OPEN ACCESS

Research Article

Comparative evaluation of automated information extraction from pathology reports in three German cancer registries

Vergleichende Analyse automatisierter Informationsextraktion aus Pathologieberichten in drei deutschen Krebsregistern

Abstract

Feeding cancer registries with data extracted from textual reports, while maintaining a high level of data quality, has always been a labour-intensive task, due to the heterogeneity of the sources. The support of this task by IT solutions is expected to accelerate and optimise this process. To this end, the commercial text mining system Averbis Health Discovery was tailored to extract information from free text at the cancer registry of the federal state of Baden-Württemberg. The following entity types were extracted from German-language pathology reports: tumour localisation and morphology, pTNM, grading, (sentinel) nodes examined and affected, laterality and R-class. According to the entity type, several machine learning approaches as well as rules were used for the tumour types breast, prostate, colorectal and skin. Whereas for the pilot site, F values ranged between 0.800 and 0.996, values dropped when applying the extraction pipeline to two new sites (cancer registries Rhineland-Palatinate and Lower Saxony), for morphology from 0.950 to 0.657 and 0.933, and for localisation (topography) from 0.902 to 0.675 and 0.768. There was much less difference with R-class and lymph node counts. A thorough error analysis revealed numerous issues that explain these differences, such as different workflows between the sites, disagreements between textual and coded content as well as different handlings of missing values.

Keywords: cancer registries, text mining

Zusammenfassung

Das Anreichern von Krebsregistern mit Daten aus medizinischen Texten, bei gleichzeitiger Sicherstellung der Datenqualität, ist aufgrund der Heterogenität dieser Quellen mit erheblichem Aufwand verbunden. Die Unterstützung dieser Aufgabe durch IT-Lösungen soll diesen Prozess beschleunigen und optimieren. Zu diesem Zweck wurde das kommerzielle Text-Mining-System Averbis Health Discovery darauf zugeschnitten, Freitextinformationen für das Krebsregister Baden-Württemberg zu verwerten. Die folgenden Entitätstypen wurden aus Befundberichten der Pathologie extrahiert: Tumorlokalisation und -morphologie, pTNM, Grading, untersuchte und betroffene (Sentinel-) Lymphknoten, Lateralität und R-Klassifikation. Je nach Tumortyp wurden verschiedene Ansätze des maschinellen Lernens sowie Regeln für die Tumorentitäten Brust, Prostata, Kolon/Rektum und Haut verwendet. Während für den Pilotstandort die F-Werte zwischen 0,804 und 0,996 lagen, fielen die Werte, wenn die Extraktionspipeline auf zwei neue Standorte (Krebsregister Rheinland-Pfalz und Niedersachsen) angewendet wurde, für die Morphologie von 0,950 auf 0,657 und 0,933 und für die Lokalisierung (Topographie) von 0,902 auf 0,675 und 0,768. Es gab viel weniger Unterschiede in der R-Klassifikation und bei den Lymphknoten. Eine

Stefan Schulz[1]
Sonja Fix[1]
Peter Klügl[1]
Tamira Bachmayer[2]
Tobias Hartz[3]
Martin Richter[2]
Nils Herm-Stapelberg[4]
Philipp Daumke[1]

1 Averbis GmbH, Freiburg, Germany

2 Klinische Landesregisterstelle des Krebsregisters Baden-Württemberg, Stuttgart, Germany

3 Klinisches Krebsregister Niedersachsen, Hannover, Germany

4 Krebsregister Rheinland-Pfalz, Mainz, Germany

gründliche Fehleranalyse verwies auf zahlreiche Probleme, die diese Unterschiede erklären, z.B. unterschiedliche Arbeitsabläufe zwischen den Standorten, Widersprüche zwischen Text und codiertem Inhalt sowie unterschiedlicher Umgang mit fehlenden Werten.

Schlüsselwörter: Krebsregister, Text mining

1 Introduction

1.1 Medical registries

Medical registries are databases fed by uniform data about a particular health condition. Their purposes encompass monitoring and improvement of quality of care, clinical and epidemiological research as well as questions related to health economics. Cancer registries constitute the most important type of registries. According to the European Network of Cancer Registries (ENCR), there are two major aims of population-based cancer registries, viz. (i) the collection of data on new cancer cases in a defined geographic region with the purpose to investigate the burden of specific malignant disorders, and (ii) to provide a basis for investigating cancer aetiology and outcome (incidence, prevalence or survival). Thus, cancer registries are important means to assess the impact and effectiveness of health interventions, both regarding the provision of diagnostic and therapeutic measures, and the implementation of preventive actions by public policies.

1.2 Cancer registries in Germany

In Germany, cancer registries support treatment and follow-up care [1]. They are divided into clinical cancer registries, which describe hospital populations, and population-based (aka epidemiological) ones, which track cancer cases across institutions. Cancer registries are organised at the level of the 16 German federal states. The cases underlying this study cover the federal states of Baden-Württemberg (BW), Rhineland-Palatinate (RP) and Lower Saxony (LS). Table 1 provides some basic information about these three institutions and the population they support.

1.2.1 Datasets

The three registries receive notifications by hospitals, clinics and doctor's offices. They exchange data with civil registration offices in order to facilitate record linkage and notifications of deaths. Their dependence on manual annotation of cancer notifications requires detailed workflows and annotation guidelines. Input data are mainly narratives, with varying degrees of embedded codes (ICD-10, ICD-O, TNM). Some cancer centres at university hospitals produce well-curated structured datasets for their internal quality assurance processes, which, as a side effect, constitute a valuable input for the state-level registries. Target datasets vary between tumour types but there is a common set of fields and value restrictions that are equally valid for the majority of tumour types: ICD-10 – in its German modification ICD-10-GM – is used for disease coding, enhanced by ICD-O for morphology and localisation (also known as topography) of primary tumours and metastases. The TNM system is used for primary tumour size (T), regional lymph nodes (N) and distant metastases; tumour grading is expressed by a score between G1 (low grade) and G4 (very high grade), including Gx (undetermined grade). Datasets also include lymph node count values, distinguishing between examined and affected ones, for all lymph nodes and for sentinel lymph nodes in particular. For all tumour localisations, laterality values are given (left, right, bilateral, midline, not applicable). Finally, the UICC R classification denotes absence or presence of residual tumours after treatment.

1.2.2 Workflows

In the three centres, two parallel workflows can be distinguished. The first workflow starts with the submission of structured and appropriately formatted hospital data to the registry's data management team, whose work is therefore limited to minor data curation. In contrast, the second workflow requires manual annotations. It is this workflow where our effort is centred and for which results are described in this paper. It starts with the reception of narrative pathology reports (see Table 1), which have to be read by experts, who then assign codes and values, which requires time and considerable intellectual efforts. As long as both sources describe the same patients, sources are merged and "best of" registry entries are created, according to the interpretation by the registry experts. Evaluations have shown that quality improves significantly when pathology report annotations are added to the data from the hospitals [2], [3].
However, there are still a number of bottlenecks, particularly regarding quality issues at the source. Legal notification requirements in Germany are recent (2009 for hospitals, 2011 for doctor's offices and labs, e.g. pathology), and therefore software that is in current use still lacks maturity, with some modules in beta status.
Process quality at the registries is also affected by a lack of qualified staff and by a constantly changing environment with frequent updates of standard operating procedures and coding guidelines. The creation of "best of" adjudications is complex and still error-prone, due to the lack of standardisation of textual input.

