Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3650–3660
Marseille, 20-25 June 2022

GGPONC 2.0 — The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers

Florian Borchert[1,★], Christina Lohr[2,♠], Luise Modersohn[2,♠], Jonas Witt[1], Thomas Langer[3], Markus Follmann[3], Matthias Gietzelt[4,★], Bert Arnrich[1], Udo Hahn[2,♠], Matthieu-P. Schapranow[1,★]

[1]Digital Health Center, Hasso Plattner Institute, University of Potsdam, Germany
[2]Jena University Language & Information Engineering (JULIE) Lab, Friedrich Schiller University Jena, Germany
[3]German Guideline Program in Oncology, German Cancer Society, Berlin, Germany
[4]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany
★HIGHMED Consortium of the German Medical Informatics Initiative
♠SMITH Consortium of the German Medical Informatics Initiative
{firstname.lastname}@{hpi|uni-jena|mh-hannover}.de, {lastname}@krebsgesellschaft.de

Abstract

Despite remarkable advances in the development of language resources over the recent years, there is still a shortage of annotated, publicly available corpora covering (German) medical language. With the initial release of the German Guideline Program in Oncology NLP Corpus (GGPONC), we have demonstrated how such corpora can be built upon clinical guidelines, a widely available resource in many natural languages with a reasonable coverage of medical terminology. In this work, we describe a major new release for GGPONC. The corpus has been substantially extended in size and re-annotated with a new annotation scheme based on SNOMED CT top level hierarchies, reaching high inter-annotator agreement ($\gamma = .94$). Moreover, we annotated elliptical coordinated noun phrases and their resolutions, a common language phenomenon in (not only German) scientific documents. We also trained BERT-based named entity recognition models on this new data set, which achieve high performance on short, coarse-grained entity spans ($F_1 = .89$), while the rate of boundary errors increases for long entity spans. GGPONC is freely available through a data use agreement. The trained named entity recognition models, as well as the detailed annotation guide, are also made publicly available.

Keywords: Clinical Guidelines, Clinical NLP, Annotation, Named Entity Recognition, SNOMED CT, German Medical Language

1. Introduction

In the clinical domain, the inability to share language resources and models has been a major obstacle for researchers and medical practitioners who want to gain insights from clinical documents through natural language processing (NLP). This bottleneck is due to protected health information (PHI) in these documents, which allows to identify patients, clinical personnel, or other individuals. Even thorough de-identification efforts targeting PHI items (for a survey, cf. Meystre (2015)) do not guarantee positive votes from ethical boards for allowing strictly anonymized clinical data to pass hospital walls. Likewise, machine learning models trained on these data are usually not shareable, because attackers might re-identify sensitive data from their model parameters (Carlini et al., 2021).

Whereas the distribution of real, de-identified clinical documents on a reasonable scale seems out of scope for almost all European countries for the time being, scientific research publications are not subject to such distribution restrictions. Unfortunately, the vast majority of scholarly medical articles are only written in English. Clinical guidelines are a notable exception, as they are typically issued in the national language of their target audience. In effect, freely distributable medical language resources can be created from clinical guidelines in a variety of languages. The German Guideline Program in Oncology NLP Corpus 1.0 (GGPONC) has been the first of its kind for the German language and already stands out as the largest, manually annotated, publicly accessible German-language medical text corpus (Borchert et al., 2020).

In this work, we describe our findings from the latest GGPONC release 2.0, which has increased by around 40% in volume compared to its first version. The complete corpus has been re-annotated with a new entity annotation scheme, improved in terms of annotation quality, and augmented by explicit indication of elliptical coordinated noun phrases (CNPs), a common phenomenon in scientific text (Blake and Rindflesch, 2017). In addition to the pre-processed textual data, extensive guideline metadata, and human annotations, we publish several named entity recognition (NER) models trained on GGPONC 2.0 as well as the detailed annotation guide (HPI-DHC, 2022).

This paper is organized as follows: in Section 2 we share experiences since the release of GGPONC 1.0. We review related work in Section 3 and outline our methodology and the annotation process in Section 4. In Section 5 we present results and performance of baseline NER models. We conclude with a discussion of the results in Section 6 and an outlook in Section 7.

2. Lessons Learned from GGPONC 1.0

The current GGPONC release builds on our experience from the previous annotation campaign as well as ongoing maintenance of the corpus since its creation.

2.1. Annotations

For the initial release, we created silver-standard entity annotations for the complete corpus using a dictionary based on the *Unified Medical Language System* (UMLS) semantic groups (Bodenreider, 2004). Human annotators manually reviewed these annotations for around half of the corpus (664k tokens).

As the German UMLS subset contains only a fraction of the English version (roughly 1.55 % German concept names (257k) vs. 71.06 % English ones (11,756k) from overall 16,544k concept names in the complete UMLS in its 2021AB release),[1] the silver-standard annotations missed a large portion of concept mentions, resulting in low recall (54%). Moreover, the agreement between human annotators was relatively low (F_1 = .74) due to sloppy, underspecified annotation guidelines, but also the fuzzy meaning of many UMLS semantic groups from a clinical perspective.

While integrating GGPONC into a software application for finding clinical trial reports that disagree with current guidelines, we identified further limitations of our UMLS-based dictionary (Borchert et al., 2021). For instance, we need to detect all pharmacological interventions in the guidelines. However, many concepts were assigned to the semantic type *Pharmacologic Substance* in the UMLS, which in the context of a clinical guideline are not used as such (e.g., *water*, *air*). For these reasons, we adopted new definitions of entity classes based on the SNOMED CT concept model (Donnelly, 2006), as described in Section 4.2.

2.2. Community Feedback

We received a total of 31 data use requests for our initial release of GGPONC 1.0 from July 2020 to April 2022. Nine of these users responded to a questionnaire survey accompanying the access process. 56 % of the respondents indicated that they made use of the entity annotations in GGPONC and pointed out that additional annotation formats would be desirable. Therefore, in the new version, we not only provide all plain text file formats that were included in the first release, i.e., sentence-splitted and tokenized versions, but also make available the raw WebAnno TSV files from the INCEPTION annotation tool (Klie et al., 2018), IOB-encoded named entity annotations in a simple CONLL format, dataset files in a transformers-compatible JSON format (Wolf et al., 2020), as well as serialized SPACY documents for training a SPANCATEGORIZER model (Montani et al., 2021).

