Journal of Biomedical Informatics 116 (2021) 103712

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

Automated classification of cancer morphology from Italian pathology reports using Natural Language Processing techniques: A rule-based approach

Hammami Linda[a,1], Paglialonga Alessia[b,1], Pruneri Giancarlo[c,d], Torresani Michele[e], Sant Milena[a], Bono Carlo[f], Caiani Enrico Gianluca[b,g,2], Baili Paolo[a,*,2]

[a] Analytical Epidemiology and Health Impact Unit, Fondazione IRCCS "Istituto Nazionale dei Tumori", Milan, Italy
[b] Institute of Electronics, Computer and Telecommunication Engineering (IEIIT), National Research Council of Italy (CNR), Milan, Italy
[c] Pathology Department, Fondazione IRCCS "Istituto Nazionale dei Tumori", Milan, Italy
[d] University of Milan, School of Medicine, Milan, Italy
[e] Health Direction, Fondazione IRCCS "Istituto Nazionale dei Tumori", Milan, Italy
[f] Fondazione IRCCS "Istituto Nazionale dei Tumori", Milan, Italy
[g] Politecnico di Milano, Electronics, Information and Biomedical Engineering Dpt., Milan, Italy

ARTICLE INFO

Keywords:
Natural Language Processing
Italian language
Pathology Reports
Cancer morphology

ABSTRACT

Pathology reports represent a primary source of information for cancer registries. Hospitals routinely process high volumes of free-text reports, a valuable source of information regarding cancer diagnosis for improving clinical care and supporting research. Information extraction and coding of textual unstructured data is typically a manual, labour-intensive process. There is a need to develop automated approaches to extract meaningful information from such texts in a reliable and accurate way. In this scenario, Natural Language Processing (NLP) algorithms offer a unique opportunity to automatically encode the unstructured reports into structured data, thus representing a potential powerful alternative to expensive manual processing. However, notwithstanding the increasing interest in this area, there is still limited availability of NLP approaches for pathology reports in languages other than English, including Italian, to date. The aim of our work was to develop an automated algorithm based on NLP techniques, able to identify and classify the morphological content of pathology reports in the Italian language with micro-averaged performance scores higher than 95%. Specifically, a novel, domain-specific classifier that uses linguistic rules was developed and tested on 27,239 pathology reports from a single Italian oncological centre, following the *International Classification of Diseases for Oncology* morphology classification standard (ICD-O-M). The proposed classification algorithm achieved successful results with a micro-F_1 score of 98.14% on 9594 pathology reports in the test dataset. This algorithm relies on rules defined on data from a single hospital that is specifically dedicated to cancer, but it is based on general processing steps which can be applied to different datasets. Further research will be important to demonstrate the generalizability of the proposed approach on a larger corpus from different hospitals.

1. Introduction

Hospital healthcare databases (electronic health records, imaging, laboratory, pathology reports, etc.) include all data concerning the health care provided to patients and the services delivered to improve their health and well-being. The information brought by such a large volume of data offers a unique study opportunity for researchers and physicians.

Various types of data are comprised in health sector databases, including structured variables, (e.g., lab results) and unstructured data

* Corresponding author at: Analytical Epidemiology and Health Impact unit, Fondazione IRCCS "Istituto Nazionale dei Tumori", Via Venezian 1, 20133 Milan, Italy.
E-mail address: lifetable@istitutotumori.mi.it (P. Baili).
[1] Co-first authors.
[2] Co-last authors.

https://doi.org/10.1016/j.jbi.2021.103712
Received 9 September 2020; Received in revised form 8 January 2021; Accepted 8 February 2021
Available online 18 February 2021

(e.g., text documents). A large amount of information is covered by unstructured data, such as clinical documents, patient summaries, and pathology reports, overall representing a significant source of knowledge in support of clinical decisions.

In this scenario, the importance of incorporating Natural Language Processing (NLP) methods in clinical informatics research has been increasingly recognized [1]. NLP is a multidisciplinary field at the intersection of artificial intelligence and linguistics that deals with building computational techniques to automatically analyse and meaningfully represent human language, in a machine-readable format [2].

When dealing with document classification, the first task involves capturing of natural language terms from unstructured free text. A wide range of clinical knowledge and lexical resources exist, including terminologies, vocabularies, taxonomies and ontologies (e.g., the clinical ontology for the oncology domain "*International Classification of Diseases for Oncology*" [3]). In the recent years, a number of systems applied dictionary-based and rule-based approaches for extracting clinical concepts from clinical and biomedical texts. These include MetaMap [4], MedLEE [5], and cTAKES [6] which recognize and encode information related to patient such as symptoms, diseases and treatments. However, these systems are not easily adaptable for the Italian language as they were developed based on clinical information in English.

