that includes an additional label for uncertain cases in which the analysis and rule processing throw opposing or empty results (identified as *Untagged*). The latter serves to indicate the need for a thorough review by an expert.
- Gleason Score (Sum and Group): the Gleason score, often referred to as the Gleason grading system [23] is a numerical scale used in the field of pathology to assess the severity and aggressiveness of prostate cancer glands under a microscope. It can be expressed in several ways, generally consisting of two numbers: a primary and a secondary cancer pattern, i.e. the most common pattern of cancer glands seen, and the next most common pattern, respectively (tertiary is also extracted, but it is rarely mentioned). The sum of the primary and secondary patterns along with their corresponding group are identified and conveniently computed. Whenever the Gleason score is mentioned multiple times in a document (i.e. in a pathology report describing a biopsy with a list of inspected cilinders) all the possible components are extracted but the gleason score associated to that document is the most severe among the multiple results. The pathology reports analysed in this research show certain variability due to the evolution of the Gleason grading system in the time range considered, which have been redacted following the recommendations and updates of the International Society of Urological Pathology (ISUP).
- Type of Medical Procedure: each of the reports analysed corresponds to a specific procedure conducted on the patient. The considered procedures are: Biopsy, Prostatectomy, Cystoprostatectomy, Adenomectomy and TURP (Transurethral Resection of the Prostate). An additional *Untagged* label is considered in case none of the above could be detected. This result is treated as a multilabel output.
- Organ mentions: given that the different medical procedures can affect several areas of the organism, the mentioned organs on each document are also extracted. The considered organs are: Prostate, Seminal Vesicles, Lymph Nodes and Bladder. This result is treated as a mutilabel output.
- Other informative markers: mentions to ASAP (Atypical Small Acinar Proliferation); mentions to PIN (Prostatic Intraepithelial Neoplasia); mentions of inflammatory processes and atrophies; TNM stage: standard classification for cancer staging, it refers to Tumour, Nodes and Metastasis [24]; "DUC" label related to mentions of ductal carcinoma; and neoplastic morphology mentions extracted with the NER model.

3.2.2. Validation

The modelling team, in collaboration with health professionals, found that the SNOMED labels assigned in the validation set (see section 3.1.2) are not very accurate for the actual diagnosis of PCa. They have independently made an adjustment to this set of labels, reducing it to a three-category classification (existence or non-existence of PCa plus an 'untagged' label) through semi-automated mapping and subsequent human validation by two experts.

The intersection of the text corpus and the morphology data over patient studies consists of 10249 patients and a total of 17696 studies on which it is possible to compare and validate PCa tagging results at document-level. The inter-annotator agreement (IAA) is computed between the NLP tags and the postprocessed human-assigned SNOMED labels.

4. Results

The results reported in this section correspond to the final results obtained after three cycles of compilation, execution and validation of information extraction and classification over the available pathology reports. The base outcome of the system presented is the structured dataset of prostate cancer diagnoses from highly unstructured free-text data. In conjunction with this resource, it has been possible to validate a simple yet effective method for the extraction of relevant information from these types of documents with a very promising overall performance.

Annex A contains a complete example of one of the analysed documents, corresponding to the following textual

fragment from the diagnosis section. The table 1 below includes the most relevant features extracted from the whole document. All the characteristics extracted are specified in the table 4 in the annex.

BIOPSIA DE PRÓSTATA TRANSPERINEAL, [...] - GRADO DE GLEASON: 7 (3+4) - GRADO GRUPO: 2 [...] PATOLOGÍA ADICIONAL PROSTÁTICA: NO SE OBSERVA.

Table 1
Extracted information from example document

Extracted information	Value
cancer tag	Cancer
doc tag	Biopsia
primary gleason pattern	3
secondary gleason pattern	4
gleason sum	7
gleason group	2

The distribution of cancer presence in the population studied through pathology reports is depicted in figure 3. According to the experts, it aligns with the typical average distribution of PCa diagnoses in the specific region under study with the characteristics described in the materials section, remaining about a 2% of the records (370 reports approximately) to be reviewed by doctors. ASAP, PIN, inflammation, and atrophy are all pathologies related to the possible development of prostate cancer. However, for diagnostic purposes, medical experts consider them as negative cases of PCa.

