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

Florian Borchert*,†‡, Christina Lohr*,†‡, Luise Modersohn*,†‡, Thomas Langer§, Markus Foellmann§, Jan Philipp Sachs§, Udo Hahn†¶ and Matthieu-P. Schapranow*¶

†Digital Health Center, Hasso Plattner Institute, University of Potsdam, Germany
‡Jena University Language & Information Engineering Lab, Friedrich Schiller University Jena, Germany
§German Guideline Program in Oncology, German Cancer Society, Berlin, Germany
¶HiGHmed Consortium of the German Medical Informatics Initiative

Abstract—The lack of publicly available text corpora is a major obstacle for progress in clinical natural language processing, for non-English speaking countries in particular. In this work, we present GGPONC (German Guideline Program in Oncology NLP Corpus), a freely distributable German language corpus based on clinical practice guidelines in the field of oncology. The corpus is one of the largest corpora of German medical text to date. It does not contain any patient-related data and can therefore be used without data protection restrictions. Moreover, it is the first corpus for the German language covering diverse conditions in a large medical subfield. In addition to the textual sources, we provide a large 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 medical text corpora.

Index Terms—natural language processing, clinical guidelines, clinical text, clinical guidelines, German text

I. INTRODUCTION

Evidence synthesis 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 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 the unstructured texts and transform them into a structured, or even executable, format. As CPGs are also specific to their country of origin, they are usually published in the respective native language, so NLP technology has to be adapted properly.

A major source for the progress in NLP research in the recent years is the public availability of large text corpora. For documents originating from a clinical context, the protection of personal information is a major requirement for accessibility to researchers. Some corpus initiatives, e.g., i2b2 [1], MIMIC III [2], or CLEF eHEALTH [3] make de-identified clinical 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 open-access text corpora are hardly available for the German [4] and other non-English languages. As of today, there is no viable solution for sharing even de-identified clinical texts in Germany.

In order 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 (GGPO) [5], 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 [6]. Hence, the data set is 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.

The GGPO guidelines are available free of charge and do not contain sensitive data about individual patients. However, as processing of the guidelines is by default still prohibited by the issuer’s copyright, we provide access to the pre-processed data for other researchers via a DUA.

II. RELATED WORK

Due to legal data protection regulations, the availability of German-language clinical text corpora is severely restricted — most clinical corpora are only accessible to the research staff within the lifetime of a project and remain inaccessible.
forever for the outside world. There have been a few disconnected activities in the German NLP community to create in-project clinical corpora. In Table I we list, to the best of our knowledge, all existing German-language clinical research text corpora with clinical documents or collections of case reports that have been described in scientific publications. In addition to pure clinical documents, other document types are also interesting for the NLP community, e.g. CPGs.

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 [18]. Bouffier and Poibeau [19] 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. [20] 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 Bockmann [21] who adapted APACHE CTAKES to detect German UMLS concepts and evaluated their approach on a single German breast cancer guideline. Zadrozny et al. [22] outline a system which identifies contradictions and disagreements in English CPGs.

Some authors have focused on extracting more task-specific information, such as activities [23], process structures [24, 25, 26] or negation triggers [27]. Most of these approaches work with relatively small annotated corpora and English language, only. Recently, Fazlic et al. [28] use LSTMs and fuzzy rules to extract “action takers”, “symptoms”, “actions” and “purposes” from CPGs, recognize recommendations and predict the grade of recommendation. The authors use a data set extracted from PDF versions of 45 guidelines with 1,020 recommendations. Some larger corpora of CPGs in the English language exist already: Hussain et al. [29] present the Yale Guideline Recommendation Corpus (YGRC), a sample of 1,275 guideline recommendations extracted from National Guideline Clearinghouse (NGC). Their work revealed inconsistencies in writing style and reporting of the strength of recommendations. Using a subset of YGRC, Gad El-Rab et al. [30] present a rule-based approach to detect procedures and drug recommendations. Read et al. [31] 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. [32] identify comorbidities by analyzing pairs of co-occurring conditions, using a corpus of 268 NGC guideline summaries. Leung and Dumontier [33] identify drug-disease relations via named entity recognition using a corpus of 377 NGC guideline summaries. The extracted relations are compared to structured drug product labels to assess their overlap.