For instance, in the Italian scenario recent studies attempted to exploit one of the abovementioned systems to process both Italian clinical notes and English translated version of original Italian clinical narratives [7,8]. However, these studies demonstrated important annotation failures due to the smaller coverage of Italian UMLS with respect to the English version, and showed that the possible combination of MetaMap and of an unsupervised translator is subject to substantial failures, mainly due to poor translation.

In the medical domain, ontologies can be used to describe the events and the attributes to be extracted from texts [9]. Pathology reports represent the primary source of information for cancer diagnosis, which relies on the examination of biological specimens. The specimen is examined by a pathologist who generates a report that describes the findings and specific analysis of cells, tissues, and other histopathological indicators, crucial for cancer diagnosis [10].

Over the years, several NLP systems have been developed for the processing of pathology reports in English, mainly containing information about tumour specimens. By contrast, the availability of automated methods in languages other than English is scarce [11]. The review by Burger et al. [11] identified just five papers involving NLP on pathology reports written in languages other than English, and none were in Italian. In general, to date, few efforts on pathology reports exist in the Italian scenario. For example, in the early 2000, Crocioni et al [12] proposed a Microsoft Access application to facilitate the extraction of cancer variables (e.g., morphology, topography, grading, behaviour, etc.) from pathology reports. Presently, an automated free method to analyse and classify pathology reports written in Italian is not available.

In this context, our goal was to develop and test a novel hybrid classifier which integrates NLP tools and linguistic rules to extract cancer morphology codes as defined in the Third edition of International Classification of Diseases for Oncology (ICD-O-3) [3], from pathology reports in Italian aiming at micro-performance scores higher than 95%.

2. Material and methods

2.1. Dataset

The algorithm was developed and tested on pathology reports retrieved by the Oracle Data Warehouse of *Fondazione IRCCS "Istituto nazionale dei tumori" (Istituto Nazionale dei Tumori), Milan (Italy)*. 681,161 reports were registered from the creation of the pathology department's software (January 2003) until the start of the present work (May 2019). The reports were selected through SQL code; each report included one or more specimens, whereas one free-text *diagnosis row* was available for each specimen (number of specimens = 1,043,053) (Fig. 1).

Reports of the first 6 days of April for each year were included in the *training set* (17,798 non null diagnosis rows), while reports of the first 3 days of March for each year were included in the *test set* (9594 non null diagnosis rows). Specifically, the diagnosis rows of both datasets had an average, a median and a 95th percentile length of about 26, 16 and 100 tokens, respectively (more specifically 3787 diagnosis rows with a number of tokens greater than 50). Both training and test sets were initially unannotated, so that they required specific action by a domain expert to generate the gold standard for comparison.

2.2. Ontology

In our work, besides the pathology reports dataset, we also exploited the Italian version of the *International Classification of Diseases for Oncology* (ICD-O) [13], which includes 2335 different morphology descriptions related to 1021 morphology codes (ICD-O-M).

The ICD-O's morphology codes consist of five digits (Fig. 2): the first four digits indicate the histology type, while the last digit, usually after a slash or stroke (/), represents the behaviour code: /0″ for benign, "/1" for borderline, "/2" for in-situ and "/3" for invasive cancers.

Starting from the Italian version of the ICD-O, we constructed an ad hoc ontology in which each combination of morphology tokens has its own priority code (a higher code represents higher priority). An example is shown in Table 1. The ontology was constructed according to the following rules:

- if the morphology description is fully included in another morphology description with a different ICD-O-M code, the most specific one is prioritized. As reported in Table 1, the description "Urothelial carcinoma" is fully included in the more specific description "Urothelial carcinoma in situ". Consequently, a higher code was assigned to the latter morphology;
- if a report includes terms related to more than one morphological code, according to the standard ICD-O guidelines, the numerically higher code is assigned (i.e., the code with the higher behaviour in case of different behaviour codes or the higher histology type code in case of equal behaviour codes).

2.3. ICD-O-M classification algorithm

The present work was implemented in Python version 3.7. The processing pipeline to extract ICD-O morphology (ICD-O-M) code from the *training* and *test sets* is represented in Fig. 3. The first version of the classification algorithm was tested on the *training set*, it was then improved up to the here described final version, and it was finally launched on the *test set*.

Since raw pathology reports were highly unstructured and contained much noise data, NLP techniques were utilized to clean the text (e.g., for misspelling correction) and to reformat the text into a semi-structured format. First, each diagnosis row was subdivided into sub-rows in coincidence with every new paragraph. Then, each sub-row was

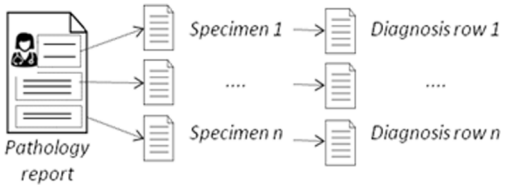

Fig. 1. Representation of pathology reports' composition. A pathology report includes one or more specimens. For each specimen, one free-text diagnosis row is available.

