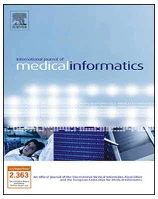
International Journal of Medical Informatics 127 (2019) 141–146

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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



# Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation

Matthias Becker[a,*], Stefan Kasper[c], Britta Böckmann[a,b], Karl-Heinz Jöckel[b], Isabel Virchow[c]

[a] Department of Computer Science, University of Applied Sciences and Arts, Dortmund, Germany
[b] Institute of Medical Informatics, Biometry and Epidemiology, University Hospital Essen, Germany
[c] West German Cancer Center, University Hospital Essen, Germany

ARTICLE INFO

*Keywords:*
Natural language processing
Electronic health records
Information extraction
Knowledge management

ABSTRACT

*Background:* Colorectal cancer is the most commonly occurring cancer in Germany, and the second and third most commonly diagnosed cancer in women and men, respectively. The therapy for this disease is based primarily on the tumor stages, which are usually documented in an unstructured form in medical information systems. In order to re-use this knowledge, the information must be extracted and annotated using the correct terminology.
*Methods:* In this study, a natural language processing pipeline is developed to identify specific guideline-based patient information and to annotate it with Unified Medical Language System concepts for manual evaluation by a physician. The gold standard for one-time evaluation is determined using the human abstraction of 2513 German clinical notes from electronic health records.
*Results:* Using this approach to process the narrative clinical notes on colorectal cancer for retrospective evaluation of the therapy recommendation, the algorithm achieves a precision value of 96.64% for tumor stage detection and 97.95% for diagnosis recognition with recall values of 94.89% and 99.54%, respectively. The average precision value across all concepts relevant to treatment decisions for patients with known cancer diagnoses (11 concept groups) achieved a precision value of 82.05% with a recall value of 82.45% and an F1-score of 81.81%, respectively.
*Conclusions:* The identification of guideline-based information from narrative clinical notes has the potential for implementation as clinical decision support tools.

## 1. Introduction

In Germany, colorectal cancer is the second most common cancer in women and the third most common in men [1]. Colon and rectal cancer have many similarities in etiology and histology [2]. However, they differ in their preoperative, operative, and adjuvant therapeutic strategies. The prognosis of patients with colorectal cancer is impacted by the tumor stage at the time of diagnosis and other biomarkers [3,4]. The therapy is based primarily on the tumor stages defined by clinical guidelines such as the German S3 guidelines [5]. In order to choose the optimal, guideline-oriented therapy for a patient, it is necessary to obtain some specific information that can usually be found as unstructured documentation in the medical information system [6].

Electronic medical information systems are now used in almost all healthcare facilities, replacing traditional handwritten medical reports [7]. Thus, structured diagnoses, clinical symptoms, lab values, radiological examinations, molecular biomarkers, and many other treatment-related data are electronically recorded and managed. However, a large amount of clinical data is documented in unstructured or undefined formats. To re-use these unstructured data for the development of individual treatment algorithms, it is necessary to extract and structure the data by natural language processing (NLP) through the enhancement and annotation of the correct terminology [8,9]. The technical requirements and further development of semantic analysis methods, such as text mining, are the basis for structured data extraction in the development of treatment algorithms. The NLP of clinical text notes or clinical outcome parameters has recently been extensively developed [10]. In the clinical setting, NLP has been found to be helpful for retrospective studies and clinical decision making [11,12]. There are numerous functional biomedical NLP applications in English [10]. In contrast, clinical NLP applications in languages other than English are rare; however, German and French are the most investigated

[*] Corresponding author at: University of Applied Sciences and Arts, Emil-Figge-Str. 42, 44227 Dortmund, Germany.
*E-mail address:* Matthias.Becker@fh-dortmund.de (M. Becker).

https://doi.org/10.1016/j.ijmedinf.2019.04.022
Received 2 February 2019; Received in revised form 7 April 2019; Accepted 24 April 2019

languages in the PubMed database [10]. In the past, various applications have been successfully implemented in German for different clinical questions [13–16].

In this study, algorithms were developed using NLP techniques to extract and summarize information from plain text clinical notes in German to analyze whether the quality of the results was sufficient to create a prerequisite for the development of a clinical decision support system (CDSS) based on health records for colon and rectal cancer.

## 2. Related research

Several studies have been performed on the staging classification and terminology extraction in pathology reports with the aim of structuring their contents [33]. Most approaches focus on English texts and use both rule-based systems and machine learning systems as well as a combination of the two [34].