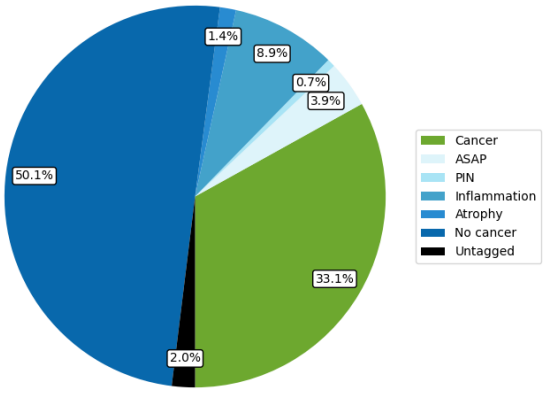

Figure 3: Distribution of cancer diagnosis.

The distributions of Gleason Score sum and group retrieved for positive PCa cases are also reproduced in table 2 and figure 4.

Regarding the compiled corpus of documents, it was found that the inclusion of the macroscopic section aided in extracting features related to procedures. The distribution of procedures found is depicted in figure 5.

Figure 6 displays the organs directly related to prostate cancer mentioned in the textual corpus. These mentions are

Table 2
Gleason sum distribution.

Gleason sum	% records
4	0.3
5	0.5
6	52.7
7	32.8
8	6.9
9	6.0
10	0.8

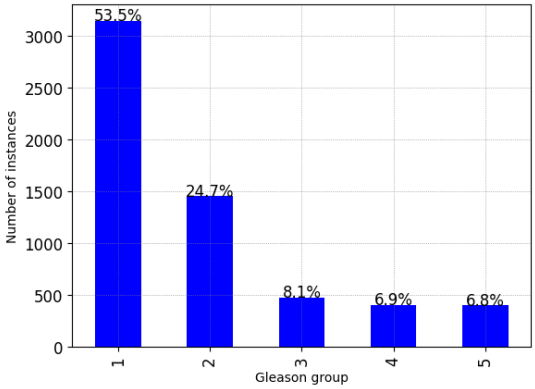

Figure 4: Gleason groups distribution.

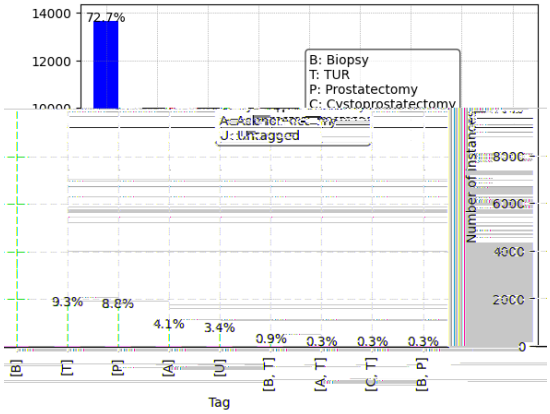

Figure 5: Procedures identified.

valuable for characterising and filtering data in downstream tasks.

Tumour morphology entities extracted with the NER model were not found to be helpful at this level of language analysis, regarding the techniques integrated within the system. Neither prostate cancer nor Gleason scores were improved, as these mentions, when correctly detected, co-occurred with specific prostate domain expressions already matched in the lexical-morphological analysis. The transformer-based model was excluded from the process in order to reduce the system's runtime.

Finally, to ensure the system's correct performance, prostate cancer diagnosis from pathology reports were validated against the validation set, i.e. the manually-annotated morphology dataset, using the predefined tags outlined in

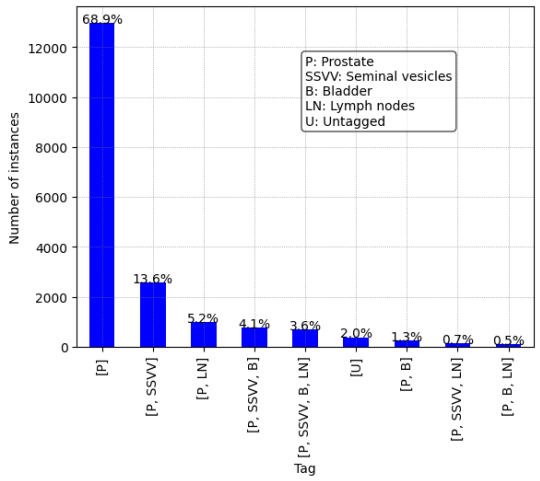

Figure 6: PCa organs identified.