### III. Methods

#### A. 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 I.

The guidelines distinguish between recommendations and background texts, and 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.

| Corpus / Data | Documents | Sentences | Tokens | Available |
|---------------|-----------|-----------|--------|-----------|
| **Framed:** clinical reports and medical textbook snippets [7] | – | 6k | 100k | ❌ |
| Reports from five medical domains [8] | 544 | – | – | ❌ |
| Radiology reports [9] | 174 | 4k | 28k | ❌ |
| Transcatheter echocardiography reports [10] | 140 | – | – | ❌ |
| Operative reports (surgery) [11] | 420 | 22k | 266k | ❌ |
| Discharge summaries from a dermatology department [12] | 1,696 | – | – | ❌ |
| Discharge summaries and clinical notes from nephrology domain [13] | 1,725 | 28k | 158k | ❌ |
| Discharge summaries and clinical notes from nephrology domain [14] | 183 | 2k | 13k | ❌ |
| X-ray reports [15] | 3,000 | – | – | ❌ |
| 3000PA: internist and ICU discharge summaries [16] | – | 3,000 | – | ❌ |
| 5000PA JENA PART | – | 1,006 | 170k | 1,421k | ❌ |
| JSYNCC: case examples from medical textbooks [17] (v1.1) | – | 903 | 29k | 368k | ✓ |
| Discharge summaries with osteoporosis diagnosis [17] | – | 1,982 | – | 2,001k | ❌ |
| **GGPO** – recommendations | 25 (4,348) | 7k | 132k | ✓ |
| **GGPO** – complete corpus | 25 (8,418) | 60k | 1,340k | ✓ |
B. Automated Annotation

Besides the XML version of the corpus, we created plain text versions of all recommendation of background text parts to facilitate processing by existing NLP pipelines. For preprocessing, like sentence splitting and tokenization, we used the JCoRE \[37\] (i.e., UIMA-based) pipelines and FRAMES models, which were developed for German clinical text.

We used the JuFit\[38\] tool, a filter for UMLS, to create a dictionary of all German words from the UMLS\[39\] (version 2019AB\[1\]) 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), as well as a list of gene names compiled from Entrez Gene and UniProt with the approach originating from Wermter et al.\[40\] and German stop words. With these dictionaries, we configured a JCoRE pipeline for a dictionary-based text search.

Finally, we detect TNM expressions\[2\], which are extracted using a rule-based approach implemented with the PYTHON library SPAC\[2\]. 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.

III. RESULTS

A. Corpus Characteristics

In total, 25 GPGs with 8,414 text segments were extracted from the CMS comprising the first version of the corpus (summarized in Table III). We report the total number of recommendations and background text segments, since they serve as the units of analysis for our automated annotation pipelines. While the number of recommendations in GGPO\[NC\] is comparable to the CREST corpus,\[31\] the amount of structured metadata and background text in our corpus is much larger.

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 CPG and the underlying medical evidence. Table IV summarizes the automated entity extraction results. The result quality and their 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

As CPG are subject to a regular update cycle, we are able to automatically repeat the data acquisition process in the future to provide a historical view on the guideline development. For instructions on how to get access to the corpus see: https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-english

B. Comparison with Other German Medical and Non-Medical Corpora

We analyze the characteristics of GGPO\[NC\] 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)\[4\], the Jena Part of the 3000PA corpus (1006 German discharge summaries)\[16\] as well as abstracts from German case reports from PUBMED. In addition, we compare the results to out-of-domain corpora consisting of German Wikipedia articles of wars (WIKI\[WARS\]DE)\[41\] and news articles from the KRAUTS corpus\[42\]. The results are summarized in Table IV.

The fraction of stop words is comparable across all medical text corpora, as is the fraction tokens that map to UMLS concepts. As expected, the guideline recommendations contain

\[2\]The UICC TNM system is a classification scheme for malignant tumors, see: https://www.uicc.org/resources/tnm

\[3\]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.

\[1\]https://www.nlm.nih.gov/research/umls/
TABLE II

METADATA ELEMENTS OF RECOMMENDATIONS OF GGPO NC

| 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 [34], SIGN [35], or GRADE [36]                          |
| Edit state                    | State (checked, new or modified) and text note regarding guideline updates    |

TABLE III

DETAILS OF THE GGPO NC TEXT CORPUS. THE NUMBERS OF TOKENS AND TYPES REFER TO THE PURE TEXTUAL CONTENT OF THE CORPUS, EXCLUDING ANY META-DATA AND HEADINGS.

| Guideline                                      | Segments | Recommendations | Sentences | Tokens    | References |
|------------------------------------------------|----------|-----------------|-----------|-----------|------------|
| 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     | 2,321      |
| 6 Colorectal cancer                            | 509      | 290             | 3,176     | 71,416    | 9,644      | 2,580      |
| 7 Prostate cancer                              | 307      | 221             | 3,090     | 67,900    | 12,660     | 2,824      |
| 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    | 9,140      | 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     | 48,076    | 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      | 1,127      |
| 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      | 43              | 779       | 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      |
| Full Corpus                                    | 8,414    | 4,348           | 59,672    | 1,340,201 | 76,252     | 37,928     |

more medical terms per token than the background text. Compared to the clinical corpora, the guiofficelines 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 GGPO NC, 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 GGPO NC 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.