McCowan et al. [35], Nguyen et al. [36], and Martinez et al. [37] used text mining to perform cancer classification according to the TNM Classification of Malignant Tumors (TNM) [3]. In the field of colorectal cancer, Martinez et al. [37] obtained F-scores of 81%, 85%, and 94% for staging tumor, node, and metastases respectively for colorectal cancer pathology reports, using 200 pathology reports for training and evaluation. A machine learning approach was applied as the solution approach to classify the pathology reports according to the TNM staging scale. The Weka toolkit and the Naïve Bayes and SVN (support-vector machine) algorithms for the best results were used. Coden et al. [38] developed a model called the Cancer Disease Knowledge Representation Model, which achieved a recall on colon cancer reports between 76% and 100% and a precision between 72% and 100% for all classes except a lower precision and recall for metastatic tumors.

Similar work exists for other types of cancer. Currie et al. [39] constructed a rule-based system to extract concepts from 5826 breast cancer and 2838 prostate cancer pathology reports. The authors obtained around 90–95% accuracy for most of the 80 extracted fields, using domain experts for the evaluation. For other tumor diseases, such as breast cancer, approaches exist for non-English texts such as the example in Weegar et al. [40]. A rule-based system for text mining of Norwegian breast cancer pathology reports was developed and achieved an average precision of 80%, a recall of 98%, and an F-score of 86%. For German-language documents there is no comparable work in the area of staging classification for colorectal cancer.

## 3. Methods

### 3.1. Study setting

This study was conducted at the West German Cancer Center, University Hospital in Essen, Germany and received ethical permission from the Research Ethics Committee in Essen. The analyzed cohort consisted of 500 patients with known colorectal cancer for whom health records were available. In order to generate, adapt, and verify an algorithm during the learning process, three different groups were needed [17,18]. From the complete dataset of 500 patients, data which later served as the test group, were retrieved before beginning the learning process. Therefore, 45% of the complete dataset was used as the training group, 45% as the one-time evaluation group of the final system, and 10% as the test group displayed in Table 1. Since the documents taken from the electronic health record (EHR) all originate from the same hospital, it can be assumed that they exhibit strong similarities. Therefore, a small test group was chosen for this approach to obtain the largest possible evaluation group in order to carry out a meaningful evaluation.

A total of 2513 German clinical notes, which included 820 medical reports, 817 radiology reports, 107 microbiology reports, 326 pathology reports, 20 virology reports, and 423 tumor board protocols were utilized for the evaluation. The tumor board protocols included the decisions about treatment in connection with the patient's disease, such as the tumor stage and whether a tumor was resectable. The clinical notes were in PDF files, which were converted into text files, and these clinical notes were either unstructured or semi-structured and only contained native electronic text.

**Table 1**
Distribution of patient groups for training, test, and evaluation.

| Groups | Absolute number of patients | Percentage |
|---|---|---|
| Training group | 225 cancer patients (2490 clinical notes) | 45% |
| Test group | 50 cancer patients (503 clinical notes) | 10% |
| Evaluation group | 225 cancer patients (2513 clinical notes) | 45% |
| Total | 500 cancer patients (5506 clinical notes) | 100% |

The training group was used in the learning process to create and adapt the model. The learning procedures were divided into different types of learning. If all the attribute values and thus also those of the attributes to be predicted were known in the training set examples—as in this work—and were used by the NLP pipeline, then they were referred to as *supervised learning* or *learning from examples* [32]. The advantage of this form of learning is that the known values of the attributes, which support the prediction of the training set examples, can be used in the construction of the algorithm; this method is generally a more effective form of learning [32]. The disadvantage of this method is that a large amount of training must be provided compared to other training methods, in which the categorization and annotation must be known.

A semi-automated rule-based system to augment the terminology was devised. Since the categorization and annotation in the training dataset must be known to train the system, the training dataset was preprocessed by pattern analysis using a rule-based system with regular expressions and morpho-syntactic rules [41] for terminology extraction that collects domain-relevant terms from a corpus of domain-specific documents. The training data set was then optimized using hand-coded rules and manual annotation, and the rule and annotation set was iteratively applied to the training set. The German terms from the Unified Medical Language System (UMLS) database were applied as the starting set, including the results from the manual analysis. Newly identified terms—for example, synonyms, common abbreviations, and adjectival forms of a word—were added to the start set. The trained system was then tested several times against the test group and the pipeline for optimization. The evaluation data set was used once after completion of the test phase to calculate the performance of the pipeline.

Table 2 depicts the evaluation distribution of the patients according to their tumor stages. The classification of the tumor and metastasis was based on the TNM criteria.