2.3. Data Access and Release Cycle

Data use requests are handled via e-mail to the German Guideline Program in Oncology Office.[2] Access is granted to non-commercial, scientific users after providing a brief description of their research purpose. Given the current rate of requests (about 2-3 per month), we can ensure timely access for users, with only minor manual overhead for reviewing requests and sending out download links. Based on the feedback from our survey, we plan to provide more frequent minor releases in the future, i.e., release upon update of individual guidelines, but at most once per month. Prior and future releases will be available via a dedicated file server.

3. Related Work

A few text corpora based on clinical guidelines exist for the English language (Hussain et al., 2009; Read et al., 2016). Yet, none of them was annotated for entities. For an overview of NLP applications in the domain of clinical guidelines, we refer to Borchert et al. (2020). To the best of our knowledge, GGPONC 2.0 is currently the largest publicly available, semantically annotated German medical text corpus.

In Table 1 we list a selection of corpora from different medical domains with comparable entity annotations. Two observations can be made. First, there is an increasing tendency to release larger-sized corpora ($\gg$ 100K tokens). Second, there is, much to a surprise, not really consensus about the named entity types to be annotated—yet, SNOMED top level categories and UMLS semantic groups might be the most reasonable candidates for such a much-needed convergence (our work here adheres to the SNOMED categorization). While *unlabeled* German medical text datasets with orders of magnitude more tokens have recently been used for unsupervised pre-training, none of these is publicly available (Bressem et al., 2020; Richter-Pechanski et al., 2021). Although GGPONC is comprised of scientific articles, rather than clinical narratives, we have previously traced a substantial overlap in terminology with clinical corpora, e.g., GGPONC 1.0 and the Jena part of 3000PA share about 40% of unique UMLS concepts (Borchert et al., 2020).

Coordination ellipses in medical documents have been the subject of a number of prior studies (Buyko et al., 2007; Chae et al., 2014; Wei et al., 2015; Blake and Rindflesch, 2017). All of them point out the potential benefits of resolving ellipses for the performance of downstream applications. Yet, none of these studies explicitly deal with German medical documents. Hence, GGPONC 2.0, to the best of our knowledge, is the first German medical text corpus with explicit annotation of (resolved) elliptical CNPs.

[1] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

[2] https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-english/

Corpus	Description	Doc.	Sent.	Tokens	Entities	Agreement
(Mostly, English-language) Corpora						
Wang (2009)	Clinical notes	0.31K	–	47K	11 SNOMED CT top level concepts	$F_{1(exact)} = .88$
Roberts et al. (2009)	CLEF (Clinical text)	0.15K	–	–	Condition, Intervention, Locus, Result, etc.	$F_1 \in [.74, .99]$
van Mulligen et al. (2012)	EU-ADR (MEDLINE)	0.30K	–	–	Drug, Disorder, Gene, Protein	$F_1 = .75$
Doğan et al. (2014)	NCBI disease (MEDLINE)	0.79K	7K	–	Disease	$F_1 \approx .90$
Patel et al. (2018)	Clinical documents	5.16K	398K	**3,825K**	Various (based on UMLS semantic types)	$\kappa \approx .97$
Nye et al. (2018)	EBM-NLP (MEDLINE)	5.00K	–	–	Participants Intervention Outcome	$\kappa \in [.50, .71]$ $\kappa \in [.59, .69]$ $\kappa \in [.51, .62]$
Miñarro-Giménez et al. (2019)	Clinical snippets (multilingual)	–	–	41K	SNOMED CT concepts	$\alpha_{loose} \in [.40, .74]$
Schulz et al. (2020)	Case reports (PUBMED Central)	0.05K	8K	168K	Case, Condition, Finding, Factor	$\kappa = .68$
German-language Corpora						
Hahn et al. (2018) Lohr et al. (2020)	3000PA Jena part (Discharge summ.)	1.11K	170K	1,421K	Diagnosis, Symptoms, Finding	$F_{1(inst)} = .65$ $F_{1(tok)} = .84$
Borchert et al. (2020)	GGPONC 1.0	8.42K	60K	1,340K	UMLS semantic groups, TNM	$F_1 = .74$
Kittner et al. (2021)	BRONCO (Discharge summ.)	0.20K	11K	90K	Diagnosis Treatment Medication	$F_1 \in [.69, .88]$ $F_1 \in [.66, .81]$ $F_1 \in [.87, .94]$
Our work	**GGPONC 2.0**	10.19K	78K	**1,877K**	Finding, Substance, Procedure	$\gamma = .94$

Table 1: Overview of medical text corpora with clinical annotations comparable to this work. GGPONC is among the largest annotated medical text corpora, in particular for the German language. Many authors have not reported numbers of sentences or tokens in the past, allowing only limited comparison by volume.

4. Materials and Methods

In the following, we elaborate on details of our methodology and available data.

4.1. Data Acquisition

The workflow to create GGPONC releases is outlined in Fig. 1. First, we retrieve the semi-structured guideline data from the GGPO's content management system. From this raw data, we then create a single XML file with the document structure and guidelines' metadata, excluding literature references which are written to a literature index file. Individual plain text documents, corresponding to text segments of the guidelines, can be linked back to their metadata (like timestamps, recommendation levels, etc.) through a metadata index file. We use a JCORE pipeline (Hahn et al., 2016) configured with FRAMED models (Wermter and Hahn, 2004) for sentence splitting and tokenization. For the current release, we have extracted all 30 available oncology guidelines, summarized in Table 2.

4.2. Annotation

For our GGPONC 2.0 release, all documents were manually annotated with named entities and elliptical CNPs by seven annotators. This team was composed of students of medicine from the Charité University Hospital in Berlin, who had already passed their first medical exam (*Physicum*, in German), supported by a medical doctor who resolved annotation conflicts during the refinement of the annotation guide (see Section 4.2.3). Annotators spent 1,142 hours, in total, for annotating individual documents with the INCEPTION platform (Klie et al., 2018) over a time span of six months. The mean annotation speed was 1.65K tokens per hour, but with high variability between individual annotators ($\sigma = 473.3$). No pre-annotations were provided to prevent biasing the annotation process. However, we enabled default string-based recommenders in INCEPTION based on previous annotations made by the same annotators.