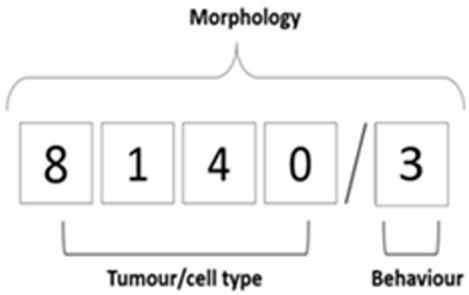

Fig. 2. Structure of morphology code in ICD-O. The first 4 digits indicate the histologic term of the tumor. The fifth digit, after a slash or stroke (/), is a behavior code which indicate whether a tumor is malignant, benign, in situ, or uncertain whether malignant or benign. In this example, the term Adenocarcinoma is coded.

Table 1
Example of ontology records for the token 'urotelial'

Priority code	Morphology (in Italian)	Morphology (in English)	Tokens	ICD-O-M
385220520	Carcinoma uroteliale papillare invasivo	Papillary urothelial invasive carcinoma	['carcinom', 'urotelial', 'papillar', 'invas']	8130/3
385220494	Carcinoma uroteliale invasivo	Urothelial invasive carcinoma	['carcinom', 'urotelial', 'invas']	8120/3
381300201	Carcinoma uroteliale papillare non invasivo	Papillary urothelial carcinoma non-invasive	['carcinom', 'urotelial', 'papillar', 'non invas']	8130/2
381300200	Carcinoma uroteliale papillare	Papillary urothelial carcinoma	['carcinom', 'urotelial', 'papillar']	8130/3
381200500	Carcinoma uroteliale infiltrante	Urothelial infiltrating carcinoma	['carcinom', 'urotelial', 'infiltr']	8120/3
381200101	Carcinoma uroteliale in situ	Urothelial carcinoma in situ	['carcinom', 'urotelial', 'situ']	8120/2
381200100	Carcinoma uroteliale	Urothelial carcinoma	['carcinom', 'urotelial']	8120/3
380000313	Neoplasia uroteliale papillare a basso potenziale di malignità	Papillary urothelial neoplasm of low malignant potential	['neoplas', 'urotelial', 'papillar', 'bass malign']	8130/1
181200200	Papilloma uroteliale	Urothelial papilloma	['papillom', 'urotelial']	8120/1

analysed using the *Natural Language Toolkit* (NLTK) in Python, so as to implement the processing steps listed within the dotted line border in Fig. 3 (i.e., tokenization, stemming, stopwords removal, misspelling correction, N-grams definition, and negation detection), as explained in detail in Sections 2.3.1–2.3.6. Finally, the text was classified into ICD-O-M categories following the procedure described in Section 2.3.7.

2.3.1. Tokenization

The NLTK tokenizer for Italian language was applied to all texts in our dataset. Specifically, tokenization was performed through the use of regular expressions using the *Regexp_tokenize()* function which checks whether an exception rule may be applied from the specified language (e.g., split criteria in presence of punctuations, hyphens etc.). Furthermore, the output of tokenization was pre-processed by setting its content to lower case letters and by filtering out any non-alphanumeric character as it has been observed that they do not influence classification tasks.

Finally, the processed output of tokenization was provided as input for further text cleaning steps.

2.3.2. Stemming

The second text processing step was focused on text normalization and, specifically, on the conversion of tokens to a common root form which can be achieved using stemming or lemmatization techniques. Although both approaches were tested, lemmatization was not performed as it strongly depends on lexical databases (i.e. WordNet [14]) that do not include the required variety of oncology-specific clinical terms in the Italian language.

On the other hand, stemming strips off any suffixes and normalizes several morphological variants of a word into the same form (i.e., stem). Stemming was performed using *SnowballStemmer*, a non-English NLTK stemmer which supports the Italian language [15].

2.3.3. Stop words removal

Stop words removal was applied to further reduce the overall vocabulary in our corpus. List of stop words provided by NLTK for the Italian language model includes 279 tokens. After normalization through the NLTK stemmer, these were reduced to 192 single stemmed stop words. In addition to this list, another 211 stop words stems where manually defined by analysing the reports and identifying frequent words that were not significant for the classification task, e.g., hospital names or typical verbs such as *to present*, *to examine*, *to obtain*, etc. Finally, the stop words removal was implemented by automatically checking whether a token from the text was included in the stop word list.