Table 1: Synopsis of cancer registries in three German federal states

	Baden-Württemberg (BW)	Rhineland-Palatinate (RP)	Lower Saxony (LS)
Population covered	11,000,000	4,100,000	8,000,000
Total notifications/year (2019)	1,126.617	331,046	428,478
Yearly notifications of newly diagnosed cases (2019)	147,788	87,074	29,134
Yearly notifications by pathology labs (2019)	340,672	44,947	65,659
Number of reporting pathology labs (2020)	56	18	40

1.2.3 Challenges

Whenever structured data is manually entered into forms, after reading and analysing free text reports, quality issues are unavoidable and need to be monitored. Despite considerable efforts regarding quality, such measures have not yet been implemented in routine by the three cancer registries. Ideally, double annotations would allow the monitoring of inter-annotator agreement (e.g. by Kappa statistics) and subsequent adjudication of diverging annotations. This would allow ongoing quality assessment at annotator level, which then triggers training and coaching activities, as well as the ongoing refinement of annotation guidelines. Inter-annotator agreement values are also important reference data for assessing the outcome of machine annotations. If language models are trained by inconsistently annotated data, performance measures achieved by information extraction performed by a text mining system cannot be expected higher than the threshold of human agreement.
Due to resource constraints in the cancer registries, this study did not allow to measure interrater agreement, which can be seen as a limitation for the evaluation of the text mining tools we describe in the following sections. However, existing inter-annotator agreement studies for cancer registry coding in a German context have resulted in 76.7% for ICD-O tumour localisation, 80.3% for ICD-O morphology, 98.5% for ICD-O behaviour and 73.3% for ICD-10, measured by Fleiss' Kappa and considering full agreement across all hierarchical levels [4].

1.3 Purpose of the study

Given the high workload required by the manual analysis of pathology texts, the purpose of this study is to assess how human language technologies, powered by state-of-the-art artificial intelligence methodologies can support this task. Such support can be capitalised by the following scenarios:

- fully automated filling of registries by identifying relevant text passages in pathology reports,
- support of manual data input by tentative pre-filling of fields,
- automated clustering by tumour type (e.g. breast, lung, colorectal, prostate),
- keyword highlighting to accelerate manual annotation,
- content-based filtering, e.g. filtering by T4-tumours.

In this paper we will assess the quality of machine-based analysis of pathology texts against a gold standard, i.e. to which extent it fits the requirements of the registries. The question which of the above scenarios can be supported would then be subject of a future implementation study.

2 Background

2.1 Human language technologies applied to medical texts

Natural language processing (NLP), particularly information extraction [5], together with terminology and ontology [6], [7] have been awarded increasing attention in healthcare and biomedical research due to the predominance of free text in electronic health records, opposed to the need for structured and standardised data, e.g. for observational research, health statistics, disease reporting, quality assurance and billing. Whereas the field of biomedical terminology systems is huge, tumour documentation has internationally converged to a relatively small set of terminological (quasi-)standards, with ICD-10 for macroscopic aspects, ICD-O for morphology and localisation, and TNM (version 7/8) for staging. Depending on the tumour type, additional scoring systems and clinical stage classifications are added, such as the Gleason score for prostate tumours, or the Clark and Breslow staging systems for skin tumours.
NLP methods that automatically assign such codes have shown considerable progress, which, however, critically depends on available resources like annotated corpora and language-specific vocabularies. Apart from English, all natural languages suffer from a lack of these resources, which explains the underuse of NLP methods despite increasing demand. For German, important medical terminology resources lack localisation (e.g. SNOMED CT [8] and the NCI thesaurus [9]). Another important reason for the underuse of NLP on clinical content is the extreme brevity of clinical texts, characterised by acronyms and abbreviations, which require disambiguation efforts, unless under a very limited scope with a specific vocabulary.
Due to data privacy and therefore the lack of public availability of clinical texts – in contradistinction to scientific text corpora like the millions of MEDLINE abstracts and PubMedCentral full texts, clinical language processing

Table 2: Document sets and properties used for this study

Cancer registry	Document set	Period based on registration date	Mean Tokens per document	Median Tokens per document	Min. Tokens per document	Max. Tokens per document
BW	46,381	Jan 2015 – Nov 2019	440.0	341	26	6,605
RP	26,894	Jan 2015 – Dec 2019	582.4	480	141	6,843
LS	19,170	July 2018 – Aug 2019	568.9	507	7	5,279

has mostly occurred inside closed information environments, with amounts of data orders of magnitude below the size of openly available corpora. This constitutes a limit to the use of recent approaches from the field of machine learning – particularly deep learning [5], which are currently revolutionising the way human language is processed by computers. Wherever sufficiently large (annotated) amounts of data are available to train specific neural networks (NNs), models trained accordingly have outperformed traditional machine learning methods. Nevertheless, there is also evidence that manually-crafted rules for specific extraction tasks can achieve better results [10]. These findings should be taken into account when processing clinical texts. In addition, manual annotation of clinical texts is labour-intensive, so that the limits of what is possible are quickly reached with supervised learning methods (i.e. those depending on human annotation), given the amount of annotations necessary to achieve the desired outcome.

3 Material and methods

3.1 Textual data

Our analysis is focused on pathology reports for four types of cancer (breast, prostate, colon/rectum, skin). These documents are usually sent to the cancer registries by pathology labs as PDF files and do not follow any standardised template. All documents are in German. Several documents may refer to the same patient, but do not necessarily arrive in the order they have been created.
An important aspect is also that the source documents used in this study are not representative for all pathology reports produced in each state. This is due to the data curation activities done by university hospitals (workflow 1 in 1.2.2). As a consequence, their source documents do not reach the cancer registries, which receive structured data instead.
Table 2 shows the free-text document sets that were provided by the cancer registries. One document corresponds to one pathology report.

3.2 Structured data

All structured data used for this study were strictly related to datasets corresponding to exactly one document. According to the workflow, the dataset belongs either to a tumour diagnosis record or a tumour progress record. The fields of the datasets as given in Table 3 were used for the comparison with the automatic text analysis in our study.

Table 3: Fields of structured datasets used for comparison with automated text analysis

Field	Description
Tumour localisation	ICD-O localisation code
Morphology	ICD-O morphology code
Diagnosis	ICD-10-GM diagnosis code
pT	Primary tumour
pN	Regional lymph nodes
pM	Distant metastases
Grading	Grading
Nodes exam.	Number of examined lymph nodes
Nodes aff.	Number of affected lymph nodes
Sentinel Nodes exam.	Number of examined sentinel lymph nodes
Sentinel Nodes aff.	Number of affected sentinel lymph nodes
Laterality	Laterality
R-Class	R classification

The following applies to cancer registry data sets: d_{all} documents and structured r_{all} records ($|d_{all}| = |r_{all}|$). In order to guarantee representativeness, the selection is done according to the following criteria:

- coverage of a predefined period,
- cancer types: breast, prostate, colon/rectum, skin,
- all documents annotated in this period (or a random sample thereof),
- only documents that were annotated or are scheduled for being annotated.

