GGPONC: A Corpus of German Medical Text with Rich Metadata Based on Clinical Practice Guidelines

Florian Borchert[1,4*], **Christina Lohr**[2,5*], **Luise Modersohn**[2,5*], **Thomas Langer**[3], **Markus Follmann**[3], **Jan Philipp Sachs**[1], **Udo Hahn**[2,5], **and Matthieu-P. Schapranow**[1,4]
[1]Digital Health Center, Hasso Plattner Institute, University of Potsdam, Germany
{firstname.lastname}@hpi.de
[2]Jena University Language & Information Engineering (JULIE) Lab
Friedrich Schiller University Jena, Germany
{firstname.lastname}@uni-jena.de
[3]German Guideline Program in Oncology, German Cancer Society, Berlin, Germany
{lastname}@krebsgesellschaft.de
[4]HIGHMED Consortium of the German Medical Informatics Initiative
[5]SMITH Consortium of the German Medical Informatics Initiative

Abstract

The lack of publicly accessible text corpora is a major obstacle for progress in natural language processing. For medical applications, unfortunately, all language communities other than English are low-resourced. In this work, we present GGPONC (German Guideline Program in Oncology NLP Corpus), a freely distributable German language corpus based on clinical practice guidelines for oncology. This corpus is one of the largest ever built from German medical documents. Unlike clinical documents, clinical guidelines do not contain any patient-related information and can therefore be used without data protection restrictions. Moreover, GGPONC is the first corpus for the German language covering diverse conditions in a large medical subfield and provides a variety of metadata, such as literature references and evidence levels. By applying and evaluating existing medical information extraction pipelines for German text, we are able to draw comparisons for the use of medical language to other corpora, medical and non-medical ones.

1 Introduction

The synthesis of validated experience in the form of Clinical Practice Guidelines (CPGs) serves as a basis for evidence-based decision making in clinical practice. To leverage the knowledge in CPGs for clinical decision support systems, e.g., for integration with electronic health records or automated evaluation of adherence to these guidelines, machine-readable versions of CPGs are necessary. However, CPGs today are disseminated mostly as free-text documents, with few formal elements. Thus, Natural Language Processing (NLP) might be helpful to automatically extract information from these unstructured texts and transform them into a structured, or even machine-executable, format. As CPGs are also specific to their country of origin, they are usually formulated in the respective native language, so NLP technology has to be adapted properly.

A major reason for the progress in NLP research in the past years is the public availability of large text corpora (and attached metadata). Yet, for documents originating from a clinical context, the protection of personal information is a major requirement imposed by legal privacy regulations. Some research initiatives, e.g., I2B2 (Uzuner et al., 2011), MIMIC-III (Johnson et al., 2016), the Shared Task of Social Media for Health (Weissenbacher et al., 2019), or CLEF EHEALTH (Goeuriot et al., 2020) make de-identified clinical text document collections available under the conditions of Data Use Agreements (DUA). Besides, databases of biomedical research articles like PUBMED provide an abundant amount of examples for medical language. However, with only few exceptions, such easily accesssible text corpora are hardly available for the German (Lohr et al., 2018) and other non-English languages. Today, there is no viable solution for sharing even de-identified clinical texts in Germany. In effect, not only are large-scaled research datasets missing but also pretrained language models for German medical language, such as an equivalent of BioBERT (Lee et al., 2020).

To address *(1)* the lack of available German medical text resources for NLP research, and *(2)* the need for machine-readable CPGs, we constructed a corpus based on a set of German CPGs for oncology. The *German Guideline Program in Oncology*

*Authors marked by * equally share first authorship.

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 38–48

https://doi.org/10.18653/v1/P17

(GGPO) (Follmann, 2020), operated by the Association of the Scientific Medical Societies in Germany, the German Cancer Society and the German Cancer Aid, is in a unique position to enable this research, as their guidelines are also provided via a mobile app (Seufferlein et al., 2019). Hence, this data set is already available in a semi-structured format with rich, formatted metadata, resulting in a much higher data quality than data extracted a posteriori from PDF versions of the guidelines.

An excerpt from the XML version of GGPONC is depicted in Listing 1. Other than clinical text relating to individual patients, we deal with scientific medical text that does not contain any privacy-sensitive data requiring de-identification. This way, we can provide access to GGPONC for other researchers via a DUA.[1]

```
<document>
 <section>
  <name>Risikofaktoren</name>
  <section>
   <name>Helicobacter pylori</name>
   <recommendation>
    <recommendation_creation_date
        value="2019-01-01"/>
    <recommendation_grade value="B"/>
    <!-- more metadata -->
    <text>Die H. pylori-Eradikation
        mit dem Ziel der Magenkarzinom
        -prävention sollte bei den
        folgenden Risikopersonen
        durchgeführt werden (siehe
        Tabelle unten).</text>
   </recommendation>
   <text>Das Magenkarzinom ist eine
       multifaktorielle Erkrankung,
       bei der die Infektion mit H.
       pylori den wichtigsten
       Risikofaktor darstellt. Seit
       1994 ist H. pylori durch die
       Weltgesund-heitsorganisation
       als Klasse I Karzinogen
       anerkannt und wurde 2009 als
       solches bestätigt <litref id="
       65327"/>.</text>
  </section> <!-- more sections -->
 </section>
</document> <!-- more documents -->
```

Listing 1: Text snippet from the XML version of the underlying GGPO corpus. Documents are structured into sections which can contain multiple recommendations. CPG recommendations can carry a multitude of metadata elements, as well as a concise text statement. Additional background text segments may contain more detailed information.