In gastrointestinal tumors, T describes the infiltration into the wall of the colon or rectum, N signifies the involved loco-regional lymph nodes, and M represents distant metastasis. The tumor stage was classified based on the TNM status using the Union for International Cancer Control (UICC) staging system [19]. The analyzed patient cohort consisted of patients with a first diagnosis and patients with a relapse or metachronous metastasis. Different TNM statuses were documented for

**Table 2**
Evaluation dataset (patients and clinical notes) for the evaluation of the NLP pipeline.

| Tumor stages: | Stage IV | Stage III | Stage II | Stage I | Total |
|---|---|---|---|---|---|
| Colon cancer patients (1483 clinical notes) | 66 | 20 | 15 | 14 | 115 |
| Rectal cancer patients (1030 clinical notes) | 57 | 14 | 33 | 6 | 110 |
| Total (2513 clinical notes) | 123 | 34 | 48 | 20 | 225 |

181 patients. For these patients, the advanced stage was chosen since an improvement of the stage could not be assumed during the course of the disease.

For the validation and evaluation of the algorithm, a retrospective manual analysis of 225 patients with colorectal cancer was performed. The sample was chosen to define the actual distribution of tumor stages and was selected according to the International Statistical Classification of Diseases and Related Health Problems (ICD) codes from the electronic health records. No other restrictions were imposed, and the sample was random.

### 3.2. Semantic standards

To standardize the communication and modeling knowledge, terminologies and ontologies are used. In medicine, the UMLS is generally used [20]; the UMLS is a meta-ontology that summarizes various medical terminologies and their translations, including important terminologies for anatomy, clinical terminologies, and oncology. The medical concepts and terms with their synonyms and translations are relationally linked in the UMLS via a semantic network. In this way, hierarchical relations between concepts, such as “a heart attack is a cardiovascular disease” or “the bronchi are part of the lungs,” can be identified by inference. Even more complex ontological relations, such as “location of” or “is adjacent to,” are modeled in the UMLS and can be used for anatomical inferences.

### 3.3. Gold standard

The gold standard has been defined in collaboration with a medical oncologist and the relevant clinical guidelines to determine the appropriate information as necessary for an accurate recommendation that conforms to the guidelines. Based on this information, two clinical pathways were developed: one for colon cancer and the other for rectal cancer (see Appendix A and Appendix B). These paths were annotated by UMLS concepts to create a link between the patient-specific data and the evidence-based knowledge (clinical guidelines) to help the physicians reach a therapeutic decision.

Table 3 lists the gold standard as the minimum guidelines-based UMLS annotation dataset required to identify the treatment of a patient on a colorectal clinical path as well as the concepts that were annotated in the 2513 evaluation documents. The annotation took place after the analysis through the NLP pipeline to avoid any bias for the comparison between the gold standard and the algorithm performance. The documents were annotated by a computer scientist and consolidated by a physician using the brat rapid annotation tool [21]. All of the German synonyms for the respective UMLS terms were marked during annotation, regardless of whether they were in the UMLS database. The overall inter-annotator agreement between the two annotators—the physician and the computer scientist—was 97%. The clinical criteria for annotation were based upon the current clinical practice guidelines.

### 3.4. Natural language processing

For the extraction of UMLS concepts from the German clinical notes, an NLP pipeline with a mapping to the UMLS database was required. In this work, a natural language processing based algorithm for entity recognition with UMLS concept mapping for the German language was developed. Based on a previous study, the pipeline was extended by various German synonyms in the UMLS database [22] and regular expressions for the preprocessing of the clinical notes. For negation detection, NegEx was used with a German database for negations [23,24,42]. During the learning process of the database, a custom database extended by German synonyms was built [25,26].

Fig. 1 presents the overview architecture of the system. For the evaluation of the pipeline, 225 evaluation cases (2513 documents) were manually annotated and used as the gold standard (lower swim lane). Similarly, the data was automatically annotated by the NLP pipeline standard (upper swim lane). The evaluation compared the results of the NLP pipeline with the gold-standard manual abstraction.

### 3.5. Statistical analysis

For this approach, the identified concepts were compared with the gold standard of the respective documents. The analysis was performed through the identification of each of several concepts at the level of individual mentions of concepts within the documents. This process can lead to several concepts appearing in a document. The following values were determined: precision or positive predictive value (PPV), recall or true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), false positive rate (FPR), accuracy (ACC), and F1-measure (F1), for which recall and precision were equally weighted. These measures were calculated as follows:

$$PPV = \frac{true\quad positive}{true\quad positive + false\quad positive} \qquad F1\quad Score = \frac{2 * PPV * TPR}{PPV + TPR}$$

$$TPR = \frac{true\quad positive}{true\ positive + false\ negative} \qquad FPR = \frac{false\quad positive}{false\quad positive + true\ negative}$$

$$TNR = \frac{true\quad negative}{true\ negative + false\quad positive} \qquad FNR = \frac{false\quad negative}{true\ positive + false\quad negative}$$

$$ACC = \frac{true\ positive + true\ negative}{true\ positive + true\quad negative + false\ positive + false\quad negative}$$