4.2.1. Annotation of Clinical Entities

Our annotation scheme consists of three coarse entity classes corresponding to top-level hierarchies in the SNOMED CT concept model. Detailed definitions of all entity classes can be found in the annotation guide. Based on these top-level concepts, we have defined the following sub classes relevant for GGPONC:

- **Finding**: Diagnosis or Pathology, Other Finding
- **Substance**: Clinical Drug, Nutrient or Body Substance, External Substance
- **Procedure**: Therapeutic, Diagnostic

Since the exact entity boundaries are often ambiguous, we consider the problems of identifying mentions and

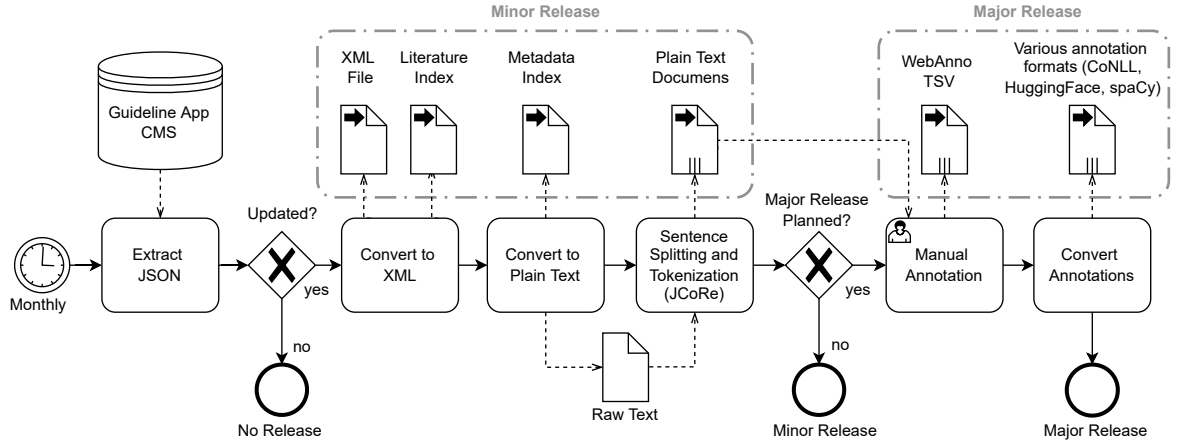


Figure 1: Semi-automated curation process for GGPONC modeled using the Business Process Model and Notation. We regularly check for updates of the guidelines and automatically process the raw JSON files from the content management system (CMS) of the Guideline Program in Oncology to create artifacts with different parts of metadata. These updates are published as minor releases. This workflow has been completely automated through Apache Airflow pipelines. Manual annotation of the resulting documents, as performed as part of this work, results in a major release.

		Guideline	Year	Files	Rec.	Sent.	Tokens	Types	Refs.
1	·	Pancreatic cancer	2013	292	158	854	18,901	3,602	1,154
2	●	Penis cancer	2020	167	94	960	20,915	4,542	561
3	·	Psycho-oncology	2014	121	47	778	21,909	4,113	835
4	○	Oral cavity cancer	2021	132	96	763	22,256	3,947	1,172
5	·	Malignant ovarian tumors	2020	195	97	1,103	27,432	5,139	1,035
6	●	Anal cancer	2020	216	93	1,248	34,429	5,246	724
7	·	Chronic lymphocytic leukemia	2018	285	138	1,417	36,811	5,680	726
8	·	Laryngeal cancer	2019	189	118	1,526	37,374	6,812	681
9	●	Follicular lymphoma	2020	296	149	1,537	38,206	6,344	761
10	·	Oesophageal cancer	2018	172	91	1,530	38,574	6,615	1,026
11	○	Hodgkin lymphoma	2020	253	168	1,710	38,899	5,886	976
12	○	Hepatocellular and biliary cancer	2021	263	146	1,599	41,125	6,552	990
13	·	Testicular tumors	2020	315	163	1,923	47,024	6,746	1,412
14	·	Prevention of cervix cancer	2020	302	103	2,058	52,351	7,928	1,388
15	○	Renal cell carcinoma	2020	293	131	2,284	52,576	8,255	1,507
16	·	Endometrial cancer	2018	317	173	2,005	53,773	8,005	1,340
17	·	Stomach cancer	2019	246	142	2,282	54,281	8,028	1,671
18	●	Adult soft tissue sarcomas	2021	407	228	2,407	56,358	8,785	1,169
19	·	Actinic keratosis	2020	193	74	2,599	57,375	6,853	1,278
20	○	Malignant melanoma	2020	297	167	2,746	65,207	9,241	1,718
21	○	Cervical cancer	2021	415	127	2,829	68,576	9,593	1,549
22	·	Colorectal cancer	2019	546	278	3,021	73,389	9,507	2,446
23	○	Prostate cancer	2021	351	238	3,403	81,899	10,154	2,357
24	·	Supportive therapy	2020	819	337	4,288	96,734	12,369	2,401
25	·	Lung cancer	2018	665	312	4,302	100,930	12,591	2,345
26	○	Breast cancer	2021	685	362	4,232	101,254	12,592	2,831
27	○	Bladder cancer	2020	364	230	4,280	102,192	11,965	2,631
28	○	Prevention of skin cancer	2021	370	141	4,298	106,488	13,817	1,578
29	○	Palliative medicine	2020	700	442	6,029	145,657	15,806	3,117
30	●	Complementary medicine	2021	327	149	8,079	184,205	15,622	3,286
		Total		**10,193**	**5,192**	**78,090**	**1,877,100**	**89,256**	**46,665**

Table 2: Overview of guidelines in the current GGPONC release, sorted by the number of **Tokens**. The number of **Files** corresponds to the number of text segments (recommendations and background text), in contrast to the number of **Rec**ommendations alone. In addition, we report the number of **Sent**ences, unique **Types**, and literature **Ref**erences. GGPONC 2.0 includes five entirely new topics (●) compared to version 1.0. Moreover, 11 guidelines have received a substantial update (○) since the last major release, whereas 14 were not updated at all (·).