[1]For instructions on how to get access to the data, see: https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-english/

2 Related work

Due to legal data protection measures, German-language clinical text corpora are extremely rare and existing ones almost impossible to (re)use. Typically, accessibility is restricted to research staff only within the lifetime of a project and blocked for the outside world. In Table 1, we list, to the best of our knowledge, all existing German-language clinical research text corpora which have been described in scientific publications up until now. With only few exceptions, these corpora are small and mostly limited to a specific medical discipline or clinical division. In addition to these pure clinical documents, other document types are also interesting for the NLP community, e.g., CPGs, which are available for a wide range of conditions.

CPGs as a target for automated text analytics have been much less utilized compared to other scientific publications and clinical documents. Most of that work took place in the context of formalizing CPGs as computer-interpretable guidelines (Peleg, 2013). Bouffier and Poibeau (2007) describe an approach to fill in a semi-structured *Guideline Elements Model* template by segmenting unstructured guidelines using linguistic patterns. An evaluation was run on 18 French guidelines. Serban et al. (2007) describe the extraction and instantiation of linguistic templates for guideline formalization, evaluated on a Dutch guideline for breast cancer treatment. German CPGs were the focus of Becker and Böckmann (2017) who adapted APACHE CTAKES to detect German UMLS concepts and evaluated their approach on a single German breast cancer guideline. Zadrozny et al. (2017) outline a system which identifies contradictions and disagreements in English CPGs.

Some authors have focused on extracting more task-specific information, such as activities (Kaiser et al., 2010), process structures (Wenzina and Kaiser, 2013; Zhu et al., 2013; Hematialam and Zadrozny, 2017) or negation triggers (Gindl et al., 2008). Taboada et al. (2013) apply a pipeline of open-source tools for parsing CPGs, Named Entity Recognition (NER) tagging and relation extraction in a case study with 171 sentences from CPGs. Most of the aforementioned approaches work with relatively small annotated corpora and English language, only. Recently, Fazlic et al. (2019) use LSTMs and fuzzy rules to extract *"action takers"*, *"symptoms"*, *"actions"* and *"purposes"* from CPGs, recognize recommendations and predict the grade

Table 1: Overview of existing text corpora of German clinical language. For GGPONC, we report the number of guidelines with the number of their individual text segments in brackets.

Corpus / Data	Documents	Sentences	Tokens	Available
FRAMED: clinical reports and medical textbook snippets (Wermter and Hahn, 2004)	–	6k	100k	✗
Reports from five medical domains (Fette et al., 2012)	544	–	–	✗
Radiology reports (Bretschneider et al., 2013)	174	4k	28k	✗
Transthoracic echocardiography reports (Toepfer et al., 2015)	140	–	–	✗
Operative reports (surgery) (Lohr and Herms, 2016)	450	22k	266k	✗
Discharge summaries from a dermatology department (Kreuzthaler et al., 2016)	1,696	–	–	✗
Discharge summaries and clinical notes from nephrology domain (Roller et al., 2016)	1,725	28k	158k	✗
Discharge summaries and clinical notes from nephrology domain (Cotik et al., 2016)	183	2k	13k	✗
X-ray reports (Krebs et al., 2017)	3,000	–	–	✗
3000PA: internistic and ICU discharge summaries 3000PA JENA PART (Hahn et al., 2018)	≈ 3,000 1,006	– 170k	– 1,421k	✗ ✗
JSYNCC: case examples from medical textbooks (Lohr et al., 2018) (v1.1)	903	29k	368k	✓
Mixed-domain, -section, and -document type ASSESS CT corpus (Miñarro Giménez et al., 2019)	60 (400–600 chars each)	–	≈ 6k	✗
Discharge summaries with osteoporosis diagnosis (König et al., 2019)	1,982	–	2,001k	✗
Technical-Laymen Corpus: social media samples (Stomach-Intestines, Kidney) (Seiffe et al., 2020)	4,000	–	438k	✓
GGPONC – recommendations	25 (4,348)	7k	132k	✓
GGPONC – complete corpus	25 (8,418)	60k	1,340k	✓

of recommendation. The authors use a data set extracted from PDF versions of 45 guidelines with 1,020 recommendations.

Some larger corpora of CPGs for the English language exist already. Hussain et al. (2009) present the Yale Guideline Recommendation Corpus (YGRC), a sample of 1,275 guideline recommendations extracted from the *National Guideline Clearinghouse* (NGC). Their work revealed inconsistencies in writing style and reporting of the strength of recommendations. Using a subset of YGRC, El-Rab et al. (2017) present a rule-based approach to detect procedures and drug recommendations. Read et al. (2016) describe the CREST corpus, consisting of 4,029 recommendations from 170 guidelines annotated with their respective recommendation strength and report a total number of 8,138 types within the recommendations. Large corpora of CPGs lend themselves to mining the state-of-the-art knowledge in a medical subfield. For instance, Leung et al. (2015) identify comorbidities by analyzing pairs of co-occurring conditions, using a corpus of 268 NGC guideline summaries. Leung and Dumontier (2016) find drug-disease relations via named entity recognition using a corpus of 377 NGC guideline summaries. The relations are compared to structured drug product labels to assess their overlap.