section 3.2: *PCa+, PCa- and Untagged*. Although, initially, the morphology coding was not entirely consistent, after the reviewing cycles it was found to be a useful baseline and provided valuable support for the detection of complex cases. These cases were then personally reviewed by medical experts, thus enhancing our methodology and allowing for a thorough evaluation of the obtained results.

Table 3
Agreement on cancer diagnosis.

NLP/Expert	PCa+	PCa-	Untagged
PCa+	5433	648	1
PCa-	12	11566	11
Untagged	1	18	6
Total agreement			17005

As shown in table 3, with respect to the evaluable pathology reports coinciding in both datasets, a final 96.095% agreement between NLP-based classification and human labelling was reached, which corresponds with a weighted Cohen's Kappa of 0.9115. During the validation cycles, our automatic analysis and processing was found to be more robust than the manual annotation registered in the hospital repositories, and both were iteratively improved. Gleason scores, medical procedures, organ mentions and the rest of the additional markers retrieved also underwent a cursory validation by human expert intervention, as there was no corresponding information conveniently categorised in the clinical repositories or in the previous work studied. The distribution of data obtained corresponds to the actual distribution of cases treated in the time period analysed, as determined by the vast experience of the clinical staff involved.

5. Conclusions and future work

In this paper we have presented our approach to retrieving, classifying and structuring information using a combination of NLP techniques over prostate cancer pathology reports in Spanish from the records belonging to the Health Sector Zaragoza II. The methodology designed, and the system implemented, set the starting point for the development of a global and homogeneous cancer risk prediction system based on the biological footprint of patients distributed in the different electronic health records available in the hospital repositories, supporting and improving the efficiency of decision support tools for early disease detection processes.

We have built an extensive structured dataset that serves to enhance the predictive capacity and explainability of advanced predictive models for PCa risk identification, such as multimodal algorithms that work with biological data and images, and enrich hospital repositories. Additionally, the validation performed along with the clinical experts throws very encouraging results, approximating to 96% of annotator agreement. Nevertheless, a comprehensive and rigorous validation of the remaining characteristics is still required, despite their initial alignment with the expectations and needs of the experts.

The developed system has been successfully adapted and executed into a colorectal cancer scenario within this project, by simply defining an in-domain thesaurus. This has allowed the processing and evaluation of pathology reports and then, extending the approach to colonoscopies and EHR, serving as an effective system to annotate customized datasets for further research.

As future work, several avenues are considered, from the NLP perspective we plan to improve the techniques explored, delving into higher levels of language analysis as the semantics, and exploring the new possibilities offered by the leading-edge Large Language Models (LLMs). Furthermore, for the use case under study, the methodology will be applied to the magnetic resonance reports of prostate cancer patients as well, with the necessary adaptations for that specific context, in order to further enhance the dataset constructed. Lastly, our objective is to extrapolate the analysis and processing to all types of reports with textual content to build a comprehensive system that allows for the identification of a patient's biological footprint and its influence on the final diagnosis.

Ethical Statement

This study and the use of patient data was approved by the regional ethics committee of Aragón.

Acknowledgments

This research was funded by project MIA.2021.M02.0007 of NextGenerationEU program and Integration and Development of Big Data and Electrical Systems (IODIDE) group of Aragon Goverment program.

References

[1] A. A. Rabaan, M. A. Bakhrebah, H. AlSaihati, S. Alhumaid, R. A. Alsubki, S. A. Turkistani, S. Al-Abdulhadi, Y. Aldawood, A. A. Alsaleh, Y. N. Alhashem, J. A. Almatouq, A. A. Alqatari, H. E. Alahmed, D. A. Sharbini, A. F. Alahmadi, F. Alsalman, A. Alsayyah, A. A. Mutair, Artificial intelligence for clinical diagnosis and treatment of prostate cancer, Cancers 14 (2022) 5595. URL: http://dx.doi.org/10.3390/cancers14225595. doi:10.3390/cancers14225595.