A chi-squared test was applied to evaluate the statistical independence of the differences in the distribution between colon and rectal cancer. The statistical significance was determined to be

**Table 3**
Gold standard annotation for evaluation based on the clinical pathways.

| | Item (textual examples from the documents) | UMLS Concepts (CUI) |
|---|---|---|
| ICD | International Statistical Classification of Diseases and Related Health Problems, Codes: C18, C19, C20 (C18.1, rectal cancer, colorectal cancer) | C0007102, C0153443, C0949022 |
| TNM | tumor stages I, II, III, IV (T3N1M1, T3 NX M0, UICC IV) | C0278474, C0278479, C0278480, C0278484 |
| ANO | cm from anus.(11 cm from ano, 13 cm a.a.) | C0178825 |
| MSI | microsatellite instability (MSI, MSS, MS stable) | C0920269, C4321493 |
| RES | resection potential (resectability, respectable) | C1514888 |
| MUT | mutation statuses (RAS, KRAS, BRAF, K-RAS) | C0034677, C2747835, C0812241 |
| IT | intensive therapy (IT, intensive chemotherapy, IC) | C0085559 |
| TU | large tumor burden (LTB, tumor load, cancer cells) | C1449699 |
| RP | rapid progress (RP, tumor regression) | C0456962 |
| TS | tumor symptoms (rectal bleeding, blood in stool, diarrhea) | C3846098 |
| OC | organ complication (OC, liver failure) | C0178784 |

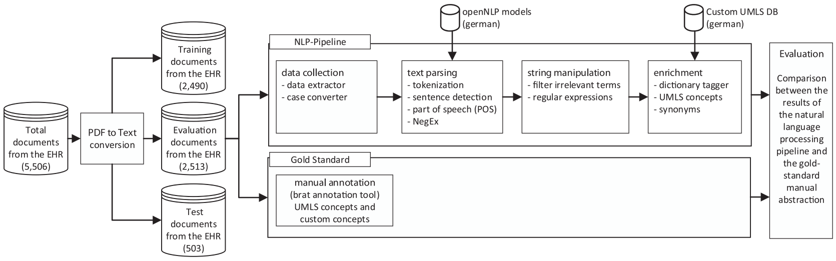


**Fig. 1.** Study design- The complete record of 5506 documents was converted by a PDF to text converter and divided into three groups. The evaluation group was analyzed by the trained NLP pipeline and compared with the same manually annotated (gold standard) documents to calculate the performance of the NLP pipeline.

$p < 0.05$.

## 4. Results

The colorectal cancer cohort dataset for evaluation contained 225 patients. There were 115 colon cancer cases and 110 rectal cancer cases. The dataset contained 2513 clinical notes. The mean age of the patients in the dataset was 62.9 years, and 45% of the subjects were female. Tables 4 and 5 describe the overall performance of the identified UMLS concepts by the NLP pipeline for colon and rectal cancer.

Using this approach for processing narrative clinical notes on colorectal cancer for retrospective evaluation of the therapy recommendation, the algorithm achieved an average precision value across all analyzed elements—11 concept groups in Table 5—of 82.05% with a recall value of 82.45% and an F1-score of 81.81%, respectively. The identification of the tumor status, the diagnosis, and the distance of the tumor from the anus for rectal cancers have a precision and recall of over 90% and, accordingly, the highest balanced F-score compared to the other results (Table 5). This information is crucial for the evaluation and recommendation of a therapy as it is used for the initial categorization within the guidelines. It can be observed that items which were rarely or never negated in the documents indicate a true negative rate of zero or very low numbers, thus presenting higher results than the other values. When identifying the diagnoses, one must consider that these are entered automatically into structured documents similar to those in the ICD catalogue of the hospital information system and are only rarely entered manually by the physician in free text. The data for the TNM classification and distance from the anus were preprocessed using regular expressions for more specific identification. The worst performance was achieved for organ complications and resection potential, in which the F1-score was below 70% with an equally low accuracy. These moderate results can be attributed to the fact that organ complications are usually complexly documented and described. Resectability is documented heterogeneously and is usually based on image data. This information is often removed from the textual context and is, therefore, difficult to identify automatically.