In summary, our work is most similar to the CREST corpus (Read et al., 2016), in the sense that we provide a corpus based on CPGs consisting of medical text and metadata. However, while the number of recommendations in GGPONC is comparable to CREST, the amount of structured metadata and background text in our corpus is much larger (see Section 4.1). Also, our corpus contains German text, addressing a scenario where available resources are much scarcer (see Table 1). While

Table 2: Metadata elements of recommendations of GGPONC

Attribute	Description
Recommendation creation date	Date the recommendation was first introduced
Type of recommendation	Evidence-based or consensus-based statement or recommendation
Recommendation grade	A (strong recommendation) B (recommendation) 0 (weak recommendation / option)
Strength of consensus	Strong Consensus Consensus Approved by majority No consensus
Total vote in percentage	Percentage of approval among the expert committee
Literature references	List of evidence backing up the recommendation
Expert opinion	Yes or absent
Level of evidence	According to Oxford, SIGN, or GRADE
Edit state	State (checked, new, modified) & note regarding guideline updates

Becker and Böckmann (2017) also apply NLP to German CPGs, we consider a large superset of the CPGs used in their work and provide access to our data as a preprocessed and analyzed text corpus.

3 Methods

3.1 Data Collection

In order to assemble the corpus of German CPGs, we acquired semi-structured JSON versions of the guidelines from the REST API of the Content Management System (CMS) that serves the backend for the mobile app provided by the GGPO. The data was subsequently transformed from JSON to an XML format. We preserved the document structure (chapters and sections), as well as recommendation metadata and literature references. An example of the resulting XML format can be found in Listing 1. The metadata elements are described in Table 2. Literature references are included with an ID number which can be resolved to a citation in the provided literature index file.

The guidelines distinguish between recommendations and background texts; we preserved this distinction in the corpus. In general, the recommendations tend to be concise statements related to a particular clinical question. For evidence-based recommendations, literature references and evidence levels are included. The background texts provide the reasoning behind the recommendations and a summary of the evidence underlying the recommendations, again backed by literature references.

3.2 Automated Annotation

Besides the XML version of the corpus, we created plain text versions of all recommendations of background text parts to facilitate processing by existing NLP pipelines. For preprocessing, like sentence splitting and tokenization, we used JCoRe (Hahn et al., 2016) (i.e., UIMA-based) pipelines and FraMed (Wermter and Hahn, 2004) models which were developed for German clinical text.

We also employed the JuFit tool (v1.1) (Hellrich et al., 2015), a filter for UMLS to create a dictionary of all German words from the UMLS (Bodenreider, 2004) (version 2019AB)[2] and the Semantic Groups ANAT (Anatomical Structure), CHEM (Chemicals & Drugs), DEVI (Devices), DISO (Disorders), LIVB (Living Beings), PHYS (Physiology), and PROC (Procedures) (without advanced JuFit rules). We have chosen only these six out of the full set of 15 UMLS Semantic Groups because we used similar categories in the named entity recognition tasks and also wanted to avoid cognitive overloading of the human annotators.

Finally, we screened TNM expressions[3] which were extracted using a rule-based approach implemented with the Python library spaCy. This part was originally developed for German pathology reports in the context of the HiGHmed consortium of the Medical Informatics Initiative of Germany. TNM expressions and genes were specifically chosen for their relevance in cancer treatment.

[2] https://www.nlm.nih.gov/research/umls/

[3] The UICC TNM system is a classification scheme for malignant tumors, see https://www.uicc.org/resources/tnm

Table 3: Details of GGPONC. We report the number of text **Seg**ments (plain text files), **Rec**ommendations and literature **Ref**erences. The numbers of **Sent**ences, **Tokens** and **Types** refer to the pure textual content of the corpus, excluding any meta-data and headings. Annotated parts of the corpus are marked with * (see also Section 4.3).

	Guideline	**Seg.**	**Rec.**	**Sent.**	**Tokens**	**Types**	**Ref.**
1	Palliative medicine*	696	445	5,956	134,489	15,795	3,065
2	Lung cancer*	666	313	4,251	93,324	12,756	2,344
3	Breast cancer	685	362	4,127	93,128	12,660	2,824
4	Supportive therapy	823	337	4,224	90,711	12,411	2,401
5	Bladder cancer	355	225	3,872	85,299	11,347	2,521
6	Colorectal cancer*	569	290	3,176	71,416	9,644	2,580
7	Prostate cancer	307	221	3,090	67,900	9,418	2,119
8	Malignant melanoma	297	167	2,715	60,354	9,318	1,256
9	Prevention of skin cancer	288	119	2,354	55,965	9,140	952
10	Actinic keratosis and SCC of the skin*	199	74	2,590	54,073	6,861	1,278
11	Stomach cancer	246	142	2,328	50,836	8,156	1,670
12	Endometrial cancer	317	173	1,999	50,056	8,154	1,340
13	Cervical cancer*	341	115	2,168	49,422	8,164	1,127
14	Prevention of cervix cancer*	302	103	2,055	48,676	7,989	1,391
15	Renal cell cancer*	276	122	2,118	48,013	8,202	1,496
16	Testicular tumors	315	163	1,917	43,726	6,774	1,412
17	Oesophageal cancer*	172	91	1,611	35,710	6,680	1,026
18	Laryngeal cancer	189	118	1,525	35,519	6,841	681
19	Chronic lymphocytic leukemia (CLL)*	290	138	1,410	34,470	5,682	725
20	Hodgkin lymphoma*	253	167	1,489	31,876	5,245	889
21	Hepatocellular cancer (HCC)*	157	88	1,296	27,852	5,704	803
22	Malignant ovarian tumors	193	94	1,136	25,807	5,110	1,013
23	Psycho-oncology*	121	47	778	19,270	4,127	835
24	Pancreatic cancer	294	158	857	16,871	3,670	1,154
25	Oral cavity cancer*	111	76	630	15,438	3,376	1,026
	Annotated Part	**4,153**	**2,069**	**29,528**	**664,029**	**50,732**	**18,585**
	Full Corpus	**8,414**	**4,348**	**59,672**	**1,340,201**	**76,252**	**37,928**