C. Evaluation of Annotation Results

The automatic annotations for a subset of the CPGs have been independently reviewed by human experts (1 medical doctor and 3 students of medicine, all of them passed their first medical exam) using the BRAT annotation tool [43]. Due to restricted resources for manual annotation work, we decided to evaluate on a subset of four (full) guidelines of a
TABLE IV
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 corpora | Non-Clinical corpora |
|--------|-----------------|---------------------|
|        | Complete | Recom. | 3000PAJ | JSynCC | PUBMED | 95,604 | 31,422 |
| Tokens | 1,340,201 | 132,145 | 1,421,713 | 368,389 | 43,110 | 95,604 | 31,422 |
| Sentences | 39,672 | 6,969 | 170,539 | 29,476 | 2,012 | 4,564 | 1,244 |
| 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 | 7.15 | 0.02 |
| ANAT (%) | 0.45 | 0.48 | 1.78 | 1.11 | 0.79 | 0.04 | 0.09 |
| CHEM (%) | 0.32 | 1.01 | 1.08 | 0.41 | 0.59 | 0.04 | 0.09 |
| 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 | 0.30 | 0.12 | 0.19 |
| 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 | 0.0 |
| Stop words (%) | 34.05 | 35.53 | 20.37 | 32.96 | 34.51 | 34.65 | 24.24 |

Fig. 1. Intersection of distinct UMLS concepts in JSynCC1.1, 3000PA (Jena part) and GGPOnc. The vertical bars indicate the size of all intersecting subsets of terminology shared between the corpora, whereas the horizontal bars denote the total number of distinct concepts per corpus.

The agreement subset consists of the five text segments with the largest amount of automatic annotations for each of the four guidelines, resulting in 20 agreement documents with a size of approx. 0.7–0.8 have shown to be normal for typical clinical entities, 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 [45]. 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) [46].

V. DISCUSSION & LIMITATIONS

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, augmenting the dictionary-based approach with well-known improvements employed in gene taggers for English texts is one of the next steps.

The German UMLS has a number of issues, which severely affect our dictionary-based entity extraction pipelines. First and foremost, its vocabulary size is very limited. For instance, 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% of the English version). Recently introduced drugs are missing in the UMLS Chemistry category, so a more up-to-date dictionary of drug names should be used for future work. Moreover, use of German umlauts is inconsistent in UMLS, e.g., a is sometimes transcribed as ae, as in eingeschraenkte Nierenfunktion, which results in a higher than necessary false negative rate. All of these factors contribute to rather low recall values, as evident in Table [V].

The agreement subset consists of the five text segments with...
The accuracy of our dictionary matches is affected by inconsistencies in the use of compounds across the corpus. For instance, *Pankreaskarzinompatienten* (patients with pancreas carcinoma) would not be detected as an entity, whereas *Pankreaskarzinom-Patienten* would be, as 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 our rule-based TNM extraction approach are high on GGPO NC, one has to be careful as certain TNM expressions can have a completely different meaning in another context (ambiguity). For instance, V1 and V2 are valid TNM components referring to venous invasion, but are also detected in the WikiWARSDE corpus when actually referring to German missiles from World War II.

### VI. Conclusion

We presented GGPO NC, one of the largest corpora of German medical text to date, assembled from the German CPGs in oncology and equipped with rich structure and metadata. We applied clinical information extraction pipelines to extract a variety of 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 [6] or in clinical decision support systems [47]. Our results indicate that GGPO NC 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. [48] 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.

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 if these CPGs might be outdated given changes in the underlying evidence base.

We make GGPO NC 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://github.com/JULIELab/GGPOnc/

### ACKNOWLEDGMENTS

Parts of this work were generously supported by the German Federal Ministry of Research and Education (BMBF) under grants (01ZZ1802H, 01ZZ1803G). We thank all annotators, as well as André Scherag and Danny Ammon from the Jena University Hospital. In addition, we thank all colleagues from the HiGHmed and SMITH consortia for their constant support and valuable input contributing to our joint research.

### References

[1] Ĭ. Uzuner, B. R. South, S. Shen, and S. L. DuVall, “2010 12B2/VA challenge on concepts, assertions, and relations in clinical text,” *JAMIA*, vol. 18, no. 5, pp. 552–556, 2011.