[2] Y.-H. Chuang, J.-H. Su, D.-H. Han, Y.-W. Liao, Y.-C. Lee, Y.-F. Cheng, T.-P. Hong, K. S.-M. Li, H.-Y. Ou, Y. Lu, C.-C. Wang, Effective natural language processing and interpretable machine learning for structuring ct liver-tumor reports, IEEE Access 10 (2022) 116273–116286. URL: http://dx.doi.org/10.1109/ACCESS.2022.3218646. doi:10.1109/access.2022.3218646.

[3] H. R. Abdulshaheed, S. A. Mohammed Al-Juboori, I. A. Al Sayed, I. A. Barazanchi, H. M. Gheni, Z. A. Jaaz, Research on optimization strategy of medical data information security and privacy, in: 2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), IEEE, 2022. URL: http://dx.doi.org/10.23919/EECSI56542.2022.9946606. doi:10.23919/eecsi56542.2022.9946606.

[4] E. H. Houssein, R. E. Mohamed, A. A. Ali, Machine learning techniques for biomedical natural language processing: a comprehensive review, IEEE Access 9 (2021) 140628–140653.

[5] C. Li, Y. Zhang, Y. Weng, B. Wang, Z. Li, Natural language processing applications for computer-aided diagnosis in oncology, Diagnostics 13 (2023). URL: https://www.mdpi.com/2075-4418/13/2/286. doi:10.3390/diagnostics13020286.

[6] M. Alawad, H.-J. Yoon, G. D. Tourassi, Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports, in: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), IEEE, 2018. URL: http://dx.doi.org/10.1109/BHI.2018.8333408. doi:10.1109/bhi.2018.8333408.

[7] D. Martinez, Y. Li, Information extraction from pathology reports in a hospital setting, in: Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11, ACM, 2011. URL: http://dx.doi.org/10.1145/2063576.2063846. doi:10.1145/2063576.2063846.

[8] J. DiBello, B. H. Li, C. Zheng, W. Yu, S. Weinmann, K. E. Richert-Boe, D. P. Ritzwoller, S. K. Vandeneeden, S. J. Jacobsen, Development of an algorithm to identify metastatic prostate cancer in electronic medical records using natural language processing., Journal of clinical oncology : official journal of the American Society of Clinical Oncology 32 30_suppl (2014) 164. URL: https://api.semanticscholar.org/CorpusID:25873512.

[9] A. Thomas, C. Zheng, H. Jung, A. Chang, B. J. Kim, J. Gelfond, J. Slezak, K. R. Porter, S. J. Jacobsen, G. W. Chien, Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results, World Journal of Urology 32 (2014) 99–103. URL: https://api.semanticscholar.org/CorpusID:8917027.

[10] O. Hamzeh, L. Rueda, A gene-disease-based machine learning approach to identify prostate cancer biomarkers, in: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2019, pp. 633–638.

[11] J. Morote, I. Schwartzman, A. Borque, L. M. Esteban, A. Celma, S. Roche, I. M. de Torres, R. Mast, M. E. Semidey, L. Regis, et al., Prediction of clinically significant prostate cancer after negative prostate biopsy: The current value of microscopic findings, in: Urologic Oncology: Seminars and Original Investigations, volume 39, Elsevier, 2021, pp. 432–e11.

[12] P. Khosravi, M. Lysandrou, M. Eljalby, Q. Li, E. Kazemi, P. Zisimopoulos, A. Sigaras, M. Brendel, J. Barnes, C. Ricketts, D. Meleshko, A. Yat, T. D. McClure, B. D. Robinson, A. Sboner, O. Elemento, B. Chughtai, I. Hajirasouliha, A deep learning approach to diagnostic classification of prostate cancer using pathology–radiology fusion, Journal of Magnetic Resonance Imaging 54 (2021) 462–471. URL: http://dx.doi.org/10.1002/jmri.27599. doi:10.1002/jmri.27599.