The following data were excluded for being considered out of scope for this study:

- structured data that were not extracted from a document by the cancer registries (including structured data that arrive at the registries together with a document, but where no data extraction needs to be done at the registries),
- structured data for which no document was available,
- data related to any cancer type other than breast, prostate, colorectal or skin,
- the number of documents per patient is not taken into account, because the study is on documents, not on patients.

The rationale for this study design is that we exclusively scrutinise the assigning of predefined codes or values to narrative pathology reports. The hypothesis to be tested is whether an NLP-based text mining approach extracts

information items in a quality comparable to human experts.

3.2.1 Development and test set

The textual and structural data provided by the cancer registry of Baden-Württemberg (BW) is utilized for the initial customization of a commercial text mining software. The overall dataset is divided into a development set with 44,324 reports and a test set with 2,057 reports. The development set is used for the engineering, adaptation and optimization of the rule-based components as well as for the training of the machine learning models. The BW test set is only applied for evaluation.
The remaining two datasets of RP (Rhineland-Palatinate state, 26,894 reports) and LS (Lower Saxony state, 19,170 reports) are only utilized for evaluating the existing text mining system and thus provide a suitable experimental setting to investigate its generalizability.

3.3 Tools

The text mining technology used is AHD (Averbis Health Discovery) [11] by Averbis GmbH [12]. It contains over 50 different text mining annotators, e.g. for the recognition of diagnostic statements, medical procedures, lab values, drugs, anatomy, morphology, scores and others. Available for several languages, including English and German, AHD bundles annotators in predefined text mining pipelines tailored to document types like discharge summaries or pathology reports. AHD has been successfully used for various use cases, e.g. for data driven patient recruitment for clinical trials [13], automated coding and billing [14], documentation support in private practices [15], rare disease identification [16], antibiotic resistance monitoring [17], radiology report analysis [18], and health data de-identification [19]. Health Discovery is based on the Unstructured Information Management Architecture (UIMA) [20], an extensible framework for text analysis and text processing. It stresses the interoperability of components (analysis engines), which communicate in a pipeline by adding or modifying the meta information stored in the Common Analysis System (CAS), which contains the currently processed document. This information is represented by typed feature structures and indexed for efficient access. The most common type of a feature structure is the annotation, which assigns its type and additional features to a span of text. Most analysis engines create new annotations or modify existing ones in order to represent the result of their analysis.

3.4 Experimental approach

3.4.1 Pathology pipeline

AHD's pathology pipeline is designed for extracting information from pathology reports and addresses challenges typical for this text genre, which often contains sensitive patient data or other protected health information. This explains why there are no public datasets for applying machine-learning-based approaches. Information needed for pathology coding may not be explicitly present in the document, but needs to be derived from different, interrelated information. Finally, coding rules may differ from cancer type to another.
To address these challenges, the pathology pipeline combines different approaches and sources to deliver high-quality coding results such as

- terminology concept mapping (ICD-10, ICD-O...) and wordlists,
- rule-based annotations,
- machine learning,
- algorithms for combining the approaches.

The AHD pathology pipeline is composed of several general purpose text mining components (e.g. for disambiguation) as well as specialized ones for clinical documents. On top of this, a specialized component addresses cancer-specific coding on document level.
In this section, we first describe the structure of the pathology pipeline concerning the more general components and then the component for cancer coding in more detail. Table 4 describes the pipeline, with its components, which are sequentially processed. Figure 1 shows a screenshot of the interface of the pathology pipeline within the AHD text mining system.
The pipeline components implement methodologies ranging from rules and dictionaries to different kinds of machine learning approaches. Terminology-based annotators for tumour morphology and localisation apply dictionary lookup with linguistic pre-processing, abbreviation resolving and other extensions. Machine learning includes shallow approaches like Maximum entropy models, Conditional Random Fields and Support Vector Machines on the one hand and deep learning approaches like Convolutional Neural Networks on the other hand. Many components rely on manually engineered rules, which are still the method of choice where training data are scarce or annotations too expensive. The rules are created using UIMA Ruta [21], which provides a mixture of declarative and imperative paradigms to specify patterns of arbitrary annotations and their consequences. Finally, some components simply are composed of specific algorithms implemented in a programming language.

3.4.2 Pathology document classification

Our document classification component for cancer registry documents assigns the 13 output fields (see Table 3) and is composed of multiple subcomponents. For an initial categorization, support vector machines (SVMs) classify the pathology report into one of the supported cancer types, i.e. breast, prostate, skin, rectum/colon. This information feeds downstream processing steps for cancer-specific fall-back values or coding rules.
Morphology and tumour localisation are then separately classified using convolutional neural networks (CNNs), independently of the cancer type identified. If no valid

Table 4: Pathology pipeline

Component	Description
Language detection	Identification of the language (German, English) is required for the selection of dictionaries and machine learning models, e.g., for part-of-speech tagging
Linguistic pre-processing	Different components providing annotations and information for downstream processing steps, e.g. tokenization, sentence splitting, abbreviation resolution, stopword and invariant tagging, stemming, decompounding, part-of-speech tagging, shallow parsing, temporal expression detection, numeric value parsing
Sectionizing	Segmentation of the document into different sections
Laterality	Identification of anatomic laterality mentions (right, left, bilateral) and their context
Patient information	Patient-specific information like birth, death and admission/discharge dates
TNM	Extraction of TNM notations, mentions of grading, lymph nodes and similar entities
LabValues	Combines the annotation of measurements with dictionaries like LOINC in order to extract laboratory values and similar text entities
Diagnoses	Identification of diagnostic statements and their codes according to ICD-10
Localisation	Identification of tumour localisation (topography) mentions based on ICD-O, additional parsing of breast quadrant mentions
Morphology	Identification of tumour morphology mentions based on ICD-O
GleasonScore	Detection of Gleason scores for prostate cancer
Enumerations	Parsing of different styles of enumerations
Negation detection	Identification of negated text passages, also concerning enumerations
Additional oncology components	Extraction of text entities and relations like tumour staging, different kinds of receptors and specimens, e.g. breast cancer receptor status (-,+,++,+++, Measurement, unit "%") for estrogen, progesterone and HER2/neu; tumour thickness for skin cancer.
Disambiguation	Additional consolidation of text entities to which more than one code can be assigned
Pathology document classification	The actual components performing the cancer coding

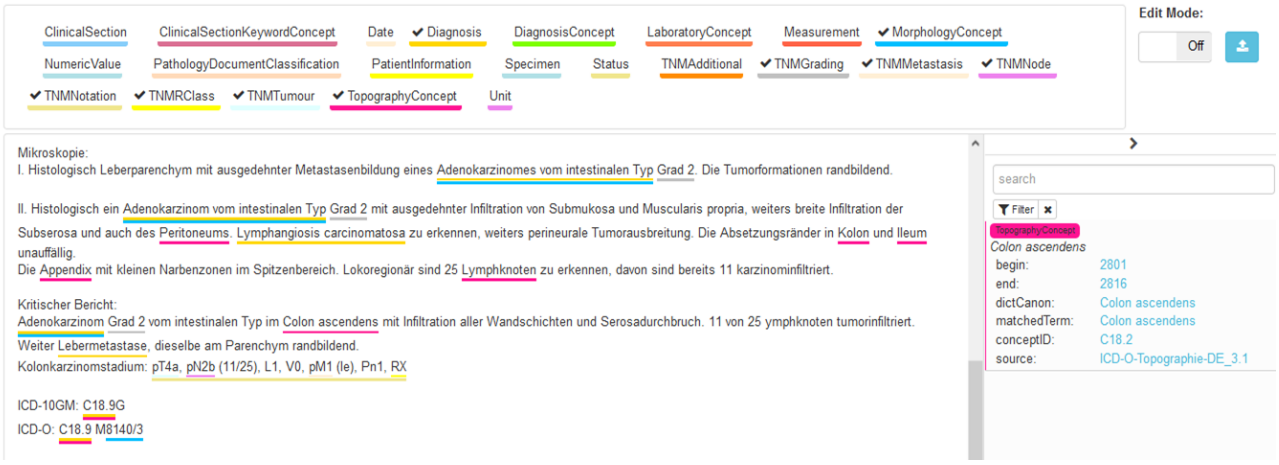

Figure 1: Annotations of the pathology pipeline, displayed in the AHD annotation editor