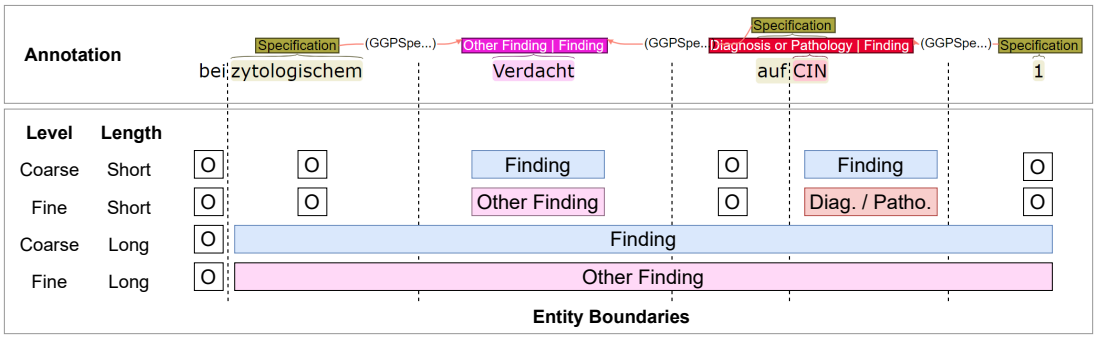
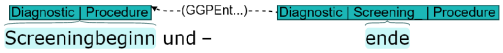
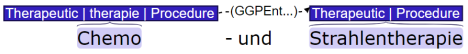


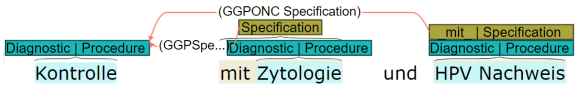
Figure 2: Annotation example for entities and specifications. When we convert the entity annotations to a flat file format with IOB tags, specifications can be handled according to individual requirements. For GGPONC 2.0, we provide two different views on entity spans: the shortest possible spans, where specifications are completely ignored, and the longest possible spans, where all specifications are merged into the class of the head noun. Combined with two different levels of granularity (coarse or fine), this results in four different dataset configurations.



(a) Compound with fragment for forward ellipsis



(b) Compound with fragment for backward ellipsis



(c) Ellipses outside of compound without fragment

Figure 3: Annotation examples for ellipses in CNPs. The omitted parts on the left / right side are annotated as prefix / suffix attributes of the entity span. When parts of a compound have been omitted, we link from the incomplete conjunct to the corresponding complete conjunct with a *fragment* relation.

their boundaries separately: in general, annotators were instructed to identify the shortest possible span for each entity mention. For each part of the enclosing noun phrase specifying the entity in more detail, we annotate *Specification* spans (similar to our previous work in Hahn et al. (2012)) and connect them through a relation with the head entity as outlined in Fig. 2. Therefore, the purpose of specifications is very similar to the *Modifier* entity class introduced by Patel et al. (2018). While each token typically belongs to only one entity class, specifications can be arbitrarily nested and chained.

4.2.2. Annotation of Elliptical CNPs

Due to the prevalence of elliptical CNPs in GGPONC, we decided to explicitly annotate them for entity mentions by providing the omitted parts through *prefix* or *suffix* attributes (see Fig. 3). We follow Chae et al. (2014) who distinguish *forward ellipses* and *backward ellipses*, as well as their combination in *complex ellipses*. A special case of coordination ellipses that occurs frequently in German are *elliptical compounds* (Aepli and Volk, 2013). These are often, but not always, indicated by a suspensive hyphen ("Ergänzungsstrich"). Examples for different types of ellipses and their resolution annotations are depicted in Table 3. Since the annotated information is sufficient to recover the incomplete conjuncts in CNPs, we have a ground truth for evaluation of automated systems performing this task. Note that we do not annotate ellipses outside of entity mentions, so some elliptical CNPs still fall out of the scope of these conventions.

4.2.3. Annotation Guide and Agreement

We follow the protocol suggested by Roberts et al. (2009) for annotation guide development, where multiple annotators work on the same sets of documents until an acceptable level of stability of the inter-annotator agreement (IAA) is reached. We sampled these documents from GGPONC 2.0 aiming at a representative distribution of entity mentions identified by the dictionary-based approach, which we had successfully implemented for GGPONC 1.0. Afterwards, annotators worked on the full guidelines independently, whilst potential questions were discussed in a group chat.

The first iteration (1a) was performed by 3 annotators with coarse entity classes only. As a follow-up, the first version of our annotation guide was created and finer-grained entity classes were incorporated. The same documents where annotated in the next iteration (1b), followed by a workshop to resolve disagreements and further refine the annotation guide. This process was repeated for two more iterations. For GGPONC, the IAA reached satisfactory levels already after iteration two, a third one was performed to assure the stability of agreement (backed up by the data in Table 4).

For measuring the IAA metric, we used the γ-method (Mathet et al., 2015), implemented in the *pygamma-agreement* package (Titeux and Riad, 2021) which is included in the INCEPTALYTICS toolkit (Hamacher and Zesch, 2021). Computation of γ took 156h for the

	Example	Resolution	*English Translation*
		(1) Forward ellipsis	
a)	Krebs-Vorsorge / -Früherkennung	Krebs-Vorsorge / Krebs-Früherkennung	*cancer prevention / screening*
b)	HPV31, 33, 45 und 51	HPV31, HPV33, HPV45 und HPV51	*HPV31, 33, 45 and 51*
c)	Vitamin C, E und A_1	Vitamin C, Vitamin E und Vitamin A_1	*vitamin C, E and* A_1
		(2) Backward ellipsis	
a)	Chemo- und Strahlentherapie	Chemotherapie und Strahlentherapie	*chemotherapy and radiotherapy*
b)	BRAF- und MEK-Inhibitor	BRAF-Inhibitor und MEK-Inhibitor	*BRAF and MEK inhibitor*
c)	zielgerichtete und Immuntherapien	zielgerichtete Therapien und Immuntherapien	*targeted and immunotherapy*
		(3) Complex ellipsis	
a)	HPV-16- und/oder -18-Positivität	HPV-16-Positivität und/oder HPV-18-Positivität	*HPV-16 and/or -18 positivity*
b)	BRCA1/2-Mutation	BRCA1-Mutation / BRCA2-Mutation	*BRCA1/2 mutation*
c)	Zweitlinien- oder Drittliniensystem- bzw. -chemotherapie	Zweitliniensystemtherapie oder Drittliniensystemtherapie bzw. Zweitlinienchemotherapie oder Drittlinienchemotherapie	*second line or third line systemic or chemotherapy*

Table 3: Examples of different types of ellipses in coordinated compound noun phrases found in GGPONC

small set of files of the agreement sets, since it is based on a complex alignment algorithm and scales poorly with the number of annotators.