**Table 4**
Performance of the identified UMLS concepts of the evaluation documents (for abbreviations, see Table 3).

| | TP | TN | FP | FN | Total |
|---|---|---|---|---|---|
| ICD | 430 | 4 | 9 | 2 | 445 |
| TNM | 576 | 8 | 20 | 4 | 608 |
| MSI | 61 | 20 | 19 | 12 | 112 |
| ANO | 89 | 1 | 14 | 9 | 113 |
| RES | 130 | 31 | 65 | 54 | 280 |
| MUT | 112 | 65 | 26 | 32 | 235 |
| IT | 16 | 3 | 5 | 5 | 29 |
| TU | 11 | 2 | 4 | 2 | 19 |
| RP | 10 | 7 | 2 | 9 | 28 |
| TS | 21 | 2 | 6 | 2 | 31 |
| OC | 35 | 6 | 5 | 9 | 55 |

TP = true positives, FN = false negatives, FP = false positives, TN = true negatives.

**Table 5**
Statistics of the UMLS concepts identified by the NLP pipeline (for abbreviations, see Table 3).

| | PPV | TPR | TNR | FNR | FPR | ACC | F1-Score | CHI (p) |
|---|---|---|---|---|---|---|---|---|
| ICD (%) | 97.95 | 99.54 | 30.77 | 0.46 | 69.23 | 97.53 | 98.74 | 0.8889 |
| TNM (%) | 96.64 | 94.89 | 28.57 | 5.11 | 71.43 | 91.97 | 95.76 | 0.9617 |
| MSI (%) | 76.25 | 83.56 | 51.28 | 16.44 | 48.72 | 72.32 | 79.74 | 0.9963 |
| ANO (%) | 90.18 | 91.82 | 35.29 | 8.18 | 64.71 | 84.25 | 90.99 | 0.7292 |
| RES (%) | 66.67 | 70.65 | 32.29 | 29.35 | 67.71 | 57.50 | 68.60 | 0.8996 |
| MUT (%) | 81.16 | 77.78 | 71.43 | 22.22 | 28.57 | 75.32 | 79.43 | 0.9509 |
| IT (%) | 76.19 | 76.19 | 37.50 | 23.81 | 62.50 | 65.52 | 76.19 | 0.9345 |
| TU (%) | 73.33 | 84.62 | 33.33 | 15.38 | 66.67 | 68.42 | 78.57 | 0.9980 |
| RP (%) | 83.33 | 52.63 | 77.78 | 47.37 | 22.22 | 60.71 | 64.52 | 0.9937 |
| TS (%) | 77.78 | 91.30 | 25.00 | 8.70 | 75.00 | 74.19 | 84.00 | 0.9344 |
| OC (%) | 87.50 | 79.55 | 54.55 | 20.45 | 45.45 | 74.55 | 83.33 | 0.9734 |

PPV = precision or positive predictive value, TPR = recall or true positive rate, TNR = true negative rate, FNR = false negative rate, FPR = false positive rate, ACC = accuracy, F1 = F-measure; CHI (p) = chi-squared test.

The chi-squared test under the null hypothesis that there is no association between colon and rectal cancer ($p < 0.05$) indicates that there is a strong correlation between the results of the NLP pipeline for colon and rectal cancer. The null hypothesis that there is no association can be rejected.

## 5. Discussion

Using this approach for processing narrative clinical notes on colorectal cancer for retrospective evaluation of the therapy recommendation, the algorithm achieved an average precision value across all analyzed elements of 82.05% with a recall value of 82.45% and an F1-score of 81.81%, respectively. The best values were achieved for ICD and TNM recognition, and the worst performance was achieved for organ complications and resection potential. The differences in performance indicate that language heterogeneity and idiosyncrasies introduce challenges for clinical NLP. By examining the measured values and the chi-squared test, it is evident that there are no significant differences between the two diseases observed here. This result can be attributed to the fact that both colon and rectal cancers have many similarities in etiology and histology. However, they differ in preoperative, operative, and adjuvant therapy strategies.

The results suggest that the preprocessing of clinical documents by NLP can positively impact the improvement of the CDSS. The identification of guideline-based information from narrative clinical notes has the potential for implementation as decision support tools. However, it can only be used as a tool for monitoring, since no error-free statements can be made about the therapy decision at the specified performance values.

A further analysis reveals that the descriptions pertaining to the possibility of metastasis resectioning and organ complications in the documents are complex. In terms of resectability, the German S3 guidelines propose a pragmatic division of stage IV patients into three groups based on the primary objective of personalized therapy [4]. The first group consists of the primarily technically resectable liver or lung metastases, the second group requires an intensified systemic therapy, and the third group consists of patients with multiple metastases without the option for resectioning after metastasis regression as well as those without tumor-related symptoms or organ complications. These three groups cannot be reliably identified by NLP techniques alone. For example, it is necessary to create a decision tree that can divide the results into groups similar to those carried out and outlined in the solution approach of tumor classification. The decisions regarding the resectability of distant metastases are the task of interdisciplinary tumor boards in which the resectability is usually determined by the interpretation of radiologic images. A combination of text mining and image mining may bring a significant improvement in the results [27,28]. Furthermore, using a combination of risk factors and laboratory parameters that were not considered in the current approach may improve the results.