2.3.4. Misspelling correction

In order to correct any misspellings in the database, we developed a *spell checker* function able to identify misspelled tokens and to return the correct version. To perform this task, we first identified the misspellings. In order to define rare and common tokens, we calculated the number of occurrences of all tokens in the entire dataset (total number of tokens = 14,761,611). Specifically, after a preliminary frequency analysis, we defined an occurrence threshold of 20 on the tokens total. Rare and common tokens were respectively considered as those with an occurrence below and above the defined threshold (30,350 rare tokens were identified). The proposed spell checker scans the rare tokens and checks if they are present in the SNOMED-CT [16] and ICD-O dictionaries [4,7]. For rare tokens that are not coded within these standards, the function computes the edit distance with tokens from the abovementioned dictionaries and from non-rare tokens in our database. Specifically, to calculate the similarity/dissimilarity of two strings, we used the normalized *Levenshtein* edit distance [17], defined as the minimum number of operations required to transform one string into the other. Each of these rare tokens was therefore matched with the word that had the minimal edit distance. Finally, the spell checker function corrects the misspellings in the input text by identifying rare tokens that are not coded in the standard dictionaries thereby replacing each one of them with the matched word on the list.

2.3.5. N-Grams definition for multi-token expressions

This step regards the identification of multi-token expressions so in order for these to be mapped to single tokens (N-Grams). Specifically, during the evaluation of the training set, it was found that many tokens included in specific ICDO-O-M entries might suggest a wrong classification if considered as single tokens. For instance, common sentences in pathology reports such as "*Adenocarcinoma con regressione tumorale*" (Adenocarcinoma with tumor regression) would lead to classification failures as the algorithm would recognize the "*tumor*" stemmed token and wrongly suggest the "*Tumor NOS*" classification (NOS: not otherwise specified) instead of the correct "*Adenocarcinoma*" classification. In this scenario, the identification of short sequences of contiguous words with defined meaning (e.g., bi-grams, tri-grams) was a crucial step for the classification task. Accordingly, specific linguistic rules were introduced

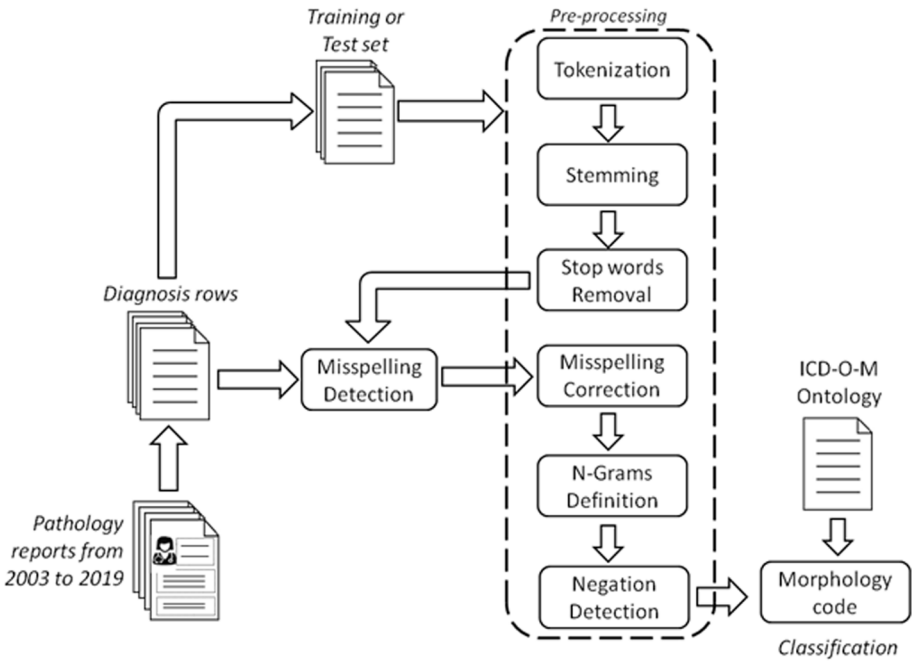

Fig. 3. ICD-O-M Classification Algorithm. Pathology reports (Training set and Test set) are selected from the overall dataset of reports from 2003 to 2019. The pre-processing of raw texts is schematized within the dotted line: tokenization, stemming and stopwords removal performed through the NLTK library tools; misspellings correction performed through a newly developed spell-checker function; definition of N-Grams from the training set, negation detection and definition of negation N-grams. Finally, pathology reports are classified into morphology code according to the ICD-O-M coding schema, through a rule-based approach.

so as to create unique tokens (N-grams) based on the manually identified tokens patterns in the training set (i.e., definition of the "*regression tumor*" bi-gram according to the linguistic pattern represented by the "*regression*" token together with the following token). Specifically, more than 1500 N-grams were identified in the training set, including those representing negation patterns described in the following paragraph.