output is identified above a given threshold, different fall-back procedures are applied for both tumour morphology and localisation: First, concepts for either type that have been annotated by dictionary lookup are ranked using SVMs, and the code with the highest confidence in its context is selected for further processing. If no concepts have been extracted, a fall-back value depending on the cancer type is chosen. The final ICD-O codes are then assigned given a predefined mapping to ICD-10-GM diagnosis codes.

The remaining entity types like tumour, node, metastasis, grading, laterality and lymph nodes are selected using only the pre-processing functionality of the AHD pipeline. Specialized annotators detect potentially all related text mentions. For multiple mentions, either the highest value (e.g., grading) is selected or a majority voting is applied (e.g., laterality).

Finally, a set of additional rules perform a final modification of the values by using all given information. This post processing adapts the coding to the cancer registry coding guideline and induces values not directly represented in the pathology report. One example is the rule that prostate grading is not provided by TNM but by the Gleason score, or that laterality only matters for breast cancer (the other tumour sites are not bilateral), but that for both prostate and colorectal cancer the guideline requires the laterality value set to "T". Figure 2 depicts the workflow.

For machine learning models, SVMs and CNNs are used, supported by the following formative evaluations on the BW dataset. The first ones, in particular multiple one-vs-

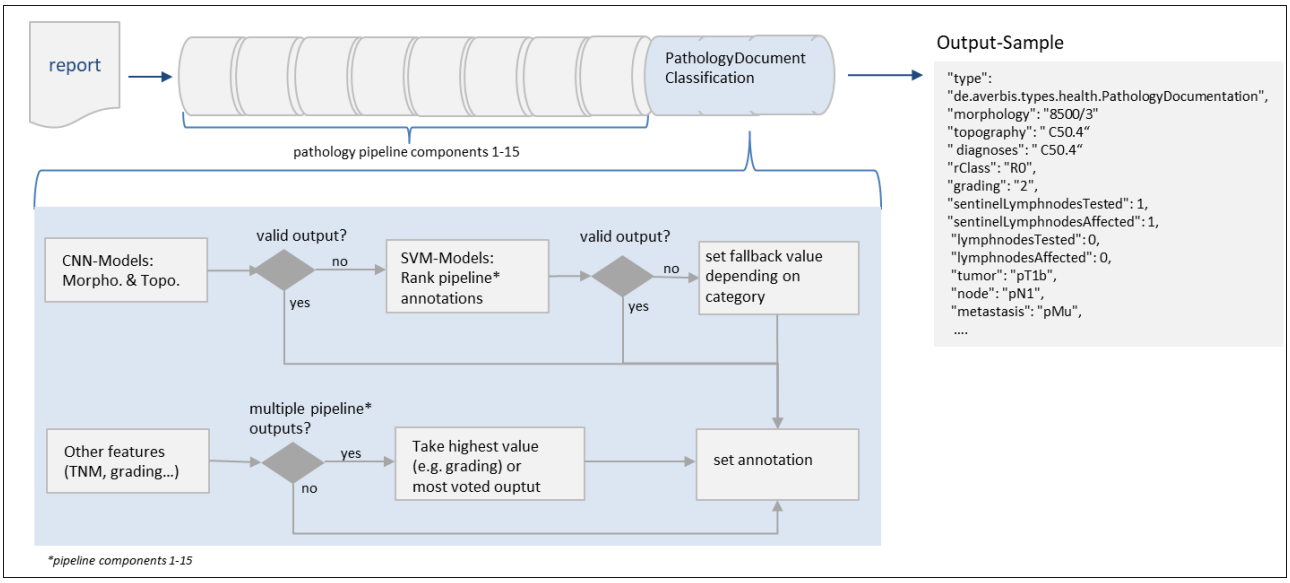

Figure 2: Approach to determine the final output fields of the pathology pipeline

Table 5: Results of a five-fold cross evaluation for cancer type classification using SVMs

Class	Count	F1	Precision	Recall	True Pos.	False Pos.	False Neg.
Melanoma	15,953	0.9993	0.9996	0.9991	15,938	6	15
Colorectal	6,870	0.9969	0.9954	0.9984	6,859	32	11
Breast	3,868	0.9984	0.9984	0.9984	3,862	6	6
Prostate	11,189	0.9991	0.9986	0.9986	11,173	4	16

all linear SVMs are applied for the initial categorization of the cancer type by single-label multi-class document classification. The features only consist of bag-of-stems and are weighted using logarithmic frequencies with redundancy [22] with L2-normalization. The classification model for the pipeline was trained using a subset of the BW training set (37,880 reports). A separate five-fold cross evaluation on this collection highlights the applicability of the model for the given task. The results in Table 5 indicate that classical ("shallow") machine learning is sufficient for this task.

A "deep" neural ML approach, viz. CNNs are applied to the classification of tumour morphology and localisation via single-label multi-class document classification. The network architecture is based on Kim et al. [23] using TensorFlow [24]. As an input space, fastText word embeddings with subword information provide a good trade-off between classification accuracy and prediction speed. The embeddings are tuned on the given data and classification task during the training of the model. The hyperparameters of the CNNs are optimized using BOHB [30], a combination of Bayesian Optimization and Hyperband. The models are trained on the complete train set of BW (d_{train}). Evaluation results (F1 score) on the test set comparing the performance to a baseline support vector machine (SVM) are depicted in Table 6. This clearly supports our option for using CNNs for this task.

Table 6: Comparison between SVM and CNN on tumour localisation and morphology classification

Feature	Support Vector Machine (SVM)	Convolutional Neural Network (CNN)
Tumour localisation	0.786	0.907
Morphology	0.925	0.966

4 Results

Text mining quality was evaluated for the features of Table 4 by comparing the automatic recognition with the manual coding results of the cancer registries. For each feature, the F1 score (a harmonic mean of precision and recall) was calculated. It was assumed that the manual coding results were correct and could therefore be considered the gold standard. However, we will see that this assumption is relativized by a detailed error analysis, as an important aspect of the results we report.

4.1 Overview

Figure 3 shows the evaluation results of the three cancer registries across all features. For 9 of 13 features, the F1 scores are highest for the BW state.

In the following, the results for each of the three cancer registries (BW, RP, LS) are explained and interpreted in detail.