4 Results

4.1 Corpus Characteristics

In total, 25 GPGs with 8,414 text segments were extracted from the CMS comprising the first version of GGPONC. In Table 3, we give an overview of the CPGs in terms of the number of tokens and types, as well as the number of literature references. We also report the total number of recommendations and background text segments, since they serve as the units of analysis for our automated annotation pipelines. The CPGs cover a wide range of indications and anatomical locations. They also differ significantly in their extent, e.g., there is much more text for broad topics, such as palliative medicine, or indications with many treatment options, such as lung cancer. Of the approximately 38k literature references in the corpus, around 20k are unique with roughly 9k explicit links to PUBMED. We provide bibliographic details on these references alongside the corpus to facilitate research on the relationships between CPGs and the underlying medical evidence.

Table 4 contains the automated named entity extraction results. Their quality and interpretation in comparison to other German (clinical and non-clinical) text corpora will be discussed in the next section. The whole corpus consists of:

- a single XML file, including the document structure and all mentioned metadata,
- a file for the complete literature index,
- individual plain text versions of the text segments, sentences, and tokens,
- automatically created entity annotations and a subset of manually corrected annotations in standoff format.

Table 4: Comparison of GGPONC with 3000PA (Jena part), JSYNCC, German PUBMED abstracts of case reports and two non-clinical corpora (German Wikipedia articles of wars (WIKIWARSDE) and news articles from the KRAUTS corpus)

	GGPONC		Clinical			Non-Clinical	
	Complete	**Recom.**	**3000PA-J**	**JSYNCC**	**PUBMED**	**WIKI**	**KRAUTS**
Documents	8,418	4,348	1106	903	336	22	142
Sentences	60k	7k	171k	29k	3k	5k	1k
Tokens	1,340k	132k	1,421k	368k	43k	96k	31k
Tokens / Sentence	22.5	19.0	8.8	12.5	16.5	20.9	25.3
UMLS* (%)	6.42	8.93	8.72	5.71	7.59	0.75	0.02
ANAT (%)	0.45	0.48	1.78	1.11	0.79	0.04	0.09
CHEM (%)	0.82	1.01	1.08	0.41	0.59	0.04	0.07
DEVI (%)	0.12	0.17	0.20	0.55	0.18	0.06	0.04
DISO (%)	1.42	2.02	2.96	1.21	2.80	0.08	0.13
LIVB (%)	1.07	1.32	0.38	0.35	0.82	0.38	0.37
PHYS (%)	0.37	0.43	0.76	0.60	0.50	0.12	0.10
PROC (%)	2.18	3.50	1.56	1.49	1.90	0.01	0.12
Genes (%)	1.28	1.41	2.21	0.87	0.97	0.94	0.55
TNM (%)	0.19	0.37	0.07	0.07	0.04	0.003	0
Stop words (%)	34.05	35.53	20.37	32.96	34.51	34.65	24.24

As CPGs are subject to a regular update cycle, we are able to automatically redo the data acquisition process in the future in order to provide a historical view on the guideline development.

4.2 Comparison with Other German Medical and Non-Medical Corpora

We analyze the characteristics of GGPONC by comparing the entity matches with three German medical text corpora, namely version 1.1 of the JSYNCC corpus (case examples from clinical text books) (Lohr et al., 2018), the Jena Part of the 3000PA corpus (1106 German discharge summaries) (Hahn et al., 2018),[4] and abstracts of German case reports from PUBMED. In addition, we compare the results to out-of-domain corpora, namely German WIKIPEDIA articles of wars (WIKIWARSDE) (Strötgen and Gertz, 2011) and news articles from the KRAUTS corpus (Strötgen et al., 2018). The results are depicted in Table 4.

The fraction of stop words is comparable across all medical text corpora, as is the fraction of tokens that map to UMLS concepts. As expected, the guideline recommendations contain more medical terms per token than the background text. Compared to the clinical corpora, the CPGs have more instances of the class *Living Beings*, as they often describe treatment recommendations for certain populations. Notably, the average sentence length is much greater in the clinical guidelines, and in particular in the background text, pointing at the more scientific style of writing prevalent in the guidelines as compared to clinical narratives. TNM expressions occur much more frequently in GGPONC, which can be attributed to its focus on the oncology domain.

Both out-of-domain corpora contain only small amounts of UMLS concepts (apart from the semantic class *Living Beings*), which indicates a high precision of our entity tagging approach. In Figure 1 we visualize the overlap of unique medical concepts from UMLS found in each of the corpora.

While there is a significant overlap between GGPONC and the clinical corpora, a major fraction of concepts is unique to each corpus. These results suggest that our corpus combined with other clinical text corpora can provide a more comprehensive view on the use of medical language, in general, than each of the corpora alone.

4.3 Evaluation of Annotation Results

The automatic annotations for a subset of the CPGs have been independently reviewed by human experts (4 students of medicine, all of them passed their first medical exam, supervised by a medical

[4]Based on the approval by the local ethics committee (4639-12/15) and the data protection officer of Jena University Hospital discharge summaries were extracted from the HIS of the Jena University Hospital and further transformed.

Table 5: Pair-wise average F1-score and standard deviation (σ) for instance and token based inter-annotator-agreement (IAA), precision and recall per entity class. Genes had to be excluded from IAA analysis due to the large number of false positives.