[13] C. Breischneider, S. Zillner, M. Hammon, P. Gass, D. Sonntag, Automatic extraction of breast cancer information from clinical reports, in: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2017. URL: http://dx.doi.org/10.1109/CBMS.2017.138. doi:10.1109/cbms.2017.138.

[14] H.-J. Yoon, J. Gounley, M. T. Young, G. Tourassi, Information extraction from cancer pathology reports with graph convolution networks for natural language texts, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019. URL: http://dx.doi.org/10.1109/BigData47090.2019.9006270. doi:10.1109/bigdata47090.2019.9006270.

[15] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv preprint arXiv:1904.03323 (2019).

[16] H. Lin, J. Ginart, W. Chen, Y. Interian, H. Gong, B. Liu, T. Upadhaya, J. Lupo, J. Hong, S. Braunstein, Oncobert: Building an interpretable transfer learning bidirectional encoder representations from transformers framework for longitudinal survival prediction of cancer patients, 2023. doi:10.21203/rs.3.rs-3158152/v1.

[17] S. S. Seda, F. d. P. P. León, J. M. Conde, M. C. G. Ruiz, J. M. Sánchez, G. Rodríguez, J. A. P. Simón, C. L. P. Calderón, Plataforma para la extracción automática y codificación de conceptos dentro del ámbito de la oncohematología (proyecto coco), Procesamiento del Lenguaje Natural 61 (2018) 65–71.

[18] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results., IberLEF@ SEPLN (2020) 303–323.

[19] O. Solarte-Pabón, O. Montenegro, A. García-Barragán, M. Torrente, M. Provencio, E. Menasalvas, V. Robles, Transformers for extracting breast cancer information from spanish clinical narratives, Artificial Intelligence in Medicine 143 (2023) 102625. URL: https://www.sciencedirect.com/science/article/pii/S0933365723001392. doi:https://doi.org/10.1016/j.artmed.2023.102625.

[20] M. M. Van Berkum, Snomed ct® encoded cancer protocols, in: Amia Annual Symposium Proceedings, volume 2003, American Medical Informatics Association, 2003, p. 1039.

[21] W. H. Organization, Icd-10 : international statistical classification of diseases and related health problems : tenth revision, 2004.

[22] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Compu-

tational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.
[23] J. I. Epstein, W. C. Allsbrook, M. B. Amin, L. L. Egevad, The 2005 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma, American Journal of Surgical Pathology 29 (2005) 1228–1242. URL: http://dx.doi.org/10.1097/01.pas.0000173646.99337.b1. doi:10.1097/01.pas.0000173646.99337.b1.
[24] R. D. Rosen, A. Sapra, Tnm classification., 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK553187/, last accessed: 2023-09-30.

A. Annex 1: Full example

Diagnosis

BIOPSIA DE PRÓSTATA TRANSPERINEAL, PROTOCOLO EXTENDIDO (BASADO EN CAP JUN): - ADENOCARCINOMA CONVENCIONAL. - AI4: ÁPEX IZQUIERDO, TRANSICIONAL: - CILINDROS AFECTADOS / REMITIDOS: 1/1 - GRADO DE GLEASON: 6 (3+3) - GRADO GRUPO: 1 - PORCENTAJE DE PATRÓN GLEASON 4 O 5: 0% - PORCENTAJE DE TEJIDO PROSTÁTICO AFECTADO POR TUMOR: 16,6% - MM DE CARCINOMA / MM DE CILINDRO: 1/6 MM - AI5: ÁPEX IZQUIERDO, ANTERIOR: - CILINDROS AFECTADOS / REMITIDOS: 2/2 - GRADO DE GLEASON: 6 (3+3) - GRADO GRUPO: 1 - PORCENTAJE DE PATRÓN GLEASON 4 O 5: 0% - PORCENTAJE DE TEJIDO PROSTÁTICO AFECTADO POR TUMOR: 61,5% - MM DE CARCINOMA / MM DE CILINDRO: 8/13 MM - AD4: ÁPEX DERECHO, TRANSICIONAL: - CILINDROS AFECTADOS / REMITIDOS:2/2 - GRADO DE GLEASON: 7 (3+4) - GRADO GRUPO: 2 - PORCENTAJE DE PATRÓN GLEASON 4 O 5: 5% - PORCENTAJE DE TEJIDO PROSTÁTICO AFECTADO POR TUMOR: 47,3% - MM DE CARCINOMA / MM DE CILINDRO: 9/19 MM - INFILTRACIÓN GRASA PERIPROSTÁTICA: NEGATIVA. - INFILTRACIÓN DE VESÍCULA SEMINAL: NO VALORABLE POR AUSENCIA DE VESÍCULA SEMINAL EN EL MATERIAL REMITIDO. - INVASIÓN LINFOVASCULAR: NEGATIVA. - INVASIÓN PERINEURAL: NEGATIVA. - PATOLOGÍA ADICIONAL PROSTÁTICA: NO SE OBSERVA.