Overall IAA reached a value of $\gamma = .94$ and exceeds this value for most fine-grained entity classes. Major sources of disagreement were *Specification* boundaries ($\gamma = .89$) and the distinction between *Other Findings* ($\gamma = .91$) and *Diagnostic Procedures* ($\gamma = .93$). The latter distinction is ambiguous in many cases since terms often refer to either diagnostic procedures or the properties they measure, e.g., "Blutbild" (*complete blood count*). We resolved such cases by defining precedence rules in the annotation guide.

	Iteration			
	1a	**1b**	**2**	**3**
Number of annotators	3	7	7	7
Number of documents	5	5	6	3
Number of sentences	149	149	158	67
Number of tokens	4206	4206	3725	1814
IAA (γ)	.75	.89	.93	**.94**
Specification	.71	.87	**.91**	.89
Finding	.82	.93	.95	**.97**
Diagnosis/Pathology	-	.91	.94	**.96**
Other Finding	-	.85	.87	**.91**
Substance	.92	.99	.98	**.99**
Clinical Drug	-	.97	.98	**1.00**
Nutrient/Body Subs.	-	.99	**.99**	.98
External Substance	-	.96	-	**1.00**
Procedure	.82	.93	.96	**.96**
Therapeutic	-	.95	.96	**.96**
Diagnostic	-	.89	**.98**	.93
IAA ($_u\alpha$)	.56	.71	.79	**.85**

Table 4: IAA across multiple iterations calculated using the γ-measure. We report Krippendorff's α (unitizing, character-based) for comparison, yet the results are less meaningful due to the prevalence of overlapping spans. Note that the same documents were used during iterations 1a and 1b, as fine-grained entity subclasses were consented during the first review round.

Although the use of a simpler and less resource-eager IAA measure, e.g., F_1-score or Krippendorff's α, would have been desirable, these alternatives are unable to properly account for arbitrarily nested and overlapping specifications and entity mentions. Just for comparison reasons, we report Krippendorff's α in Table 4 reaching $\alpha = .85$ in the last iteration.

4.3. Baseline NER Models

We created four datasets from the annotated documents, one for each of the combinations of granularity and span length shown in Fig. 2. Each file, i.e., text segment of a guideline, is randomly assigned to either the training (70 %), development (15 %), or test (15 %) set. Annotations are converted to JSON files with IOB-encoded NER labels compatible with the HUGGINGFACE TRANSFOMER library (Wolf et al., 2020).

For each of the four datasets, we train a standard HUGGINGFACE NER model that consists of a pre-trained BERT encoder and a token classification head (Devlin et al., 2019). We initialize the encoder from the publicly available BERT checkpoint *deepset/gbert-base* (Chan et al., 2020), pre-trained on German texts taken from general and legal domains. Each NER model is trained for 100 epochs, and we keep the final model checkpoint that maximizes the F_1 score on the development set. On a single NVIDIA A40 GPU, one training run takes around 6 h. We manage experimental configurations with the HYDRA framework (Yadan, 2019) and performed a parallel grid search on a cluster with 6 GPUs over the following hyperparameters: learning rate, learning rate schedule, weight decay, and label smoothing. For the final models, we choose the set of hyperparameters that maximize the micro-averaged F_1-score on the development set for each dataset configuration. All HYDRA configurations and optimal hyperparameters for each dataset are published in our GitHub repository (HPI-DHC, 2022).

	Avg. tok./ ment.	Number of entities				Test set NER results					
		Total	Train	Dev	Test	Coarse			Fine		
						P	R	F1	P	R	F1
Short Spans	**1.1**	**246,490**	**171,358**	**36,823**	**38,309**	**.88**	**.90**	**.89**	**.86**	**.87**	**.86**
Finding	1.1	132,756	92,053	19,743	20,960	.88	.90	.89			
Diagnosis or Pathology	1.1	81,380	56,216	12,200	12,964				.90	.91	.91
Other Finding	1.1	51,376	35,837	7,543	7,996				.76	.75	.75
Substance	1.2	24,871	17,288	4,017	3,566	.87	.91	.89			
Clinical Drug	1.2	19,478	13,647	3,105	2,726				.90	.92	.91
Nutrient or Body Subst.	1.3	4,348	2,995	694	659				.69	.73	.71
External Substance	1.2	1045	646	218	181				.66	.54	.59
Procedure	1.1	88,863	62,017	13,063	13,783	.89	.91	.90			
Therapeutic	1.0	61,034	42,675	8,557	9,802				.90	.91	.90
Diagnostic	1.1	27,829	19,342	4,506	3,981				.82	.87	.85
Long Spans	**2.4**	**201,838**	**140,228**	**30,024**	**31,586**	**.75**	**.76**	**.75**	**.72**	**.73**	**.72**
Finding	2.5	105,035	72,860	15,470	16,705	.75	.75	.75			
Diagnosis or Pathology	2.3	60,898	42,098	9,042	9,758				.76	.78	.77
Other Finding	2.7	44,521	31,036	6,469	7,016				.64	.61	.63
Substance	1.9	18,169	12,566	2,976	2,627	.72	.76	.74			
Clinical Drug	2.0	14,092	9,851	2,257	1,984				.71	.76	.73
Nutrient or Body Subst.	1.7	3,277	2,219	547	511				.61	.64	.62
External Substance	1.8	809	500	172	137				.45	.34	.38
Procedure	2.5	77,686	54,121	11,455	12,110	.75	.77	.76			
Therapeutic	2.4	52,887	36,906	7,415	8,566				.76	.76	.76
Diagnostic	2.5	25,354	17,618	4,122	3,614				.68	.74	.71

Table 5: Counts of manually annotated entities (columns 3 to 6) and automatic NER results of four different BERT-based token classification models (columns 7 to 12), evaluated on fine-grained vs. course-grained entity classes, as well as long vs. short spans. Short spans cover mostly single-token entities (1.1 **tok**ens per **ment**ion on average), while long spans can be considerably longer.