[2] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “MIMIC-III, a freely accessible critical care database,” *Scientific Data*, vol. 3, p. #160035, 2016.

[3] L. Kelly, H. Suominen, L. Goereuit et al., “Overview of the CLEF eHealth evaluation lab 2019,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2019, pp. 322–339.

[4] C. Lohr, S. Buechel, and U. Hahn, “Sharing copies of synthetic clinical corpora without physical distribution: a case study to get around IPRs and privacy constraints
“German guideline program in oncology,” in LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7–12, 2018, 2018, pp. 1259–1266.

“German guideline program in oncology,” https://www.leitlinienprogramm-onkologie.de/english-language/2020, accessed: 2020-03-11.

T. Seufferlein, I. Kopp, S. Post et al., “Onkologische Leitlinien – Herausforderungen und zukünftige Entwicklungen,” Forum, vol. 34, no. 3, pp. 277–283, Jun 2019.

J. Wermter and U. Hahn, “Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora,” in Proceedings of the 11th World Congress on Medical Informatics, ser. Studies in Health Technology and Informatics, no. 107. Amsterdam: IOS Press, 2004, pp. 560–564.

G. Fette, M. Ertl, A. Wörner, P. Klügl, S. Stöck, and F. Puppe, “Information extraction from unstructured electronic health records and integration into a data warehouse,” in INFORMATIK 2012: Was bewegt uns in der/die Zukunft? Proceedings der 42. Jahrestagung der Gesellschaft für Informatik e.V. (GI). Braunschweig, Germany, September 16–21, 2012, ser. GI-Edition - Lecture Notes in Informatics (LNI), no. P-208, Gesellschaft für Informatik e.V. (GI). Bonner Köllen Verlag, 2012, pp. 1237–1251.

C. Bretschneider, S. Zillner, and M. Hammon, “Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach,” in BioNLP 2013 — Proceedings of the Workshop on Biomedical Natural Language Processing @ ACL. Sofia, Bulgaria, 2013, pp. 27–35.

M. Toepfer, H. Corovic, G. Fette, P. Klügl, S. Stöck, and F. Puppe, “Fine-grained information extraction from German transthoracic echocardiography reports,” BMC Medical Informatics and Decision Making, vol. 15, p. #91, 2015.

C. Lohr and R. Herm, “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, 2016, pp. 20–23.

M. Kreuzthaler, M. Oleynik, A. Avian, and S. Schulz, “Unsupervised abbreviation detection in clinical narratives,” in ClinicalNLP 2016 — Proceedings of the Clinical Natural Language Processing Workshop @ COLING 2016. Osaka, Japan, December 11, 2016, 2016, pp. 91–98.

R. Roller, H. Uszkoreit, F. Xu et al., “A fine-grained corpus annotation schema of German nephrology records,” in ClinicalNLP 2016 — Proceedings of the Clinical Natural Language Processing Workshop @ COLING 2016. Osaka, Japan, December 11, 2016, 2016, pp. 69–77.

V. Cotik, R. Roller, F. Xu, H. Uszkoreit, K. Buddle, and D. Schmidt, “Negation detection in clinical reports written in German,” in BioTtxM 2016 — Proceedings of the 5th Workshop on Workshop on Building and Evaluating Resources for Biomedical Text Mining @ COLING 2016. Osaka, Japan, December 12, 2016, 2016, pp. 115–124.

J. Krebs, H. Corovic, G. Dietrich et al., “Semi-automatic terminology generation for information extraction from German chest X-ray reports,” in Proceedings of the 62nd Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology (GMDS e.V.). Oldenburg, no. 243. IOS Press, 2017, pp. 80–84.

U. Hahn, F. Matthies, C. Lohr, and M. Löfler, “3000PA: Towards a national reference corpus of German clinical language,” in MIE 2018 — Proceedings of the 29th Medical Informatics in Europe Conference. Gothenburg, Sweden, April 23-25, 2018, pp. 26–30.

M. König, A. Sander, I. Demuth, D. Diekmann, and E. Steinhagen-Thiessen, “Knowledge-based best of breed approach for automated detection of clinical events based on german free text digital hospital discharge letters,” PloS one, vol. 14, no. 11, 2019.

M. Peleg, “Computer-interpretable clinical guidelines: a methodological review,” JBI, vol. 46, no. 4, pp. 744–763, 2013.

A. Bouffier and T. Poibeau, “Automatically restructuring practice guidelines using the GEM DTD,” in Biological, translational, and clinical language processing, Prague, 2007, pp. 113–120.

R. Serban, A. ten Teije, F. van Harmelen, M. Marcos, and C. Polo-Conde, “Extraction and use of linguistic patterns for modelling medical guidelines,” Artificial Intelligence in Medicine, vol. 39, no. 2, pp. 137–149, 2007.