Macroscopic findings

A.- BD1: BASE DERECHO, PERIFÉRICO POSTERIOR: SE RECIBE UN CILINDRO. INCLUSIÓN TOTAL EN BLOQUE A1. B.- BD2: BASE DERECHO, PERIFÉRICO EXTERNO: SE RECIBE UN CILINDRO. INCLUSIÓN TOTAL EN BLOQUE B1. C.- BD3: BASE DERECHO, PERIFÉRICO ANTERIOR: SE RECIBE UN CILINDRO. INCLUSIÓN TOTAL EN BLOQUE C1. D.- BD4: BASE DERECHO, TRANSICIONAL: SE RECIBE UN CILINDRO FRAGMENTADO. INCLUSIÓN TOTAL EN BLOQUE D1. E.- BD5: BASE DERECHO, ANTERIOR: SE RECIBE UN CILINDRO. INCLUSIÓN TOTAL EN BLOQUE E1. F.- MD1: MEDIO DERECHO, PERIFÉRICO POSTERIOR: SE RECIBE UN CILINDRO MÁS UN FRAGMENTO. INCLUSIÓN TOTAL EN BLOQUE F1. G.- MD2: MEDIO DERECHO, PERIFÉRICO EXTERNO: SE RECIBE UN CILINDRO. INCLUSIÓN TOTAL EN BLOQUE G1. H.- MD3: MEDIO DERECHO, PERIFÉRICO ANTERIOR: SE RECIBE UN CILINDRO MÁS FRAGMENTO. INCLUSIÓN TOTAL EN BLOQUE H1. I.- MD4: MEDIO DERECHO, TRANSICIONAL: SE RECIBE UN CILINDRO. INCLUSIÓN TOTAL EN BLOQUE I1. J.- MD5: MEDIO DERECHO, ANTERIOR: SE RECIBE UN CILINDRO. INCLUSIÓN TOTAL EN BLOQUE J1. K.- AD1: ÁPEX DERECHO, PERIFÉRICO POSTERIOR: SE RECIBE UN CILINDRO. INCLUSIÓN TOTAL EN BLOQUE K1. L.- AD2: ÁPEX DERECHO, PERIFÉRICO EXTERNO: SE RECIBE UN CILINDRO PEQUEÑO. INCLUSIÓN TOTAL EN BLOQUE L1. M.- AD3: ÁPEX DERECHO, PERIFÉRICO ANTERIOR: SE RECIBE UN CILINDRO (EN VARIOS FRAGMENTOS). INCLUSIÓN TOTAL EN BLOQUE M1. N.- AD4: ÁPEX DERECHO, TRANSICIONAL: SE RECIBE UN CILINDRO. INCLUSIÓN TOTAL EN BLOQUE N1. O.- AD5: ÁPEX DERECHO, ANTERIOR: SE RECIBE UN CILINDRO MUY FINO. INCLUSIÓN TOTAL EN BLOQUE O1.

Table 4
Extracted information from example document in Spanish. Each component of the extracted gleason patterns corresponds to the first gleason pattern, the second gleason pattern, the third gleason pattern, the gleason sum and the gleason group, respectively.

Extracted information	Value
cancer tag	Cancer
doc tag	Biopsia
organs detected	Próstata, Vesículas Seminales
extracted gleason patterns	(3, 3, 0, 6, 1), (3, 3, 0, 6, 1), (3, 4, 0, 7, 2)
primary gleason pattern	3
secondary gleason pattern	4
gleason sum	7
gleason group	2