5. Results

In the following, we share results and findings from our work on the latest release GGPONC 2.0.

5.1. Corpus and Annotations

GGPONC 2.0 consists of 78,090 sentences with 1,877,100 tokens (see Table 2), an increase of around 40% in tokens compared to GGPONC 1.0. However, the most noticeable increment occurred at the level of medical contents—the number of manually created medical NE annotations increased by a factor of more than three (GGPONC version 1.0 contained 73,799 manually supplied annotations). In total, version 2.0 now contains 246,490 (short span) entity mentions manually annotated in the gold standard. When considering long spans, the number is reduced to 201,838, as entities can be subsumed as parts of specifications of other entities (see Fig. 2). Entity counts for the splits used for training and evaluating the NER models are also shown in Table 5 (columns 4 to 6).

As clinical guidelines are developed around the PICO framework (population, intervention, comparison, outcome) (Boudin et al., 2010), the most common entity mentions are related to populations (*Diagnosis or Pathology*, *Other Finding*), and interventions (*Clinical Drugs*, *Therapeutic Procedures*). Other classes occur only rarely in the data (e.g., only 1,045 mentions of *External Substances*).

5.2. NER Results

As shown in Table 5, the best NER results on the test set are achieved on the dataset configuration with coarse-grained entity classes and short spans ($F_1 = .89$), which only slightly drops when fine-grained entity classes are considered ($F_1 = .86$). Performance is much lower for rare entity classes (e.g., $F_1 = .58$ for *External Substances*) than for frequent ones. The same is true for long entity spans, although the overall discrimination performance is substantially lower ($F_1 = .75/.72$).

To shed light on the performance differences between short and long spans, we depict the error types in Fig. 4. The vast majority of errors introduced by extending span length are (label) boundary errors, whereas the rate of actual false negatives, false positives, and labeling errors is similar or even slightly lower.

5.3. Elliptical CNPs and Impact on NER

Annotators identified 5,264 elliptical CNPs in the whole corpus related to entity mentions in 4,666 sentences, i.e., around 6% of all sentences include entities with ellipses. Out of these, 1,964 are forward ellipses, from which 844 (42.9%) were annotated with a fragment attribute, e.g., are part of an elliptical compound. In contrast, nearly all of the 2,979 backward ellipses are part of a compound (2,739 or 91.9% have a fragment attribute). In addition, there are 282 complex ellipses, from which 199 (70.6%) were part of a compound.

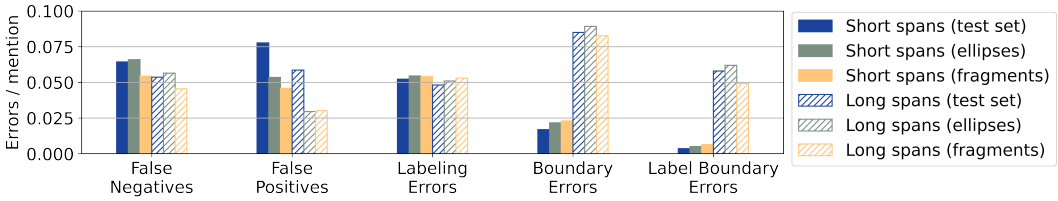


Figure 4: Number of errors per entity mention for short vs. long spans and different kinds of elliptical CNPs in the test set, using the NER models trained on fine-grained entity classes. When considering sequences of tokens, the following errors can occur in addition to *false negatives* and *false positives*: *labeling errors* (the span is predicted correctly, but a wrong label is assigned), *boundary errors* (the correct label is assigned, but boundaries of the predicted span do not match), and *label boundary errors* (both boundaries and label are incorrect).

We also assessed the impact on NER performance in the presence of elliptical CNPs by comparing the error rate on the complete test set with only test set sentences that contain an elliptical CNP with or without a compound fragment. As shown in Fig. 4, there is hardly any difference in the error rate, with only a slight increase of the rate of boundary errors for short spans and even a decrease in the false positive rate. In fact, from all examples in Table 3, the best long-span fine-grained NER model only fails to detect separate entities in example 1b), where a single long entity is predicted instead of two distinct ones, i.e., a boundary error occurs.

6. Evaluation and Discussion

6.1. NER Performance Improvements

While our short-span NER models achieve performance on a par with results reported for clinical NER in the literature (Kittner et al., 2021), performance drops considerably when considering long spans. Part of this discrepancy is explainable by label noise, since IAA has been lowest on *Specification* spans. We also note that a substantial portion of errors are boundary errors (Fig. 4), which may or may not be problematic for different downstream applications, but can disproportionately affect token-based metrics like the F_1-score.

Although the problem of predicting longer spans is inherently harder, methodological improvements seem possible. We have used, e.g., a BERT encoder pre-trained on general domain texts. Language models pre-trained on in-domain texts have shown to be applicable for various downstream tasks (Lee et al., 2020; Bressem et al., 2020; Gu et al., 2022). Yet, a German medical text corpus of the required size was not available for public use up until now. Our work using clinical guidelines may complement such large, unlabeled datasets for unsupervised pre-training in the future.

6.2. Assessment of the γ-measure

We used γ as a measure of IAA for categorical, unitizing and potentially overlapping labels. While, in theory, γ is thus the most appropriate metric, its results are hard to compare with previous studies due to the lack of comparable measurements in other (medical) entity annotation projects. In the few cases where γ was already used with unitizing semantic annotations, values between .24 and .76 were observed (Da San Martino et al., 2019; O'Gorman et al., 2021; Zehe et al., 2021). Our manual curation of annotated documents indeed suggests a much higher agreement, consistent with higher γ values, with slight disagreement for the entity classes where γ is also lowest.