		IAA Instances		IAA Token		Precision	Recall
		avg. F-score	σ	avg. F-score	σ		
UMLS Anatomy	(ANAT)	.718	.122	.720	.125	.872	.571
UMLS Chemicals	(CHEM)	.839	.045	.850	.041	.917	.600
UMLS Devices	(DEVI)	.414	.366	.409	.368	.465	.209
UMLS Disorders	(DISO)	.747	.097	.773	.096	.919	.453
UMLS Living Being	(LIVB)	.848	.066	.846	.067	.985	.698
UMLS Physiology	(PHYS)	.534	.183	.576	.177	.607	.310
UMLS Procedures	(PROC)	.706	.099	.727	.100	.944	.506
TNM (rule based)		.820	.073	.749	.120	.965	.881
Overall w/o Genes		**.742**	**.094**	**.758**	**.094**	**.945**	**.528**
Genes		-	-	-	-	.022	.589

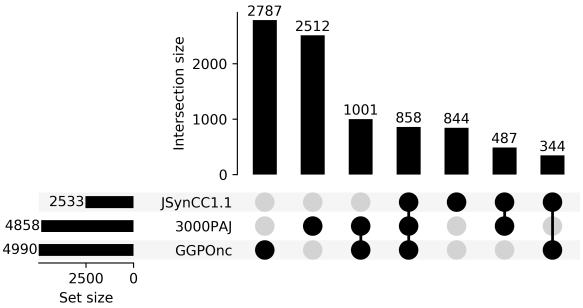

Figure 1: UpSet (Lex et al., 2014) visualization of the intersection of distinct UMLS concepts in JSynCC 1.1, 3000PA (Jena part) and GGPOnc. Each vertical bar indicates the size of one intersecting subset of UMLS terminology shared between the corpora, whereas the horizontal bars denote the total number of distinct UMLS concepts per corpus.

doctor) using the Brat annotation tool (Stenetorp et al., 2012). Due to restricted resources for manual annotation work, we decided to evaluate on a subset of 13 (full) guidelines (see Table 3), which amounts to half of the corpus. The CPGs were chosen such that they cover a diverse range of topics and percentages of token matches, with a rather high rate of around 6–8% results per token for HCC as opposed to a lower rate of roughly 5–6% for CLL and psycho-oncology. Additional guidelines were chosen for manual annotation based on project requirements.

We calculate the inter-annotator-agreement (IAA) using the pair-wise average F-score (Hripcsak and Rothschild, 2005) of instances and tokens. An instance is a single composite annotation unit which consists of one or more tokens, e.g., *"eingeschränkte Nierenfunktion"* (limited renal function) denotes an instance with two (German) tokens, *"eingeschränkte"* and *"Nierenfunktion"*.

The agreement subset consists of 20 text segments with the largest amount of automatic annotations for each of four guidelines (HCC, CLL, Pancreatic cancer, Psycho-oncology) and 20 random-sampled text segments, resulting in 40 agreement documents with a size of approx. 19k tokens annotated by all annotators. We excluded the gene category from the IAA analysis, due to an apparently large number of false positive pre-annotations. The IAA achieved an average F-score of 0.742 on instances and 0.758 on tokens. Furthermore, we calculated micro-averaged precision and recall values for the automated annotation results, using the complete set of manually reviewed annotations as gold standard. The results are depicted in Table 5.

In another annotation study of diagnoses, symptoms and findings on the Jena part of the 3000PA corpus, average F-score values converged in the range of around 0.7–0.8 for typical clinical entities as well, e.g., anatomy or disorders in comparison to diagnoses (approx. 0.7), also for pre-annotations. The low IAA value of *Physiology* is similar to the IAA of 0.5 on the symptoms category of the named study (Lohr et al., 2020). The UMLS category *Living Beings* contains a lot of information similar to personal health information. The average IAA value of around 0.9 is similar to average values of an annotation study for the anonymization of German discharge summaries (F-score > 0.95) (Kolditz et al., 2019).

5 Discussion & Limitations of this Study

While the initial results of the information extraction pipelines we employed are promising, there is much room for improvement. The extraction of genes suffers from a large number of false positives, as there are many common German words (e.g., *"gilt"*, *"dar"*) and three-letter-acronyms (e.g., *"CLL"*, *"HCC"*) with strings identical with gene names in our large dictionary (around 562k entries). Thus, supplying an improved gene tagger which balances German lexical noise with advanced capabilities of gene taggers for English texts will be a desideratum of future research.

The German UMLS has a number of issues, which severely affect our dictionary-based entity extraction pipelines. First and foremost, its vocabulary size is extremely limited. The English UMLS contains over 6.5M entries and the Spanish one around 750k, whereas there are only around 234k entries in the German version (3.6% coverage of the English version). Recently introduced drugs are missing in the UMLS *Chemistry* category, so a more up-to-date dictionary of drug names is also needed for future work. Moreover, the surface representation of German umlauts is notoriously inconsistent in UMLS, e.g., *"ä"* is sometimes transcribed as *"ae"* or even simplified as *"a"*, as in *"eingeschraenkte Nierenfunktion"*, which results in an increasing false negative rate. All of these factors contribute to rather low recall values, as shown in Table 5.

The accuracy of dictionary matches further decreases due to inconsistent handling of compounds throughout the corpus. For instance, *"Pankreaskarzinompatienten"* (patients with pancreas carcinoma) would not be detected as an entity, whereas hyphen-connected *"Pankreaskarzinom-Patienten"* would, yielding two entities (*Disorders* and *Living Beings*), respectively. In this case, we would choose to annotate the whole compound as *Living Beings* to avoid annotation on a subword level, which could be addressed using a more finely adapted tokenization algorithm. While precision and recall of the rule-based TNM extraction approach are high on GGPONC, one has to be careful as certain TNM expressions can cause context-dependent semantic ambiguities. For instance, *"V1"* and *"V2"* are valid TNM components referring to venous invasion, but are also detected in the WIKIWARSDE corpus referring in this context to German missiles from World War II.