The results indicate that the identification of specific items, such as diagnoses, that are either not or only slightly negated in the documents achieves significantly higher results than those that are often negated. In cases such as microsatellite instability, this does not always have an impact on the therapy decision because this information is not used in all treatment decisions. The most frequently misrecognized negations were those that were not negated in the same or subsequent sentence, such as "*A BRAF determination was commissioned. The mutation status was available within* 8 h *of the FFPE tissue block. The above provision was negative.*" In this case, NegEx could no longer establish a connection between the BRAF mutation and the later mentioned determination, which referred to the mutation status. The most common errors in the falsely detected negation were, for example, double mentioned concepts after negations: "There is no microsatellite instability (microsatellite stability)" where the negation detection recognized everything after the word "no" as a negation. Thus, microsatellite instability and microsatellite stability were detected as negated, which led to an increased false negative rate. Wu et al. revealed, by means of a multi-corpus analysis of negation detection, that it is easy to optimize for a single corpus but not to generalize to arbitrary clinical texts [43]. Therefore, improving the negation detection is a difficult challenge, for example, with regular expressions.

Another challenge of the project was that although the UMLS was available and contained sufficient coverage of general oncological concepts, the German translations were incomplete. The structure of specific ontologies is mentioned in the literature [29], but the results are not publicly available. Therefore, synonyms had to be introduced into the annotation dataset.

To optimize the results, the following should be addressed in future works. Further data from other hospitals should be analyzed to extend the German synonym dataset, and the pipeline should be extended by decision trees and trained with further data. Clinical practice as well as research and quality assurance benefit from clear clinical information as referred to in the use of common terminology [30]. This common terminology is necessary for the consistent reuse of data and to support semantic interoperability [31].

## 6. Conclusions

For the purpose of therapy recommendations in colorectal cancer, the algorithm in this approach of processing German clinical notes for retrospective evaluation achieved a precision value of 96.64% for the detection of tumor stages and 97.95% for the diagnosis recognition with recall values of 94.89% and 99.54%. Thus, this approach demonstrates the potential to support physicians in decision making for personalized therapy. The algorithm for the automatic identification of guideline-relevant information from the clinical texts eventually led to an improved quality of care for colorectal patients according to clinical guidelines.

## Authors contribution

Conception and design: MB, SK, BB, KHJ, IV.
Analysis and interpretation: MB, SK, IV.
Data collection: MB, IV.
Writing the article: MB, SK, BB, KHJ, IV.
Critical revision of the article: SK, BB, KHJ.
Final approval of the article: MB, IV.
Statistical analysis: MB, IV.
Obtained funding: MB, SK.
Overall responsibility: MB.

## Conflict of interest

The authors declare that there is no conflict of interest.

## Funding sources

The research reported in this publication was supported by the West German Cancer Center, University hospital Essen in Essen, Germany and Institute for Medical Informatics, Biometry and Epidemiology, University hospital Essen in Essen Germany. This study was carried out free of charge and for non-commercial purposes. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University Hospital in Essen.

Summary Points

- Electronic medical information systems are now used in almost all healthcare facilities, replacing handwritten medical reports. Thus, structured diagnoses, clinical symptoms, lab values, radiological examinations, molecular biomarkers, and many other treatment-related data are electronically recorded and managed.
- In this research, a natural language processing (NLP)-based algorithm for entity recognition with UMLS concept mapping for the German language was developed.
- The algorithm in this approach for processing German clinical notes of colorectal cancer for the retrospective evaluation of therapy recommendations achieved a precision value of 96.64% for tumor status detection and 97.95% for diagnosis recognition with recall values of 94.89% and 99.54%, respectively.
- Using this approach for processing narrative clinical notes on colorectal cancer, the algorithm achieved an average precision value across all analyzed elements (11 concept groups) of 82.05% with a recall value of 82.45% and an F1-score of 81.81%, respectively.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijmedinf.2019.04.022

## References

[1] Robert Koch Institut, Krebs in Deutschland 2007–2008, Gesellschaft der epidemiologischen Krebsregister in Deutschland, Häufigkeiten und Trends: Darm, 8. auflage, (2012), pp. 36–39.
[2] Robert Koch Institut, Bericht zum Krebsgeschehen in Deutschland November 2016, Berlin, ISBN: 978-3-89606-279-6 (2016), pp. 28–31, https://doi.org/10.17886/rkipubl-2016-014.
[3] C. Wittekind, H.J. Meyer, TNM Classification of Malignant Tumours, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2010, pp. 94–100 ISBN-10: 3527327592.
[4] O. Majek, A. Gondos, L. Jansen, et al., Survival from colorectal cancer in Germany in the early 21st century, Br. J. Cancer 106 (2012) 1875–1880, https://doi.org/10.1038/bjc.2012.189.
[5] Leitlinie Kolorektales Karzinom 2013, http://www.awmf.org/leitlinien/detail/ll/021-007OL.html, (Last Accessed: 23 February 2019).