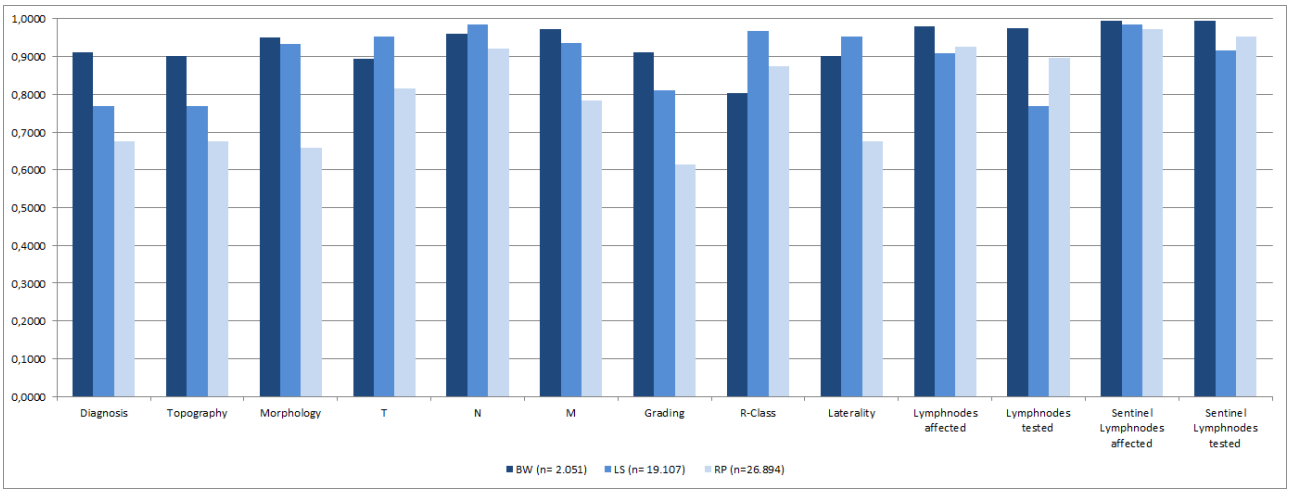

Figure 3: Overview of evaluation results. Dark blue BW (Baden Württemberg state), medium blue LS (Lower Saxony), light blue RP (Rhineland Palatinate)

4.2 Evaluation Baden-Württemberg state (BW)

4.2.1 Results

The evaluation of the BW cancer registry was performed on the test dataset n=2,057. The results show an F1 score greater than 0.8 for all features, using the manual coding results as ground truth (Table 7). The lymph node status features are distinguished by particularly high F1 score values with a match of up to 99.6%. The lowest value is found for the R-Class feature.

Table 7: Evaluation results by feature for the BW dataset. The diagnosis values (ICD-10-GM) are derived from the tumour localisation and morphology ICD-O codes.

Feature	F1 score
Diagnosis	0.9103
Localisation	0.9020
Morphology	0.9503
Tumour	0.8942
Node	0.9590
Metastasis	0.9737
Grading	0.9113
R-Classification	0.8040
Laterality	0.9010
Nodes affected	0.9790
Nodes tested	0.9761
Sentinel Nodes affected	0.9956
Sentinel Nodes tested	0.9941

4.2.2 Gap analysis

For assessing gaps in the pathology feature extraction, the top 10 mismatches of the above mentioned test data set (n=2,057) were reviewed by experts at the feature level (Table 8). Since “diagnosis” is derived from tumour morphology and localisation, it is not included in this analysis.

Table 8: Top 10 deviations (BW dataset)

ID	Count	Feature	Gold standard	Text mining
1	149	rClass	R0	null
2	146	rClass	R1	null
3	95	laterality	L	T
4	84	laterality	R	T
5	55	grading	T	3
6	55	tumour	pT0-pT3x (specific value)	pTu
7	34	tumour	pTu	pT0-pT3x (specific value)
8	33	metastasis	pMT	pMu
9	32	tumour	pTT	pTu
10	32	tumour	cT1	pT1c

The deviations can be summarized in six categories (Table 9).

Table 9: Deviations by categories (BW dataset)

Category	Issue	Count	Interpretation	ID
1	Text mining not detecting the R-class information	295	This is almost exclusively observed with skin tumours. While for other cancer types the R-classification value is usually given in the text, for skin tumours it often has to be inferred from the context. Text mining has not yet defined rules for this.	#1, #2
2	Text mining assigns "T" for "does not apply" while coders specified laterality	179	This, again, affects only skin tumours and can be explained by the fall-back value on "T" during automatic detection. While laterality recognition was implemented for breast tumours, it is still pending for skin tumours. Prostate and intestine are not paired; here a fall-back to "T" is correct.	#3, #4
3	Text mining assigns the grading value to "3" while coder assigned "T"	55	Almost exclusively in colorectal cancers, text mining misinterprets "high-grade intraepithelial neoplasia" which is seen in intestinal adenomas as grading information for the overall tumour. This can be easily fixed with a specific rule.	#5
4	TNM: Text mining assigns "u" for "unknown" while coder assigned "T" for "does not apply"	65	Not a text mining error. This information is not in the text and can therefore not be extracted. Default coding rules need to be harmonized with fall-back values of text extraction.	#8, #9
5	TNM: Text mining assigns "p" for pathology, gold standard assigned "c" for "clinical"	32	Not a text mining error. "c" for "clinical" is not in the text and can therefore not be extracted. The pathology pipeline automatically assigns "p" for TNM values (if not already given in the text) as the findings are extracted from pathology reports.	#10
6	Tumour: Text mining finds a specific value for tumour, gold standard assigned "u" for "unknown" and vice versa	98	For BW, a rule was implemented that predicts whether the report is an initial or progress report based on cues such as "biopsy" or "resection". In the former case, extracted TNM values are overwritten with "u". Samples have shown that this distinction between initial and progress reports is not always correct, which leads to these "vice-versa"-deviations.	#6, #7

4.3 Evaluation Rhineland-Palatinate state (RP)

4.3.1 Results

The evaluation of the RP dataset shows F1 score values between 0.6153 (Grading) and 0.9725 (Sentinel Nodes affected), cf. Table 10. In particular, the scores for tumour morphology and localisation, as well as ICD-10 diagnosis are up to 0.29 lower when compared with the BW results.

Table 10: Evaluation results by feature (RP dataset). The diagnosis values (ICD-10-GM) are derived from the tumour localisation and morphology ICD-O codes.

Feature	F1 score
Diagnosis	0.6745
Tumour localisation	0.6753
Morphology	0.6573
Tumour	0.8162
Node	0.9206
Metastasis	0.7834
Grading	0.6153
R-Classification	0.8740
Laterality	0.6758
Nodes affected	0.9268
Nodes tested	0.8972
Sentinel Nodes affected	0.9725
Sentinel Nodes tested	0.9530

4.3.2 Gap analysis

For a variance analysis in the pathology feature extraction, the top 10 mismatches of the above-mentioned data set (n=26,894) were reviewed by experts at the feature level (Table 11). Where appropriate, sources of error were aggregated. Again, "diagnosis" is not considered in the gap analysis as this feature is derived from the morphology and topography codes.