6 Conclusion

We presented GGPONC, one of the currently largest corpora composed of German medical texts, assembled from the CPGs in oncology and equipped with rich structure information and metadata. We applied information extraction pipelines to extract a variety of named entity classes. Despite the limitations we discussed, the information extracted so far can be of immediate use to enable semantic search functionalities in the guideline app (Seufferlein et al., 2019), precision medicine search engines (Faessler et al., 2020) or in clinical decision support systems (Schapranow et al., 2015).

Our results indicate that GGPONC shares many characteristics with existing clinical text corpora. This can facilitate the development of machine learning-based NLP algorithms for German clinical text. Beam et al. (2020) suggest that combining corpora covering different parts of medical terminology can improve the utility of trained word embeddings. In addition to the German documents discussed in this work, some of the GGPO guidelines have an additional English version, which could be used to construct parallel corpora for research in multilingual medical NLP.

Extending our work to clinical guidelines from other medical specialities besides oncology will be a straightforward way to extend the volume of the corpus, provided that the document structures can be harmonized across medical societies. However, as most CPGs are distributed as PDF documents, extraction of the plain text content from these can result in quality issues not encountered in this work.

The structured metadata of the corpus provide ample opportunities for future research. For instance, the corpus can be used as a resource for evidence-based medicine summarization, as it contains mappings from literature references to recommendation statements and evidence levels. As we plan to create future versions of the corpus based on updated guideline versions, the extracted concepts can also be used to track changes in CPGs, like the emergence of new treatments and other changes in recommended clinical practice. We envision to combine information extracted from scientific articles, such as study reports, or clinical trial registers with information from CPGs to automatically detect whether these CPGs might be outdated given changes in the underlying evidence base.

In addition to the existing annotations for a wide selection of UMLS semantic types, we can easi-

ly extend the employed pipelines with different dictionaries, e.g., derived from other subsets of the German part of the UMLS, more comprehensive official lists of drug names, or the German version of the *International Classification of Diseases*.

We make GGPONC available for researchers under the conditions of a Data Use Agreement. For instructions on how to access the corpus and the human annotated data see: https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-english/. The code to reproduce our experiments is available at: https://doi.org/10.5281/zenodo.4067994.

Acknowledgments. This work was partially supported by the German Federal Ministry of Research and Education (BMBF) under grants (01ZZ1802H, 01ZZ1803G). We thank our annotators, André Scherag and Danny Ammon from the Jena University Hospital, and all colleagues from the HiGHMED and SMITH consortia for their constant support.

References

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2020. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Proceedings of the Pacific Symposium on Biocomputing 2020. Big Island, Hawaii, USA, January 3-7, 2020*, pages 295–306.

Matthias Becker and Britta Böckmann. 2017. Semi-automatic mark-up and UMLS annotation of clinical guidelines. In *MedInfo 2017 — Proc. of the 16th World Congress on Medical and Health Informatics. Hangzhou, China, 21-25 Aug. 2017*, pages 294–297.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Amanda Bouffier and Thierry Poibeau. 2007. Automatically restructuring practice guidelines using the GEM DTD. In *BioNLP 2007 — Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing @ ACL 2007. Prague, Czech Republic, June 29, 2007*, pages 113–120.

Claudia Bretschneider, Sonja Zillner, and Matthias Hammon. 2013. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. In *BioNLP 2013 — Proceedings of the 2013 Workshop on Biomedical Natural Language Processing @ ACL 2013. Sofia, Bulgaria, August 8, 2013*, pages 27–35.

Viviana Cotik, Roland Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde, and Danilo Schmidt. 2016. Negation detection in clinical reports written in German. In *BioTxtM 2016 — Proceedings of the 5th Workshop on Building and Evaluating Resources for Biomedical Text Mining @ COLING 2016. Osaka, Japan, December 12, 2016*, pages 115–124.

Wessam Gad El-Rab, Osmar R. Zaïane, and Mohammad El-Hajj. 2017. Formalizing clinical practice guideline for clinical decision support systems. *Health Informatics Journal*, 23(2):146–156.

Erik Faessler, Michel Oleynik, and Udo Hahn. 2020. What makes a top-performing precision medicine search engine? Tracing main system features in a systematic way. In *SIGIR '20 — Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. July 25-30, 2020 (Virtual Event)*, pages 459–468.

Lejla Begic Fazlic, Ahmed Hallawa, Anke Schmeink, Arne Peine, Lukas Martin, and Guido Dartmann. 2019. A novel NLP-FUZZY system prototype for information extraction from medical guidelines. In *MIPRO 2019 — Proceedings of the 42nd Intl. Convention on Information and Communication Technology, Electronics and Microelectronics. Opatija, Croatia, 20-24 May 2019*, pages 1025–1030.

Georg Fette, Maximilian Ertl, Anja Wörner, Peter Klügl, Stefan Störk, and Frank Puppe. 2012. Information extraction from unstructured electronic health records and integration into a data warehouse. In *Proceedings der 42. Jahrestagung der Gesellschaft für Informatik (GI). Braunschweig, Germany, Sept. 16-21, 2012*, pages 1237–1251.