[6] H.J. Schmoll, D. Aderka, E. Van Cutsem, et al., ESMO consensus guidelines for management of patients with colon and rectal cancer: a personalized approach to clinical decision making, Ann. Oncol. 23 (2012) 2479–2516, https://doi.org/10.1093/annonc/mdw235.

[7] A. Winter, R. Haux, E. Ammenwerth, B. Birgl, N. Hellrung, F. Jahn, Hospital Information Systems, Health Information Systems: Architectures and Strategies (Health Informatics), Springer, 2010, pp. 33–36.

[8] B.T. McInnes, T. Pedersen, J. Carlis, Using UMLS Concept Unique Identifiers (CUIs) for word sense disambiguation in the biomedical domain, AMIA Annual Symposium Proceedings, (2007), p. 533.

[9] F. Xie, J. Lee, C.E. Munoz-Plaza, E.E. Hahn, W. Chen, Application of text information extraction system for real-time cancer case identification in an integrated healthcare organization, J. Pathol. Inform. 8 (2017) 48, https://doi.org/10.4103/jpi.jpi_55_17.

[10] A. Névéol, H. Dalianis, S. Velupillai, G.P. Zweigenbaum, Clinical natural language processing in languages other than English: opportunities and challenges, J. Biomed. Semant. 9 (2018) 12, https://doi.org/10.1186/s13326-018-0179-8.

[11] L. Cheng, J. Zheng, G. Savova, B. Erickson, Discerning tumor status from unstructured MRI reports-completeness of information in existing reports and utility of automated natural language processing, J. Digit. Imaging 23 (2) (2010) 119–132, https://doi.org/10.1007/s10278-009-9215-7.

[12] A. Pham, A. Névéol, T. Lavergne, D. Yasunaga, O. Clément, G. Meyer, R. Morello, A. Burgun, Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings, BMC Bioinformatics 15 (2014) 266, https://doi.org/10.1186/1471-2105-15-266.

[13] U. Hahn, M. Romacker, S. Schultz, medSynDiKATe—a natural language system for the extraction of medical information from findings reports, Int. J. Med. Inform. 67 (2002) 63–74.

[14] C. Weissenberger, S. Jonassen, J. Beranek-Chiu, M. Neumann, D. Müller, S. Bartelt, S. Schulz, J. Mönting, K. Henne, G. Gitsch, G. Witucki, Breast cancer: patient information needs reflected in English and German web sites, Br. J. Cancer 91 (8) (2004) 1482–1487.

[15] S. Schulz, J. Ingenerf, S. Thun, P. Daumke, German-language content in biomedical vocabularies, Proc CLEF 2013 Evaluation Labs and Workshop – CLEF-ER, (2013).

[16] C. Bretschneider, S. Zillner, M. Hammon, Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach, Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (2013) 27–35.

[17] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning - Data Mining, Inference, and Prediction, second edition, Springer, 2009, https://doi.org/10.1007/978-0-387-84858-7 ISBN 978-0-387-84858-7.

[18] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques (The Morgan Kaufmann Series in Data Management Systems), ISBN-10: 0123748569, auflage: 3, Morgan Kaufmann, 2011.

[19] Leslie H. Sobin, Mary K. Gospodarowicz, Christian Wittekind: TNM Classification of Malignant Tumours, John Wiley & Sons, 2011 31.08.2011, ISBN 978-1-4443-5896-4.

[20] Unified Medical Language System Terminology Services, A service of the U.S. National Library of Medicine | National Institutes of Health, https://uts.nlm.nih.gov/home.html, (Last Accessed: 23 February2019).

[21] brat rapid annotation tool, http://brat.nlplab.org/, (Last Accessed: 23 February 2019).

[22] M. Becker, B. Böckmann, Extraction of UMLS®;concepts using apache cTAKESTM for German language, Stud. Health Technol. Inform. (2016) 71–76.

[23] NegEx - A Python module to implement Wendy Chapman's NegEx algorithm, https://github.com/mongoose54/negex/tree/master/negex.python, (Last Accessed 23 February 2019).

[24] 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 (2001) 301–310.