Table 11: Top 10 deviations (RP dataset)

ID	Count	Feature	Gold standard	Text mining
1	3,263	metastasis	CM0	pMu
2	2,631	grading	U	T
3	2,399	laterality	L (count=1,207)	T
4	2,070	localisation	R (count=1,192)	More specific C50-codes, e.g. C50.4, C50.8, C50.2, ...
5	1,772	localisation	C50.9	C50.9
6	1,671	grading	More specific C50-codes,	T
7	1,632	laterality	e.g. C50.4, C50.8, C50.2, ...	L (count=880), R (count=752)
8	1,368	tumour	empty	pT1c
9	1,137	morphology	empty	8097/3
10	824	morphology	CT1c	8500/3

The deviations can be summarized in eight categories (Table 12).

Table 12: Deviations by categories (RP dataset)

Category	Issue	Count	Interpretation	ID
1	Metastasis: rules apply "u" although specific value is given in the text	3,263	For the BW data, a rule predicts whether the report is an initial or progress report based on cues such as "biopsy" or "resection". In the former case, extracted TNM values are overwritten with "u". It must be checked whether the rule also applies to the coding procedure in RP.	#1
2	Grading: text mining assigns "T" while coder assigned "u" or left the field empty	4,302	Not a text mining error, as this information is not given in the text. Samples suggest that the gold standard is not consistent in assigning grading values such as "unknown" (u), "not applicable" (T), "not determinable" (X), which is also indicated by the empty fields.	#2, #6
3	Laterality: coder assigned a laterality value (left or right), while text mining assigns "T"	2,399	Same root cause as described for the BW data, Table 9, Category 2, only affecting skin tumours, explainable by the fall-back value on "T" during automatic detection. Laterality recognition for skin tumours is not yet implemented.	#3
4	Laterality: text mining extracts a laterality value, coder left the field empty	1,632	This concerns reports of breast tumours only: random sample inspection of documents showed that laterality information was given in the pathology report and correctly annotated by text mining. The gold standard seems incomplete regarding this feature.	#7
5	Text mining annotates a specific tumour localisation code for breast cancer, while coder assigned C50.9 ("not further specified") and vice versa.	3,842	The fact that the deviations occur vice-versa indicate gold standard inconsistencies. C50.9 should only be coded if there is no further information about the location of the tumour. The inspection of text samples has shown that in most cases the localization is described in detail, whereas the pathologist erroneously assigns C50.9. Thus, the machine learning system is trained on incorrect data. A hybrid approach that includes rules for greatest localisation recognition could minimise this problem.	#4, #5
6	Morphology: Text mining annotates 8500/3 (Invasive ductal carcinoma, NOS) while gold standard assigned 8140/3 (Adenocarcinoma, NOS)	824	Again a deviation that concerns breast tumours only. An inspection of pathology reports has shown that pathologists tend to disagree regarding these two morphology codes. This is also evident in the gold standards and affects the predictions of text mining. With more than 70% of breast tumours being invasive ductal [28], it should be further investigated whether the prediction of text mining is more precise than the Gold Standard.	#8
7	Tumour: Text mining assigns "p" for pathology, gold standard assigned "c" for "clinical"	1,368	Not a text mining error, problem analogous to the one described for BW (#10). "c" for "clinical" is not present in the texts and can therefore not be extracted. The pathology pipeline automatically sets a "p" for TNM values (if not already given in the text), because all findings are extracted from pathology reports.	#9
8	Morphology: Text mining annotates 8097/3 (Basal cell carcinoma, NOS) contrasting with 8090/3 (Nodular basal cell carcinoma) in the gold standard	1,137	Text sample analysis has shown that in various reports the morphology code 8090/3 was assigned by the pathologist and accordingly coded by text mining. In some of these cases text contained references to a nodular carcinoma, which may explain that the coder chose 8097/3. In case of conflicting information about a code, the automatic coding follows the decision of the machine learning algorithm.	#10

4.4 Evaluation Lower Saxony state (LS)

4.4.1 Results

The evaluation of the LS dataset shows F-values between 0.768 (Nodes tested) and 0.9836 (Sentinel Nodes affected), cf. Table 13. Similar to RP data, particularly for tumour morphology, localisation and diagnosis, values are significantly lower compared to BW (up to 0.141), although not to the same extent as in the RP dataset.

4.4.2 Gap analysis

Again, the top 10 mismatches of the above-mentioned data set (n=19,170) were reviewed by experts at the feature level (Table 14). Where appropriate, sources of error were aggregated. The feature "diagnosis" is not considered in the gap analysis as it is derived from the morphology and topography codes.
The deviations can be summarized in six categories (Table 15).

Table 13: Evaluation results by feature for the LS dataset. The diagnosis values (ICD-10-GM) are derived from the tumour localisation and morphology ICD-O codes.

Feature	F1 score
Diagnosis	0.7695
Localisation	0.7684
Morphology	0.9332
Tumour	0.9522
Node	0.9838
Metastasis	0.9350
Grading	0.8117
R-Classification	0.9665
Laterality	0.9524
Nodes affected	0.9088
Nodes tested	0.7680
Sentinel Nodes affected	0.9836
Sentinel Nodes tested	0.9162

Table 14: Top 10 deviations (LS dataset)

ID	Count	Feature	Gold standard	Text mining
1	4,326	NodesTested	*empty*	*Numeric value*
2	1,672	NodesAffected	*empty*	*Numeric value*
3	1,577	SentinelNodesTested	*empty*	*Numeric value*
4	1,134	Localisation – Breast	C50.9	More specific C50-codes, e.g. C50.4, C50.8, C50.2, ...
5	1,150	Grading	U	G1-G3
6	1,055	Metastasis	pMx	*specific pM-value*
7	968	Grading	U	T
8	786	Localisation – Breast	C50-codes, e.g. C50.4, C50.8, C50.2, ...	C50.9
9	314	laterality	U	T
10	218	Localisation	C18.9	C18.7

Table 15: Deviations by categories (LS dataset)

Category	Issue	Count	Interpretation	ID
1	Node status and sentinel node status was empty in the gold standard, while text mining finds a numeric value	7,575	Not a text mining error. Sample inspection showed that this information was available in the texts, hence the gold standard is incomplete.	*#1, #2, #3*
2	Text mining assigns a more specific breast tumour localisation, while coders set C50.9 (“not further specified”) and vice versa.	1,920	Same root cause as described for RP, Table 12, Category 5: Bidirectional disagreement reveals inconsistencies in the gold standard.	*#4, #8*
3	Grading: Text mining finds specific grading value, while gold standard says “u” for “unknown”	1,150	Samples have shown that in most cases (7 of 8) the information was present in the texts and correctly recognized by text mining. There are isolated cases where text mining mistakenly recognizes regression grade as grading.	*#5*
4	Grading: Text mining sets “T” while manual coder set “U”	1,282	This deviation only affects skin tumours and is explained by the fall-back value on “T” during automatic detection, in absence of a grading or laterality value. Thus it is not a text mining error; rather coding rules need to be aligned.	#3, #7
5	Metastasis: Manual coder set “pMx” while text mining sets “pM0”	960	In the reports often text patterns were found like “clinical M0”, “under the assumption M0”, “cM0” which was recognized by text mining and returned with "pM0". Since clinical assessments are not taken into account in pathology coding, text recognition must be adapted accordingly.	*#6*
6	Tumour localisation – Colon Text mining sets “C18.7” (sigmoid) while manual coders set “C18.9” (not further specified)	218	Samples show that in most cases, there is no indication in the texts for a location of the carcinoma in the sigma. Since statistically half of all colon cancers occur in the sigmoid [27], we assumed a bias in the train data set and the localisation classifiers trained on this datasets inherits this bias.	*#10*

5 Discussion

Figure 3 shows for most of the features that the BW dataset produced the best result, followed by LS and RP. This is not surprising because BW is the first cancer registry that is already using AHD in their daily business. In several iterations, all models and rules had been tailored to these data. Data also show that the results of the features extracted by rules were more robust compared to those based on models trained on the BW dataset. Tumour localisation and morphology F-values suffered severe drops up to 30%. Only the rule-based grading and laterality values showed a similar behaviour.