M. Follmann. 2020. German Guideline Program in Oncology. www.leitlinienprogramm-onkologie.de/english-language.

José Antonio Miñarro Giménez, Ronald Cornet, Marie Christine Jaulent, Heike Dewenter, Sylvia Thun, Kirstine Rosenbeck Gøeg, Daniel Karlsson, and Stefan Schulz. 2019. Quantitative analysis of manual annotation of clinical text samples. *International Journal of Medical Informatics*, 123:37–48.

Stefan Gindl, Katharina Kaiser, and Silvia Miksch. 2008. Syntactical negation detection in clinical practice guidelines. In *MIE 2008 — Proceedings of the 21st International Congress of the European Federation for Medical Informatics. Gothenburg, Sweden, 25-28 May 2008*, pages 187–192. IOS Press.

Lorraine Goeuriot, Hanna Suominen, Liadh Kelly, Antonio Miranda-Escalada, Martin Krallinger, Zhengyang Liu, Gabriella Pasi, Gabriela Gonzalez Saez, Marco Viviani, and Chenchen Xu. 2020. Overview of the CLEF eHealth Evaluation Lab 2020. In *CLEF 2020 — Proceedings of the 11th Intl. Conference of the CLEF Association. Thessaloniki, Greece, September 22-25, 2020*, pages 255–271.

Udo Hahn, Franz Matthies, Erik Faessler, and Johannes Hellrich. 2016. UIMA-based JCoRe 2.0 goes GitHub and Maven Central: state-of-the-art software resource engineering and distribution of NLP pipelines. In *LREC 2016 — Proceedings of*

the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016, pages 2502–2509.

Udo Hahn, Franz Matthies, Christina Lohr, and Markus Löffler. 2018. 3000PA : towards a national reference corpus of German clinical language. In *MIE 2018 — Proceedings of the 29th Conference on Medical Informatics in Europe. Gothenburg, Sweden, 24-26 April 2018*, pages 26–30. IOS Press.

Johannes Hellrich, Stefan Schulz, Sven Buechel, and Udo Hahn. 2015. JUFIT : a configurable rule engine for filtering and generating new multilingual UMLS terms. In *AMIA 2015 — Proceedings of the 2015 Annual Symposium of the American Medical Informatics Association. San Francisco, California, USA, Nov 14-18, 2015*, pages 604–610.

Hossein Hematialam and Wlodek W. Zadrozny. 2017. Extracting *condition-action* statements in medical guidelines. In *AMIA 2017 — Proceedings of the 2017 Annual Symposium of the American Medical Informatics Association. Washington, D.C., USA, November 4-8, 2017*.

George M. Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Tamseela Hussain, George Michel, and Richard N. Shiffman. 2009. The Yale Guideline Recommendation Corpus: a representative sample of the knowledge content of guidelines. *International Journal of Medical Informatics*, 78(5):354–363.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad M. Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:#160035.

Katharina Kaiser, Andreas Seyfang, and Silvia Miksch. 2010. Identifying treatment activities for modelling computer-interpretable clinical practice guidelines. In *KR4HC 2010 — Selected Papers of the International Workshop on Knowledge Representation for Health Care @ ECAI 2010. Lisbon, Portugal, August 17, 2010*, pages 114–125. Springer.

Tobias Kolditz, Christina Lohr, Johannes Hellrich, Luise Modersohn, Boris Betz, Michael Kiehntopf, and Udo Hahn. 2019. Annotating German clinical documents for de-identification. In *MEDINFO 2019 — Proceedings of the 17th World Congress on Medical and Health Informatics. Lyon, France, 25-30 August 2019*, pages 203–207. IOS Press.

Maximilian König, André Sander, Ilja Demuth, Daniel Diekmann, and Elisabeth Steinhagen-Thiessen. 2019. Knowledge-based best of breed approach for automated detection of clinical events based on German free text digital hospital discharge letters. *PLoS ONE*, 14(11):#e0224916.

Jonathan Krebs, Hamo Corovic, Georg Dietrich, Maximilian Ertl, Georg Fette, Mathias Kaspar, Markus Krug, Stefan Störk, and Frank Puppe. 2017. Semi-automatic terminology generation for information extraction from German chest X-ray reports. In *GMDS 2017 — Proceedings of the 62nd Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology. Oldenburg, Germany, 17-21 Sept. 2017*, pages 80–84. IOS Press.

Markus Kreuzthaler, Michel Oleynik, Alexander Avian, and Stefan Schulz. 2016. Unsupervised abbreviation detection in clinical narratives. In *ClinicalNLP 2016 — Proceedings of the 1st Workshop on Clinical Natural Language Processing @ COLING 2016. Osaka, Japan, December 11, 2016*, pages 91–98.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan So, and Jaewoo Kang. 2020. BIOBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tiffany I. Leung and Michel Dumontier. 2016. Overlap in drug-disease associations between clinical practice guidelines and drug structured product label indications. *Journal of Biomedical Semantics*, 7:#37.

Tiffany I. Leung, Hawre Jalal, Donna Zulman, Michel Dumontier, Douglas Owens, Mark Musen, and Mary Goldstein. 2015. Automating identification of multiple chronic conditions in clinical practice guidelines. In *Proceedings of the AMIA Joint Summits on Translational Science 2015. San Francisco, California, USA, March 23-27, 2015*, pages 456–460.

Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. UPSET : visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992.

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing copies of synthetic clinical corpora without physical distribution: a case study to get around IPRs and privacy constraints featuring the German JSYNCC corpus. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 1259–1266.

Christina Lohr and Robert Herms. 2016. A corpus of German clinical reports for ICD and OPS-based language modeling. In *CLAW 2016 — Proceedings of the 6th Workshop on Controlled Language Applications @ LREC 2016. Portorož, Slovenia, 28 May 2016*, pages 20–23.