[25] J. Franke, G. Nakhaeizadeh, I. Renz, XML retrieval and information extraction, Text Mining – Theoretical Aspects and Applications, Physica-Verlag, Germany, 2003, pp. 29–32.

[26] S. Bloehdorn, P. Haase, Z. Huang, Y. Sure, J. Völker, F. van Harmelen, R. Studer, Ontology management, Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies, 1st edition, Springer, Berlin Heidelberg, 2009, pp. 3–20.

[27] Y. Zhou, Y. Tong, R. Gu, H. Gall, Combining text mining and data mining for bug report classification, J. Software: Evol. Process 28 (2016) 150–176, https://doi.org/10.1002/smr.1770.

[28] N. Gonçalves, E. Oja, R. Vigário, Medical document mining combining image exploration and text characterization, in: S. Džeroski, P. Panov, D. Kocev, L. Todorovski (Eds.), Discovery Science. DS 2014. Lecture Notes in Computer Science, vol. 8777, Springer, Cham, 2014.

[29] R. Nicolas, A. Fornells, E. Golobardes, G. Corral, S. Puig, J. Malvehy, DERMA: a melanoma diagnosis plattform based on collaborative multilabel analog reasoning, Sci. World J. 2014 (2014), https://doi.org/10.1155/2014/351518.

[30] G. Divita, Q. Zeng, A. Gundlapalli, S. Duvall, J. Nebeker, M. Samore, Sophia: a expedient UMLS concept extraction annotator, J. Am. Med. Inform. Assoc. (2014) 467–476.

[31] R. Lenz, R. Blaser, et al., IT support for clinical pathways—lessons learned, Stud. Health Technol. Inform. 124 (2006) 645–650.

[32] F. Reginald, Information Retrieval. SuchmodelleUnd Data-mining-Verfahren Für Textsammlungen Und Das Web, Dpunkt; Auflage: 1., (01.03.2003), ISBN-10: 3898642135 (2003), p. 112.

[33] I. Spasić, J. Livsey, J.A. Keane, G. Nenadić, Text mining of cancer-related information: review of current status and future directions, Int. J. Med. Inform. 83 (9) (2014) 605–623, https://doi.org/10.1016/j.ijmedinf.2014.06.009.

[34] H. Dalianis, Clinical Text Mining: Secondary Use of Electronic Patient Records, Springer Open, Cham Switzerland, 2018, p. 124, https://doi.org/10.1007/978-3-319-78503-5.

[35] I.A. McCowan, D.C. Moore, A.N. Nguyen, R.V. Bowman, B.E. Clarke, E.E. Duhig, M.-J. Fry, Collection of cancer stage data by classifying free-text medical reports, J. Am. Med. Inf. Assoc. (JAMIA) 14 (2007) 736–745.

[36] A.N. Nguyen, M.J. Lawley, D.P. Hansen, R.V. Bowman, B.E. Clarke, E.E. Duhig, S. Colquist, Symbolic rule-based classification of lung cancer stages from free-text pathology reports, J. Am. Med. Inf. Assoc. (JAMIA) 17 (2010) 440–445.

[37] D. Martinez, Y. Li, Information extraction from pathology reports in a hospital setting, CIKM' 11 Proceedings of the 20th ACM International Conference on Information and Knowledge Management (2011) 1877–1882, https://doi.org/10.1145/2063576.2063846.

[38] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, P.C. de Groen, Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model, J. Biomed. Inform. 42 (2009) 937–949.

[39] A.M. Currie, T. Fricke, A. Gawne, R. Johnston, J. Liu, B. Stein, Automated extraction of free-text from pathology reports, AMIA Annu. Symp. Proc. 2006 (2006) 899.

[40] R. Weegar, J.F. Nygård, H. Dalianis, Efficient encoding of pathology reports using natural language processing, proceedings of the international conference recent advances in natural language processing, RANLP (2017) 778–783, https://doi.org/10.26615/978-954-452-049-6_100.

[41] R.H. Baud, A.M. Rassinoux, P. Ruch, C. Lovis, J.R. Scherrer, The power and limits of a rule-based morpho-semantic parser, Proc. AMIA Symp. (1999) 22–26.

[42] V. Cotik, R. Roller, F. Xu, H. Uszkoreit, K. Budde, D. Schmidt, Negation detection in clinical reports written in German, Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), Held in Conjunction with Coling, (2016), pp. 115–124.

[43] S. Wu, T. Miller, J. Masanz, M. Coarr, S. Halgrim, D.S. Carrell, C. Clark, Negation's not solved: generalizability versus optimizability in clinical natural language processing, PLoS One 13 (November (11)) (2014), https://doi.org/10.1371/journal.pone.0112774.