6.3. Resolution of Coordination Ellipses

Error analysis has shown that the impact of elliptical CNPs on NER performance is surprisingly small—an effect which we explain by the adequate representation of this phenomenon in the training data, combined with the flexible encoding of subwords in BERT's WordPiece vocabulary. For downstream tasks, such as entity linking, reliable resolution of ellipses might still be necessary. With the detailed annotations of elliptical CNPs in this work, we may be able to train language models that can perform this task automatically.

7. Conclusion and Future Work

Our work on the new major release of GGPONC 2.0 features high-quality entity annotations, strong baseline NER models, and an analysis of the (surprisingly small) impact of elliptical CNPs on these models. Freely available corpora based on clinical guidelines allow us to study medical language use on a large scale in many different language communities. In the English-speaking world, it has been become common practice to evaluate biomedical NLP approaches on a range of benchmark datasets and tasks (Gu et al., 2022). We envision GGPONC to become part of such a benchmark for German-language biomedical NLP tasks, as well.

We plan to extend our guideline corpus to other medical specialties (besides oncology) to cover an even wider spectrum of the medical domain. Moreover, we will add additional layers of annotation, e.g., grounding entities in SNOMED CT or devising (temporal) relations to derive machine-readable process models from guidelines (Peleg, 2013; Schlegel et al., 2019). GGPONC 2.0 and previous releases are available on demand from the German Guideline Program in Oncology. Our annotation guide and source code are available online to foster the reproducibility of our experimental results (HPI-DHC, 2022).

Acknowledgements This work was generously supported by the German Federal Ministry of Research and Education (BMBF) under grants 01ZZ1802H and 01ZZ1803G. We would like to thank all colleagues from the HiGHmed and Smith consortia of the German Medical Informatics Initiative for their constant support and valuable input contributing to our joint research. We also want to thank our annotators for their hard work, fruitful discussions, and contributions to the annotation guide: Luise Chassol, Jana Esper, Nathalie Gottwald, Robin Kempkens, Saskia Nitschke, Yasmin Stoppe, and Silvia Winkler.

8. Bibliographical References

Aepli, N. and Volk, M. (2013). Reconstructing complete lemmas for incomplete German compounds. In *Language Processing and Knowledge in the Web. GSCL 2013 — Proceedings of the 25th International Conference*, pages 1–13. Springer, Berlin.

Blake, C. L. and Rindflesch, T. C. (2017). Leveraging syntax to better capture the semantics of elliptical coordinated compound noun phrases. *Journal of Biomedical Informatics*, 72:120–131.

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

Borchert, F., Meister, L., Langer, T., Follmann, M., Arnrich, B., and Schapranow, M.-P. (2021). Controversial trials first: identifying disagreement between clinical guidelines and new evidence. In *AMIA 2021 — Proceedings of the 2021 Annual Symposium of the American Medical Informatics Association*, pages 237–246. American Medical Informatics Association (AMIA).

Boudin, F., Nie, J.-Y., and Dawes, M. (2010). Clinical information retrieval using document and PICO structure. In *NAACL-HLT 2010 — Human Language Technologies: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 822–830. Association for Computational Linguistics (ACL).

Bressem, K. K., Adams, L. C., Gaudin, R. A., Tröltzsch, D., Hamm, B., Makowski, M. R., Schüle, C.-Y., Vahldiek, J. L., and Niehues, S. M. (2020). Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics*, 36(21):5255–5261.

Buyko, E., Tomanek, K., and Hahn, U. (2007). Resolution of coordination ellipses in biological named entities using conditional random fields. In *PACLING 2007 — Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 163–171. Pacific Association for Computational Linguistics.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium]*, pages 2633–2650. The USENIX Association.

Chae, J., Jung, Y., Lee, T., Jung, S., Huh, C., Kim, G., Kim, H., and Oh, H. (2014). Identifying non-elliptical entity mentions in a coordinated NP with ellipses. *Journal of Biomedical Informatics*, 47:139–152.

Chan, B., Schweter, S., and Möller, T. (2020). German's next language model. In *COLING 2020 — Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796.

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., and Nakov, P. I. (2019). Fine-grained analysis of propaganda in news article. In *EMNLP-IJCNLP 2019 — Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing*, pages 5636–5646. Association for Computational Linguistics (ACL).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. N. (2019). Bert: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1: Long and Short Papers, pages 4171–4186. Association for Computational Linguistics (ACL).

Donnelly, K. (2006). Snomed-CT: the advanced terminology and coding system for eHealth. In *Medical and Care Compunetics 3*, number 121 in Studies in Health Technology and Informatics, pages 279–290, Amsterdam etc. IOS Press.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2022). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):#2 (2:1–2:23).

Hahn, U., Beisswanger, E., Buyko, E., and Faessler, E. (2012). Active learning-based corpus annotation: the PathoJen experience. In *AMIA 2012 — Proceedings of the 2012 Annual Symposium of the American Medical Informatics Association. Informatics: Transforming Health and Health Care*, pages 301–310, Philadelphia/PA. Hanley & Belfus.

Hahn, U., Matthies, F., Faessler, E., and Hellrich, J. (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*, pages 2502–2509.

Hamacher, M. and Zesch, T. (2021). INCEpTALYTICS 0.1.0. `https://github.com/ltl-`

ude/inceptalytics/. (Last accessed: April 26th, 2022) DOI: 10.5281/zenodo.5654690.

HPI-DHC. (2022). GGPONC annotation repository. https://github.com/hpi-dhc/ggponc_annotation. (Last accessed: April 27th, 2022) DOI: 10.5281/zenodo.6473122.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., and Gurevych, I. (2018). The INCEPTION platform: machine-assisted and knowledge-oriented interactive annotation. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. International Committee on Computational Linguistics (ICCL).

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C., and Kang, J. (2020). BIOBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mathet, Y., Widlöcher, A., and Métivier, J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Meystre, S. M. (2015). De-identification of unstructured clinical data for patient privacy protection. In Aris Gkoulalas-Divanis et al., editors, *Medical Data Privacy Handbook*, pages 697–716. Springer, Cham, Switzerland.