2.3.6. Negation detection

Negation detection was performed by manually applying a set of defined linguistic rules. First, the *training set* narratives were analysed to discover the negation patterns and to define a list of negation terms. All negation terms commonly used in clinical writings were analysed and assigned to the relevant negation type according to their negation location, which is necessary for a meaningful sentence partition into bi-grams of negation. Specifically, two categories of negation locations were defined: (1) pre-negation and mainly pre-negation terms, and (2) post-negation and mainly post-negation terms.

Once all negation locations had been defined, the algorithm identified whether a sentence included a negation term and, in that case, it merged it with the negated term in a unique bi-gram following the characteristics of negation location (i.e., the negation term "*senza*" (without) is a pre-negation term, consequently it is merged with the negated term represented by the following token).

2.3.7. Classification

To classify the semi-structured text obtained by the above described pre-processing steps into ICD-O-M codes, a rule-based classification algorithm was implemented, as shown in Fig. 3. The proposed algorithm is based on a comparison between the semi-structured, pre-processed texts of the input diagnosis row and the ICD-O-M descriptions. The ICD-O-M standard includes, overall, 2,281 descriptions. For this reason, a list of rules was defined to perform comparisons only where it was necessary, limiting computational complexity. In fact, a number of diagnosis rows are evidently not correlated with some of the ICD-O-M descriptions, for example "*Infezione da Papillomavirus*" (HPV infection), "*Gastrite cronica*" (Chronic gastritis), etc. do not refer to cancer pathologies such as *Lymphoma, Neoplasm, Mesothelioma, Melanoma, Adenocarcinoma* etc.

The following rules were defined:

1. Only ICD-O-M descriptions in which all tokens match the tokens of the pre-processed diagnosis row are considered;
2. If tokens of the pre-processed diagnosis row do not match tokens in the ICD-O-M descriptions, the report is classified as "Not Found/Benign".
3. If the pre-processed diagnosis row matches more than one ICD-O-M description, ICD-O-M standard rules are applied, i.e., the ICD-O-M code with higher order of priority in the ontology is considered (example in Table 1).

2.4. Evaluation of the classification algorithm

The evaluation was performed by comparing the classification proposed by the algorithm with the ICD-O-M classification manually determined by a domain expert with multi-year experience in cancer registration, in line with the available literature for unannotated reports [9]. Specifically, the domain expert annotated both training and test sets following the ICD-O-M standard guidelines also used by the NLP classifier. Based on this gold standard classification, accuracy (A), precision (P), recall (R) and F_1 scores are computed as follows for each classification category:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

where TP = true positives, i.e., the number of records correctly classified by the algorithm; FP = False positives, i.e., the number of morphological classifications detected by the system but not by the expert; and FN = False negatives, i.e., the number of morphological classifications made by the expert that were not detected by the algorithm. Both the above-mentioned performance scores were first computed over all categories (Micro-average) and locally over each category, and then considering the average value (Macro-average).

3. Results

Once the classification model was defined and optimized on the training set (training set performance scores: micro-F_1 = 99.23% and macro-F_1 = 94.28%)), the algorithm was run on a test set including 9,594 diagnosis rows to evaluate the performance of the presented

approach. After manual validation by the domain expert, it was found that the overall dataset (training and test set) contained pathology reports related to 377 different morphologies and characterized by a strongly unbalanced distribution due to the presence of rare morphologies and the highly detailed ICD-O-M classification (i.e., 284 pathology reports referred to the morphology "Adenocarcinoma, NAS" 8140/3, whereas only one referred to the morphology "Siringoma" 8407/0). For these reasons, both micro and macro-averaged performances scores were computed. Specifically, the rule-based approach achieved a micro-F_1 = 98.14% and a macro-F_1 = 84.96%, correctly identifying the morphology in 9,412 diagnosis rows. Considering the ICD-O-M codes in the test set, a high number of pathologies (170 out of 287) achieved F_1 score = 1 (i.e., they were correctly identified in all diagnosis rows) while 28 were rare (less than 5 occurrences in the database) and were never correctly identified, leading to an F_1 score = 0 (e.g., Ductal carcinoma in situ, Papillary cystadenocarcinoma, etc.). Table 2 lists the overall micro and macro-averaged performance measures achieved on the test set as well as for the training set.

To study the performance of the classification model, we analysed the ICD-O-M results according to their behaviour code (fifth digit of ICD-O-M code). The analysis showed a slightly better performance of the presented algorithm when classifying tumour behaviours instead of the first five digits of ICD-O-M, achieving a micro-F_1 of 99.21% with respect to the 98.14% obtained by classifying both morphologies and behaviours.