The BW error analysis shows that automatic highly accurate text recognition, with deviations from the manual coding results are within the range of interrater reliability. Most deviations do not result from text mining errors, but from the application of coding rules by the coding staff. Some of these rules can easily be reengineered using rule-based NLP methods.

In the LS and RP datasets, many errors could also be traced back to a lack of compliance with coding rules, or slightly different local preferences. In addition, it was evident that the diversity of language in pathology reports prevented that ML procedures performed equally well. Inconsistencies in the gold standards, especially for morphology, reinforce this. In addition, differences in fall-back values (which are assigned in case of missing data) explain disagreements (T, U, X are not always used congruently). Nevertheless, we have to admit that the results for ICD-O morphology and localisation codes are suboptimal. Larger and more representative training corpora will be required, as well as a better implementation of quality mechanisms and rules that manage borderline cases. Just to cite one example: the use of the residual category C50.9 (unspecified location of breast cancer) should raise suspicion, since it is hardly plausible that this important information is missing in the original data. Grading, node and metastasis status are other frequent sources of gold standard errors. To quantify the influence of these factors and, as a consequence, to assess the overall performance of the AHD pipeline would, however, require inter-coder agreement values for each dataset. Given the absence of such data, the only source of information was the qualitative inspection, which nevertheless provided strong evidence that

1. the current gold standard has its shortcomings, so that the information extracted by AHD was often correct whilst the gold standard was wrong,
2. a limited set of adjustments should significantly improve the text mining performance, and that
3. difficult boundary issues between tumour localisation and morphology codes, with known disagreements between human coders will always set a limit to the F-values to be expected.

In summary, it is noticeable that the ML methods trained on the BW dataset perform better on the LS than on the RP dataset, and for the feature “morphology” even comparably well as for BW. A large part of the deviations is due to missing information in the gold standard or diverging fall-back values. If the evaluation is adjusted for these factors, there are good reasons to assume that the actual recognition quality in all features is already 90% and higher.

The institutional contexts and workflows make it difficult to compare our results to others’. A very recent work [25] compared multitask learning with single-task learning using CNNs and achieved better results for the multitask approach with up to 60% correctly classified cases according to morphology, localisation, laterality, grade and behaviour, together. Taking ICD-O tumour localisation and morphology, alone, micro-averaged F1 scores amounted to 0.915 and 0.776, respectively, using single-task CNNs. Multitask CNNs (applied to all the five tasks) increased the results by about three percentage points. Multitask CNN results cannot be compared to our study, where CNNs were only used for tumour localisation and morphology. More important, however, is the fact that Alawad et al. [25] used only three-character codes for tumour localisation and four character codes for morphology. Thus, they used a restricted set of about 130 classes, which eliminated many of the boundary issues we discussed. No inter-coder agreement values were provided.

CNNs also showed the best performance in a study on identification of ICD-O localisation codes for breast and lung cancer. With a small set of 12 different 4-character codes [26], with F1 results between 0.722 and 0.811.

6 Conclusion and outlook

From a technical perspective, the results of this study demonstrate that a hybrid solution that includes machine learning, terminology mapping and rules produces the best results. Whereas challenges such as clinical coding itself are not explicitly addressed, we could demonstrate that the several pieces of information to be extracted had to be interpreted in combination, so that a layered, pipeline-based approach turned out to be the most appropriate.

Technical challenges included different and non-standardised spelling variants and notations, particularly mentions of location and laterality, as well as the correct identification of information items and the way they were related, e.g. locations with tumour types.

The work was done under field conditions, which means that the need for high quality reference datasets – a fundamental requirement for effective training of machine learning models – had to be satisfied by routine data that had grown from practice over long periods, during which annotation rules have changed, many coding staff were involved, and strict quality control measures were only partly put in place. Although no systematic inter-coder agreement data were available from this data, other studies have shown major disagreements between coders for both ICD-O tumour morphology and localisation.

Pathologists usually show little interest in ICD-O coding and structured data collection in general, as long as no financial or scientific incentives are given. This explains conflicting information between textual and semi-structured representations in the same pathology dataset. Fine-grained and coarse-grained information coexists, e.g., a breast quadrant specified in the text along with the "unspecified" C50.9 code, or a precise textual description of a tumour morphology along with the unspecific code "adenocarcinoma or the like". Trained with this kind of data, even the best possible ML system will underperform and good ML decisions will be classified as wrong. Language models and rules tailored to one cancer registry will continue underperforming at another cancer registry as long as coding rules and heuristics (e.g. how to deal with missing data) are not the same. Harmonisation at a national level is a desideratum, regardless of the use of manual, automated or hybrid methods. Lack of harmonisation affects data comparability. Nevertheless, we can state that text mining has proven a very suitable approach to support resource-intensive manual coding and to contribute to quality control and coding standardisation. To this end, text mining tools are of high heuristic value, as they bring quality gaps and inconsistent coding rules to the surface.

Along these lines, a next step would then consist in iterative cycles that combine text mining with manual inspection and disagreement analysis. The scope of model training and validation should be broadened by including data from other cancer registries and institutions. Once such a new workflow has proven its usefulness by increasing performance values, text mining routines can be adapted to new functionalities and additional tumour types.

Notes

Funding

Development, maintenance and customisation of Averbis Health Discovery have been co-funded by the German Ministry for Education and Research (BMBF) through the research networks MIRACUM (01ZZ1801C) and SMITH (01ZZ2003).

Competing interests

Philipp Daumke is CEO of Averbis GmbH, the company that develops the tool used for this study. Some of the authors (Sonja Fix, Peter Klügl and Stefan Schulz) are or have been employed by Averbis GmbH.

Ethics statement

The ethical foundations for collecting and analysing patient-specific tumour data is covered by the state cancer registry laws of the German federal states that participated in the studies. Averbis' work was carried out as part of a data processing contract with the three cancer registries. Only de-identified texts were used for the study.