Christina Lohr, Luise Modersohn, Johannes Hellrich, Tobias Kolditz, and Udo Hahn. 2020. An evolutionary approach to the annotation of discharge summaries. In *MIE 2020 — Proceedings of the 30th Conference on Medical Informatics Europe. Geneva, Switzerland, April 28 - May 1, 2020*, pages 28–32.

Mor Peleg. 2013. Computer-interpretable clinical guidelines: a methodological review. *Journal of Biomedical Informatics*, 46(4):744–763.

Jonathon L. Read, Erik Velldal, Marc Cavazza, and Gersende Georg. 2016. A corpus of clinical practice guidelines annotated with the importance of recommendations. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 1724–1731.

Roland Roller, Hans Uszkoreit, Feiyu Xu, Laura Seiffe, Michael Mikhailov, Oliver Staeck, Klemens Budde, Fabian Halleck, and Danilo Schmidt. 2016. A fine-grained corpus annotation schema of German nephrology records. In *ClinicalNLP 2016 — Proceedings of the 1st Workshop on Clinical Natural Language Processing @ COLING 2016. Osaka, Japan, December 11, 2016*, pages 69–77.

Matthieu-P. Schapranow, Milena Kraus, Cindy Perscheid, Cornelius Bock, Franz Liedke, and Hasso Plattner. 2015. The Medical Knowledge Cockpit: real-time analysis of big medical data enabling precision medicine. In *BIBM 2015 — Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine. Washington, DC, USA, 9-12 November 2015*, pages 770–775.

Laura Seiffe, Oliver Marten, Michael Mikhailov, Sven Schmeier, Sebastian Möller, and Roland Roller. 2020. From witch's shot to music making bones: resources for medical laymen to technical language and vice versa. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation. Marseille, France, May 11-16, 2020*, pages 6185–6192.

Radu Serban, Annette ten Teije, Frank van Harmelen, Mar Marcos, and Cristina Polo-Conde. 2007. Extraction and use of linguistic patterns for modelling medical guidelines. *Artificial Intelligence in Medicine*, 39(2):137–149.

Thomas Seufferlein, Ina Kopp, Stefan Post, Walter Jonat, Rolf Kreienberg, Monika Nothacker, Annika Marks, Gerd Nettekoven, Thomas Langer, Markus Follmann, and Michael Bamberg. 2019. Onkologische Leitlinien: Herausforderungen und zukünftige Entwicklungen. *Forum*, 34:277–283.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a Web-based tool for NLP-assisted text annotation. In *EACL 2012 — Proceedings of the 13th Conf. of the European Chapter of the Association for Computational Linguistics: Demonstrations. Avignon, France, April 25-26, 2012*, pages 102–107.

Jannik Strötgen and Michael Gertz. 2011. WIKIWARS-DE : a German corpus of narratives annotated with temporal expressions. In *GSCL 2011—Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology. Hamburg, Germany, 28-30 Sept. 2011*, pages 129–134.

Jannik Strötgen, Anne-Lyse Minard, Lukas Lange, Manuela Speranza, and Bernardo Magnini. 2018. KRAUTS: a German temporally annotated news corpus. In *LREC 2018 — Proceedings of the 11th Intl. Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 536–540.

Maria Taboada, Maria Meizoso, Diego Martínez, David Riaño, and Albert Alonso. 2013. Combining open-source natural language processing tools to parse clinical practice guidelines. *Expert Systems*, 30(1):3–11.

Martin Toepfer, Hamo Corovic, Georg Fette, Peter Klügl, Stefan Störk, and Frank Puppe. 2015. Fine-grained information extraction from German transthoracic echocardiography reports. *BMC Medical Informatics and Decision Making*, 15:#91.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 I2B2/VA Challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn R. Daughton, Karen O'Connor, Michael J. Paul, and G. Gonzalez-Hernandez. 2019. Overview of the 4th Social Media Mining for Health (#SMM-4H) Shared Task at ACL 2019. In *#SMM4H 2019 — Proceedings of the 4th Workshop on Social Media Mining for Health Applications Shared Task @ ACL 2019. Florence, Italy, August 2, 2019*, pages 21–30.

Reinhardt Wenzina and Katharina Kaiser. 2013. Identifying *condition-action* sentences using a heuristic-based information extraction method. In *KR4HC-ProHealth 2013 — Selected Papers of the Joint International Workshop on Knowledge Representation for Health Care & Process-oriented Information Systems in Healthcare @ AIME 2013. Murcia, Spain, June 1, 2013*, pages 26–38. Springer.

Joachim Wermter and Udo Hahn. 2004. Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora. In *MEDINFO 2004 — Proceedings of the 11th World Congress on Medical Informatics. San Francisco, California, USA, September 7-11, 2004*, volume 1, pages 560–564. IOS Press.

Wlodek W. Zadrozny, Hossein Hematialam, and Luciana Garbayo. 2017. Towards semantic modeling of contradictions and disagreements: a case study of medical guidelines. In *IWCS 2017 — Proceedings of the 12th International Conference on Computational Semantics. Montpellier, France 19-22 September 2017*, volume 2: Short Papers, page #43.

Huijia Zhu, Yuan Ni, Peng Cai, and Feng Cao. 2013. Automatic information extraction for computerized clinical guideline. In *MEDINFO 2013 — Proceedings of the 14th World Congress on Medical and Health Informatics. Copenhagen, Denmark, 20-23 August 2013*, page 1023. IOS Press.