An analysis of classification mismatches showed that the largest number of classification failures (i.e., about 54.4%, corresponding to 99 out of 182 diagnosis rows without correct ICD-O-M classification) were mostly due to the choice of the morphology among multiple matched descriptions, which corresponded to the numerically highest ICD-O-M code, as established by ICD-O-M guidelines. About 31.3% of failures (57 out of 182 diagnosis rows) were due to the co-presence of single relevant tokens and common or non-relevant tokens. The above-mentioned errors could be limited by introducing new N-grams and stop words, based on sentence patterns that initially caused failure. About 9.3% of errors (17 out of 182 diagnosis rows) were related to failure to detect negations, due to a relatively large distance between the negation term and the negated term. For example, in Italian we observe long patterns such as in "*L'ipotesi di un melanoma non può comunque essere confermata con certezza*" (corresponding to the English translation "The hypothesis of melanoma cannot be confirmed with certainty"). In the example, negated term (*melanoma*) and negation (*non*) are separated by multiple tokens. This leads to the wrong generation of a negation bigram and subsequent failure to detect the negation (i.e. the algorithm wrongly associates the sentence to the presence of a melanoma). The last category of observed errors were due to failure to correct misspellings (5%, 9 out of 182 diagnoses). This was mainly caused by common misspellings that occurred more than 20 times and were thus not addressed by the spell checker algorithm.

4. Discussion

Although it has long been recognized that structured and encoded medical data are useful in clinical practice and research, a large part of clinical information is mostly stored in textual form and is therefore not suitable for use by automated applications. In the oncology domain, one of the main challenges is represented to the automatic extraction of information from free-text pathology reports [11]. In this scenario, NLP systems offer a unique opportunity to automatically encode the unstructured reports into structured data, and represent a powerful alternative to expensive manual processing.

In our study, we developed a new automated *rule-based* approach to classify Italian pathology reports into the ICD-O-M coding schema. NLP methods were applied on a large set of pathology reports written in Italian. These were characterized by a high variability, including diagnoses with complex explanations, different terminologies to label the same cancer type, information about multiple cancers included in a single report, and a highly unstructured nature.

The presented algorithm was developed with the aim to be integrated into the *data warehouse* of the cancer centre *Istituto Nazionale dei Tumori* in order to automatically classify the pathology reports for research purposes. In order to achieve this intended aim, it was necessary to reach at least 95% of micro-averaged performance scores. Therefore, this study exploited ad hoc approaches instead of existing NLP tools, based on available literature that reports higher performances of *in-house* developed NLP methods when encoding pathology reports [11]. Although we acknowledge that a limit of this work is that it relies on rules defined on data from a single hospital specifically dedicated to cancer, high performance was prioritized. However, the generalizability of the approach was addressed, e.g by ensuring variability in the composition of the training set. Specifically, the training set was constructed including a huge amount of diagnosis rows (more than 17,500) covering a specific period (first 6 days of April) for all calendar years. This way, the analysed pathology reports were filled in by different pathologists, with diverse expertise, thus ensuring no bias and inter-observer variability. It should be noted, however, that cancer pathological terms are rather uniform across all types of hospitals in Italy, including general hospitals.

A pre-processing pipeline was developed and applied on pathology reports and on ICD-O-M ontology descriptions. Due to the scarcity of NLP tools for the Italian language for specific pre-processing phases (e.g. negation detection and misspellings correction), we developed specific ad-hoc algorithms.

We considered standard well-established frameworks for the negation detection phase, such as *Negex* [18,19], which rely on the Unified Medical Language System (UMLS) [20] and on a list of negation phrases for the English language, but it was not possible to directly exploit any of them to analyse the narratives of this work. For these reasons, we developed specific algorithms based on linguistic rules (N-grams) that were manually defined by analysing the reports of the training set. The framework here developed could be the basis for developing a future "*Negex* framework" using the negation phrases identified for the Italian language during this work and the ICD-O-M ontology instead of the UMLS.

Furthermore, the overall defined N-grams of this work, including not only those for the negation detection but also those explained in Section 2.3.5, provide a basis for further research with datasets from other institutes (e.g. introducing new N-grams), as pathological linguistic patterns are rather uniform across Italian hospitals.