M. Becker and B. Bockmann, “Semi-automatic mark-up and UMLS annotation of clinical guidelines,” Studies in health technology and informatics, vol. 245, pp. 294–297, 2017.

W. W. Zadrozny, H. Hempl, and L. Garbayo, “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, C. Gardent and C. Retoré, Eds., vol. 2: Short Papers. Association for Computational Linguistics (ACL), 2017, p. #43.

K. Kaiser, A. Seyfang, and S. Miksch, “Identifying treatment activities for modelling computer-interpretable clinical practice guidelines,” in International Workshop on Knowledge Representation for Health Care. Springer, 2010, pp. 114–125.

R. Wenzina and K. Kaiser, “Identifying condition-action sentences using a heuristic-based information extraction method,” in Process Support and Knowledge Representation in Health Care. Springer, 2013, pp. 26–38.

H. Zhu, Y. Ni, P. Cai, and F. Cao, “Automatic information extraction for computerized clinical guideline,” Studies in health technology and informatics, vol. 192, p. 1023,
H. Hematialam and W. Zadrozny, “Extracting condition-action statements in medical guidelines,” in AMIA 2017, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 4–8, 2017.

S. Gindl, K. Kaiser, and S. Miksch, “Syntactical negation detection in clinical practice guidelines,” Studies in health technology and informatics, vol. 136, pp. 187–192, 2008.

L. B. Fazlic, A. Hallawa, A. Schmeink, A. Peine, L. Martin, and G. Dartmann, “A novel NLP-FUZZY system prototype for information extraction from medical guidelines,” in 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019, pp. 1025–1030.

T. Hussain, G. Michel, and R. N. Shiffman, “The importance of recommendations,” in Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL), 09 2011, pp. 129–134.

J. Strögen and M. Gertz, “WikiWarsDE: a german corpus of narratives annotated with temporal expressions,” in Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL), 09 2011, pp. 129–134.

W. Gad El-Rab, O. R. Zaíane, and M. El-Hajj, “Formalizing medical practice guideline for clinical decision support systems,” Health Informatics Journal, vol. 23, no. 2, pp. 146–156, 2017.

J. Read, E. Velldal, M. Cavazza, and G. Georg, “A corpus of clinical practice guidelines annotated with the importance of recommendations,” in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, 2016, pp. 1724–1731.

T. I. Leung, H. Jalal, D. M. Zulman et al., “Automating identification of multiple chronic conditions in clinical practice Guidelines.” AMIA Joint Summits on Translational Science proceedings, pp. 456–460, 2015.

T. I. Leung and M. Dumontier, “Overlap in drug-disease associations between clinical practice guidelines and drug structured product label indications.” Journal of biomedical semantics, vol. 7, p. 37, 2016.

“Ocebpm levels of evidence working group. the oxford 2011 levels of evidence,” http://www.cebm.net/index.aspx?o=5653, 2011.

R. Harbour and J. Miller, “A new system for grading recommendations in evidence based guidelines,” Bmj, vol. 323, no. 7308, pp. 334–336, 2001.

H. Balshem, M. Helfand, H. J. Schünemann, A. D. Oxman, R. Kunz, J. Brozek, G. E. Vist, Y. Falck-Ytter, J. Meerpohl, S. Norris et al., “Grade guidelines: 3. rating the quality of evidence.” Journal of clinical epidemiology, vol. 64, no. 4, pp. 401–406, 2011.

U. Hahn, F. Matthies, E. Faessler, and J. Hellrich, “Uima-based jcore 2.0 goes github and maven central — state-of-the-art software resource engineering and distribution of nlp pipelines,” in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, 2016, pp. 2502–2509.

J. Hellrich, S. Schulz, S. Buechel, and U. Hahn, “JUFIT: a configurable rule engine for filtering and generating new multilingual UMLS terms,” in AMIA 2015, 2015, pp. 604–610.

O. Bodenreider, “The unified medical language system (umls): integrating biomedical terminology.” Nucleic Acids Research, vol. 32, pp. D267–D270, 01 2004.

J. Wermter, K. Tomanek, and U. Hahn, “High-performance gene name normalization with GeNo,” Bioinformatics, vol. 25, no. 6, pp. 815–821, 2009.

T. I. Leung, H. Jalal, D. M. Zulman, T. I. Leung, H. Jalal, D. M. Zulman et al., “Automating identification of multiple chronic conditions in clinical practice guidelines,” in AMIA Joint Summits on Translational Science proceedings, pp. 456–460, 2015.