Montani, I., Honnibal, M., Landeghem, S. V., Boyd, A., Peters, H., Samsonov, M., Geovedi, J., McCann, P. O., Regan, J., Orosz, G., Altinok, D., Kristiansen, S. L., Roman, R., Fiedler, L., Howard, G., Phatthiyaphaibun, W., Tamura, Y., Bot, E., Bozek, S., Murat, M., Amery, M., Böing, B., Tippa, P. K., Vogelsang, L. U., Balakrishnan, R., Mazaev, V., Dubbin, G., Fukumaru, J., and Henry, W. (2021). EXPLOSION/SPACY: V3.1.0: new pipelines for Catalan & Danish, SPANCATEGORIZER for arbitrary overlapping spans, use predicted annotations during training, bug fixes & more. https://doi.org/10.5281/zenodo.5079800 (Last accessed: April 26th, 2022).

O'Gorman, T., Jensen, Z., Mysore, S., Huang, K., Mahbub, R., Olivetti, E., and McCallum, A. K. (2021). MS-Mentions: consistently annotating entity mentions in materials science procedural text. In *EMNLP 2021 — Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1337–1352. Association for Computational Linguistics (ACL).

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

Richter-Pechanski, P., Geis, N. A., Kiriakou, C., Schwab, D. M., and Dieterich, C. (2021). Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. *Digital Health*, 7:1–10.

Schlegel, D. R., Gordon, K., Gaudioso, C., and Peleg, M. (2019). CLINICAL TRACTOR: a framework for automatic natural language understanding of clinical practice guidelines. In *AMIA 2019 — Proceedings of the 2019 Annual Symposium of the American Medical Informatics Association. Informatics: From Data to Knowledge to Action*, pages 784–793. American Medical Informatics Association (AMIA).

Titeux, H. and Riad, R. (2021). pygamma-agreement: gamma γ measure for inter/intra-annotator agreement in PYTHON. *Journal of Open Source Software*, 6(62):#2989.

Wei, C.-H., Leaman, R., and Lu, Z. (2015). SIMCONCEPT: a hybrid approach for simplifying composite named entities in biomedical text. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1385–1391.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: state-of-the-art natural language processing. In *EMNLP 2020 — Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Systems Demonstrations*, pages 38–45. Association for Computational Linguistics (ACL).

Yadan, O. (2019). Hydra - a framework for elegantly configuring complex applications. GitHub. https://github.com/facebookresearch/hydra (Last accessed: April 26th, 2022).

Zehe, A., Konle, L., Dümpelmann, L., Gius, E., Hotho, A., Jannidis, F., Kaufmann, L., Krug, M., Puppe, F., Reiter, N., Schreiber, A., and Wiedmer, N. (2021). Detecting scenes in fiction: a new segmentation task. In *EACL 2021 — Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3167–3177. Association for Computational Linguistics (ACL).

9. Language Resource References

Borchert, F., Lohr, C., Modersohn, L., Langer, T., Follmann, M., Sachs, J. P., Hahn, U., and Schapranow, M.-P. (2020). GGPONC: a corpus of German medical text with rich metadata based on clinical practice guidelines. In *LOUHI 2020 — Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis @ EMNLP 2020*, pages 38–48. Association for Computational Linguistics (ACL).

Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI Disease Corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Hahn, U., Matthies, F., Lohr, C., and Löffler, M. (2018). 3000PA: towards a national reference corpus of German clinical language. In *Building Continents*

of Knowledge in Oceans of Data: The Future of Co-Created eHealth. MIE 2018 — Proceedings of the 29th Conference on Medical Informatics in Europe, number 247 in Studies in Health Technology and Informatics, pages 26–30, Amsterdam. IOS Press.

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

Kittner, M., Lamping, M., Rieke, D. T., Götze, J., Bajwa, B., Jelas, I., Rüter, G., Hautow, H., Sänger, M., Habibi, M., Zettwitz, M., de Bortoli, T., Ostermann, L., Ševa, J., Starlinger, J., Kohlbacher, O., Malek, N. P., Keilholz, U., and Leser, U. (2021). Annotation and initial evaluation of a large annotated German oncological corpus. *JAMIA Open*, 4(2):ooab025.

Lohr, C., Modersohn, L., Hellrich, J., Kolditz, T., and Hahn, U. (2020). An evolutionary approach to the annotation of discharge summaries. In *Digital Personalized Health and Medicine. MIE 2020 — Proceedings of the 30th Conference on Medical Informatics Europe*, number 270 in Studies in Health Technology and Informatics, pages 28-32, Amsterdam etc. IOS Press.

Miñarro-Giménez, J. A., Cornet, R., Jaulent, M. C., Dewenter, H., Thun, S., Rosenbeck Gøeg, K., Karlsson, D., and Schulz, S. (2019). Quantitative analysis of manual annotation of clinical text samples. *International Journal of Medical Informatics*, 123:37–48.

Nye, B. E., Li, J. J., Patel, R., Yang, Y., Marshall, I. J., Nenkova, A., and Wallace, B. C. (2018). A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 197–207. Association for Computational Linguistics (ACL).

Patel, P., Davey, D., Panchal, V., and Pathak, P. (2018). Annotation of a large clinical entity corpus. In *EMNLP 2018 — Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042. Association for Computational Linguistics (ACL).

Read, J. L., Velldal, E., Cavazza, M., and Georg, G. (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*, pages 1724–1731. European Language Resources Association (ELRA).

Roberts, A., Gaizauskas, R. J., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., and Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950–966.

Schulz, S., Ševa, J., Rodriguez, S., Ostendorff, M., and Rehm, G. (2020). Named entities in medical case reports: corpus and experiments. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 4495–4500. European Language Resources Association (ELRA).

van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J. A., and Furlong, L. I. (2012). The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884.

Wang, Y. (2009). Annotating and recognising named entities in clinical notes. In *Proceedings of the Student Research Workshop @ ACL-IJCNLP 2009*, pages 18–26. Association for Computational Linguistics (ACL).

Wermter, J. and Hahn, U. (2004). An annotated German-language medical text corpus as language resource. In *LREC 2004 — Proceedings of the 4th International Conference on Language Resources and Evaluation. In Memory of Antonio Zampolli*, volume 2, pages 473–476. European Language Resources Association (ELRA).