Similarly, the misspelling pipeline presented in the present work did not follow the existing standard ways for spelling correctors implementation, which use an error model combined with a language model. As a matter of fact, the creation of a robust spelling corrector for the Italian language was beyond the purpose of this work, where the adopted approach was focused on capturing the most common kinds of spelling mistakes. Due to the presence of a single eligible correction among common tokens for each misspelling, a simplified misspelling corrector is presented in order to reasonably identify eventual variations of specific oncological-specific terms in Italian pathology reports. Specifically, the proposed pipeline identified as possible misspellings all "rare" tokens of the overall dataset which were not included in clinical dictionaries, and it exploited an error model (edit distance) for the correction of these misspellings. This method allowed building a list of

Table 2
Overall rule-based classification performance scores.

		Accuracy	Precision	Recall	F_1
Training set	Micro-averaged	99.21%	99.24%	99.22%	99.23%
	Macro-averaged		94.35%	94.68%	94.28%
Test set	Micro-averaged	98.10%	98.15%	98.12%	98.14%
	Macro-averaged		85.99%	86.33%	84.96%

possible misspellings and corresponding corrections.

Although we acknowledge that generalizability of the implemented misspelling corrector might be an issue, it is of note that the proposed pipeline would be easily adaptable to other scenarios where a sufficiently large training set is available.

The proposed rule-based classifier achieved very encouraging results when compared to the gold standard manual classification performed by a domain expert (micro-F_1 equal to 98.14% on the test set), with errors mostly due to the non-identification of rare pathologies (often not present in the training set) and to complex negation sentences.

To the best of our knowledge, this is the first work that has analysed such a large dataset of highly varied pathology reports and achieving such high performance scores. The algorithm presented in our study is derived from the analysis of a huge set of training data (globally, over 27,000 texts) from which we were able to extract a large set of relevant linguistic patterns, as demonstrated by the successful results obtained on the test set.

Specifically, our dataset was characterized by a strongly unbalanced distribution of pathologies across the reports (i.e., in the test set 243 morphologies were present in less than 20 reports and only 6 morphologies were present in more than 100 reports). Our proposed approach achieved better results compared to other rule-based approaches which were previously applied in other non-English pathology reports (e.g., micro-$F_1 = 98.14\%$ compared to micro-$F_1 = 91\%$ in a study on Norwegian reports [21]).

In fact, the presented work may be validated in other cases, for example by *population-based* cancer registries [22] to extract morphology in ICD-O-M when their pathological reports are unannotated. Furthermore, the classifier could be used to facilitate the direct ICD-O-M code annotation when pathologists write the diagnosis, thus submitting the morphology codes extracted by the algorithm to a validation and, eventually, a correction from pathologists.

The algorithm may also be improved by means of further research. For example, it would be useful to explore ways in which it could be adapted to classify topography according to the ICD-O coding system that describes the neoplasms sites of origin. This way, the ICD-O-3 topography and morphology extracted from a pathology report may be combined and thus even be converted in the ICD-9-CM coding used in Italy by Hospital Discharge Records [23] or in ICD-10 codes using available conversion tables [24].

5. Conclusions

The present study addressed the task of automated assignment of *International Classification of Disease in Oncology morphology codes* (ICD-O-M) to free-text pathology reports in the Italian language, for which available linguistic NLP tools are still limited.

A novel, domain-specific, NLP-based classifier which relies on ad hoc linguistic rules defined on a large dataset of 27,239 records was developed and tested, achieving a micro-F_1 score of 98.14% on a test set of 9,594 records compared to expert annotations when classifying at a level of detail as that of the ICD-O-M, i.e. including morphologies and behaviours. Further research would be needed to fully address the generalizability of the approach to pathology reports coming from different centres, e.g., cancer-specific centres as well as general hospitals, and to further develop the algorithm in order to extract additional relevant information such as topography codes from the ICD-O standard.

CRediT authorship contribution statement

Hammami Linda: Conceptualization, Methodology, Software, Writing - original draft. **Paglialonga Alessia:** Conceptualization, Methodology, Software, Writing - original draft. **Pruneri Giancarlo:** Writing - review & editing, Supervision. **Torresani Michele:** Writing - review & editing. **Sant Milena:** Writing - review & editing, Supervision. **Bono Carlo:** Software, Writing - review & editing, Supervision. **Caiani Enrico Gianluca:** Conceptualization, Methodology, Software, Writing - original draft. **Baili Paolo:** Conceptualization, Methodology, Software, Validation, Writing - original draft.

Declaration of Competing Interest

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.

Acknowledgements

This work was supported by Fondazione IRCCS "Istituto Nazionale dei Tumori", 5x1000 Ministry of Health funds under the project "Institutional data warehouse pilot platform for research".

References

[1] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, A.D. Shah, K. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, W. Chapman, R. Dutta, Using clinical Natural Language Processing for health outcomes research : Overview and actionable suggestions for future advances, J. Biomed. Inform. 88 (2018) 11–19, https://doi.org/10.1016/j.jbi.2018.10.005.