References

1. Altmann U, Garbe C, Hofstädter F, Katalinic A. Standortbestimmung: Klinische und epidemiologische Krebsregister in Deutschland. Forum. 2007;22:28-9.
2. Piñeros M, Parkin DM, Ward K, Chokunonga E, Ervik M, Farrugia H, Gospodarowicz M, O'Sullivan B, Soerjomataram I, Swaminathan R, Znaor A, Bray F, Brierley J. Essential TNM: a registry tool to reduce gaps in cancer staging information. Lancet Oncol. 2019 Feb;20(2):e103-e111. DOI: 10.1016/S1470-2045(18)30897-0
3. Tyczynski JE, Démaret E, Maxwell Parkin D, editors. Standards and Guidelines for Cancer Registration in Europe. The ENCR Recommendations Volume 1. Lyon: IARC Press; 2003. (IARC Technical Publication; 40).
4. Lappe L, Ressing M, Plachky P, Blettner M, Zeissig SR. Comparison of coding diagnosis, localisation and histology via ICD-10 and ICD-O-3 between coders and a gold standard. In: 2018 ENCR Scientific Meeting; 2018 Sep 26-28; Copenhagen, Denmark.
5. Hahn U, Oleynik M. Medical Information Extraction in the Age of Deep Learning. Yearb Med Inform. 2020 Aug;29(1):208-20. DOI: 10.1055/s-0040-1702001
6. Schulz S, Daumke P, Romacker M, López-García P. Representing oncology in datasets: Standard or custom biomedical terminology? Informatics in Medicine Unlocked. 2019; 15:100186. DOI: 10.1016/j.imu.2019.100186
7. Schulz S, Jansen L. Formal ontologies in biomedical knowledge representation. Yearb Med Inform. 2013;8:132-46. DOI: 10.1055/s-0038-1638845
8. Millar J. The Need for a Global Language - SNOMED CT Introduction. Stud Health Technol Inform. 2016;225:683-5.
9. Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization of the NCI thesaurus. Comp Funct Genomics. 2004;5(8):648-54. DOI: 10.1002/cfg.445
10. Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc. 2019 Nov;26(11):1247-54. DOI: 10.1093/jamia/ocz149
11. Pokora RM, Le Cornet L, Daumke P, Mildenberger P, Zeeb H, Blettner M. Validierung von semantischen Analysen von unstrukturierten medizinischen Daten für Forschungszwecke Validation of Semantic Analyses of Unstructured Medical Data for Research Purposes. Gesundheitswesen. 2020 Mar;82(S 02):S158-S164. DOI: 10.1055/a-1007-8540
12. Averbis. Health discovery. [last accessed 2020 Nov 14]. Available from: https://averbis.com/health-discovery/
13. TriNetX. Natural Language Processing. Extract Clinical Facts from Physician Notes and Clinical Reports. [last accessed 2020 Nov 14]. Available from: https://www.trinetx.com/nlp/
14. Meta It. Metakis. [last accessed 2020 Nov 14]. Available from: https://metait.de/metakis
15. CompuGroup Medical Deutschland AG. Ärzte erfassen 28 Stunden pro Woche Arztbriefe – Innovatives Werkzeug spart Zeit dank Volltextsuche und Textanalyse. [last accessed 2020 Nov 14]. Available from: https://www.cgm.com/de/ueber_uns_de/news_de/presse_de/presse_details_92416.de.jsp

16. Catarino C, Grandjean A, Doss S, Mücke M, Tunc S, Schmidt K, Schmidt J, Young P, Bäumer T, Kornblum C, Endres M, Daumke P, Klopstock T, Schoser B. mineRARE: Semantic text-mining of electronic medical records as diagnostic decision support tool to search for rare neurologic diseases such as Pompe disease, Fabry disease and Niemann-Pick type C disease. Eur J Neurol. 2017;24:75

17. Schober D, Choquet R, Depraetere K, Enders F, Daumke P, Jaulent MC, Teodoro D, Pasche E, Lovis C, Boeker M. DebugIT: Ontology-mediated Layered Data Integration for Real-time Antibiotics Resistance Surveillance. In: Paschke A, Burger A, Romano P, Marshall MS, Splendiani A, editors. Proceedings of the 7th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2014); 2014 Dec 9-11; Berlin, Germany. Available from: http://ceur-ws.org/Vol-1320/paper_22.pdf

18. Daumke P, Simon K, Paetzold J, Marwede D, Kotter E. Data-Mining in radiologischen Befundtexten. Rofo. 2010;182 - WS117_3. DOI: 10.1055/s-0030-1252462

19. Seuss H, Dankerl P, Ihle M, Grandjean A, Hammon R, Kaestle N, Fasching PA, Maier C, Christoph J, Sedlmayr M, Uder M, Cavallaro A, Hammon M. Semi-automatische Deidentifizierung von deutschsprachigen medizinischen Berichten mit vertraulichem Inhalt für Big Data Analysen Semi-automated De-identification of German Content Sensitive Reports for Big Data Analytics. Rofo. 2017 Jul;189(7):e1. DOI: 10.1055/s-0035-1567202

20. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng. 2004;10(3-4):327-48. DOI: 10.1017/S1351324904003523

21. Kluegl P, Toepfer M, Beck PD, Fette G, Puppe F. UIMA Ruta: Rapid development of rule-based information extraction applications. Nat Lang Eng. 2016;22(1):1-40. DOI: 10.1017/S1351324914000114

22. Leopold E, Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? Mach Learn. 2002;46:423–4. DOI: 10.1023/A:1012491419635

23. Kim Y. Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing; 2014 Oct 25-29; Doha, Quatar. p. 1746–51. DOI: 10.3115/v1/D14-1181

24. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Kudlur M. Tensorflow: A system for large-scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16); 2016 Nov 2–4; Savannah, GA, USA. p. 265-83.

25. Alawad M, Gao S, Qiu JX, Yoon HJ, Blair Christian J, Penberthy L, Mumphrey B, Wu XC, Coyle L, Tourassi G. Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. J Am Med Inform Assoc. 2020 Jan;27(1):89-98. DOI: 10.1093/jamia/ocz153

26. Qiu JX, Yoon HJ, Fearn PA, Tourassi GD. Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports. IEEE J Biomed Health Inform. 2018 Jan;22(1):244-51. DOI: 10.1109/JBHI.2017.2700722

27. Gump V, Henß H; Universitätsklinikum Freiburg. Klinisches Krebsregister. Kodierhilfe. Kolonkarzinom. 2014 [last accessed 2020 Nov 14]. Available from: https://www.uniklinik-freiburg.de/fileadmin/mediapool/09_zentren/cccf/pdf/cccf_kkr_kodierhilfe_kolontumor.pdf

28. Reiner A, et al. Pathologie des Mammakarzinoms. In: Smola MG; Arbeitsgemeinschaft für Chirurgische Onkologie der Österreichischen Gesellschaft für Chirurgische Onkologie, editors. Consensus-Bericht Mammakarzinom. Graz; 1993 [last accessed 2020 Nov 14]. Available from: http://www.aco-asso.at/publikationen/aco-asso-consensusberichte/consensus-bericht-mammakarzinom/4-pathologie-des-mammakarzinoms/

29. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Trans Assoc Comput Linguist. 2017;5:135–46. DOI: 10.1162/tacl_a_00051

30. Falkner S, Klein A, Hutter F. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10-15; Stockholm, Sweden. (Proceedings of Machine Learning Research; 80). p. 1437–46.

Corresponding author:
Stefan Schulz
Averbis GmbH, Salzstraße 15, 79098 Freiburg im Breisgau, Germany
stefan.schulz@averbis.com

Please cite as
Schulz S, Fix S, Klügl P, Bachmayer T, Hartz T, Richter M, Herm-Stapelberg N, Daumke P. Comparative evaluation of automated information extraction from pathology reports in three German cancer registries. GMS Med Inform Biom Epidemiol. 2021;17(1):Doc01. DOI: 10.3205/mibe000215, URN: urn:nbn:de:0183-mibe0002156

This article is freely available from
https://www.egms.de/en/journals/mibe/2021-17/mibe000215.shtml

Published: *2021-03-31*