[2] P.M. Nadkarni, L. Ohno-Machado, W.W. Chapman, Natural language processing: An introduction, J. Am. Med. Informatics Assoc. 18 (2011) 544–551, https://doi.org/10.1136/amiajnl-2011-000464.

[3] A. Fritz, C. Percy, A. Jack, K. Shanmugaratnam, L. Sobin, M. Parkin, S. Whelan, ICD-O International Classification of Diseases for Oncology First Revision, Third Edit, 2013. www.who.int (accessed March 1, 2020).

[4] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proc. AMIA Symp. (2001) 17–21.

[5] C. Friedman, L. Shagina, Y. Lussier, G. Hripcsak, Automated encoding of clinical documents based on natural language processing, J. Am. Med. Informatics Assoc. 11 (2004) 392–402, https://doi.org/10.1197/jamia.M1552.

[6] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C. G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications, J. Am. Med. Informatics Assoc. 17 (2010) 507–513, https://doi.org/10.1136/jamia.2009.001560.

[7] E. Chiaramello, A. Paglialonga, F. Pinciroli, G. Tognola, Attempting to use meta map in clinical practice: A feasibility study on the identification of medical concepts from Italian clinical notes, Stud. Health Technol. Inform. 228 (2017) 28–32, https://doi.org/10.3233/978-1-61499-678-1-28.

[8] E. Chiaramello, F. Pinciroli, A. Bonalumi, A. Caroli, G. Tognola, Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes, J. Biomed. Inform. 63 (2016) 22–32, https://doi.org/10.1016/j.jbi.2016.07.017.

[9] N. Viani, T.A. Miller, C. Napolitano, S.G. Priori, G.K. Savova, R. Bellazzi, L. Sacchi, Supervised methods to extract clinical events from cardiology reports in Italian, J. Biomed. Inform. 95 (2019) 103219, https://doi.org/10.1016/j.jbi.2019.103219.

[10] S. Gao, M.T. Young, J.X. Qiu, H.J. Yoon, J.B. Christian, P.A. Fearn, G.D. Tourassi, A. Ramanthan, Hierarchical attention networks for information extraction from cancer pathology reports, J. Am. Med. Informatics Assoc. 25 (2018) 321–330, https://doi.org/10.1093/jamia/ocx131.

[11] G. Burger, A. Abu-Hanna, N. De Keizer, R. Cornet, Natural language processing in pathology: A scoping review, J. Clin. Pathol. 69 (2016) 949–955, https://doi.org/10.1136/jclinpath-2016-203872.

[12] E. Crocioni, C. Sacchettini, A. Caldarella, E. Paci, Automatic coding of pathologic cancer variables by the search of strings Tuscany Cancer Registry, Epidemiol. Prev. 29 (2005) 57–60.

[13] A. Giacomin, S. Ferretti, Icd-O, Third Edit, 2000.

[14] C. Fellbaum, WordNet, Encycl. Appl. Linguist. (2012), https://doi.org/10.1002/9781405198431.wbeal1285.

[15] nltk.stem.snowball — NLTK 3.5 documentation, (n.d.). https://www.nltk.org/_modules/nltk/stem/snowball.html (accessed August 4, 2020).

[16] SNOMED - Home | SNOMED International, (n.d.). http://www.snomed.org/ (accessed August 17, 2020).

[17] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Cybern. Control Theory (1996).

[18] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, B.G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. 34 (5) (2001) 301–310, https://doi.org/10.1006/jbin.2001.1029.

[19] H. Harkema, J.N. Dowling, T. Thornblade, W.W. Chapman, ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports, J. Biomed. Inform. 42 (5) (2009) 839–851, https://doi.org/10.1016/j.jbi.2009.05.002.

[20] D.A.B. Lindberg, B.L. Humphreys, A.T. McCray, The unified medical language system, Methods Inform. Med. 32 (4) (1993) 281–291, https://doi.org/10.1055/s-0038-1637976.

[21] R. Weegar, J.F Nygård, H. Dalianis, Efficient encoding of pathology reports using natural language processing, (2017) 778–783. https://doi.org/10.26615/978-954-452-049-6_100.
[22] "The population-based cancer registries | ECIS." https://ecis.jrc.ec.europa.eu/info/cancer_registries.html (accessed Nov. 04, 2020).
[23] "Ricoveri ospedalieri (SDO)." http://www.salute.gov.it/portale/temi/p2_4.jsp?lingua=italiano&area=ricoveriOspedalieri (accessed Nov. 02, 2020).
[24] "ICD Conversion Programs - SEER." https://seer.cancer.gov/tools/conversion/ (accessed Nov. 02, 2020).