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]

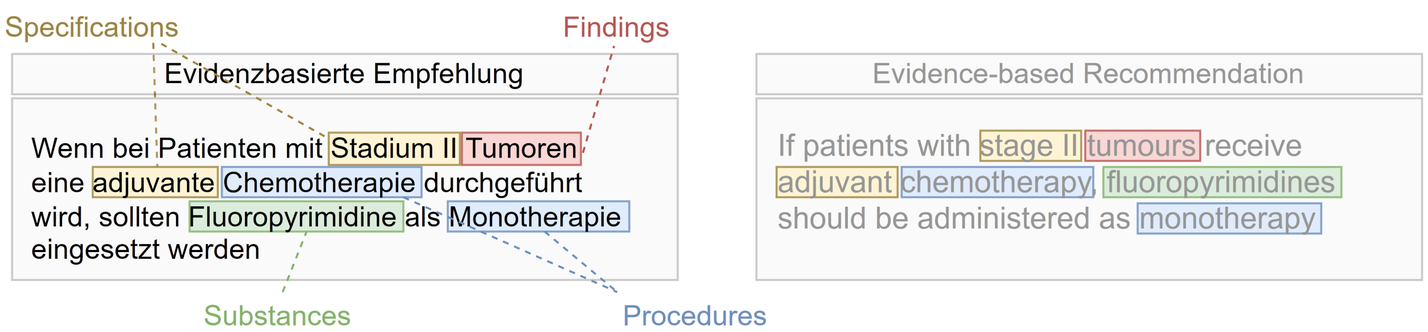
[1] HPI Digital Health Center, Hasso Plattner Institute, Potsdam
[2] JULIE Lab, Friedrich Schiller University, Jena
[3] German Guideline Program in Oncology, German Cancer Society, Berlin
[4] Peter L. Reichertz Institute for Medical Informatics, Hannover

1. Motivation

Most medical texts contain sensitive data, inaccessible for NLP researchers

Clinical guidelines are freely distributable and available in many languages!

- Limited access to clinical texts with protected health information (PHI)
- Other medical text, e.g., scientific publications, mostly available in English



Excerpt from a German clinical guideline on colorectal cancer, containing medical terminology, but no PHI

2. Corpus

GGPONC 2.0 is the largest annotated, freely distributable German medical text corpus

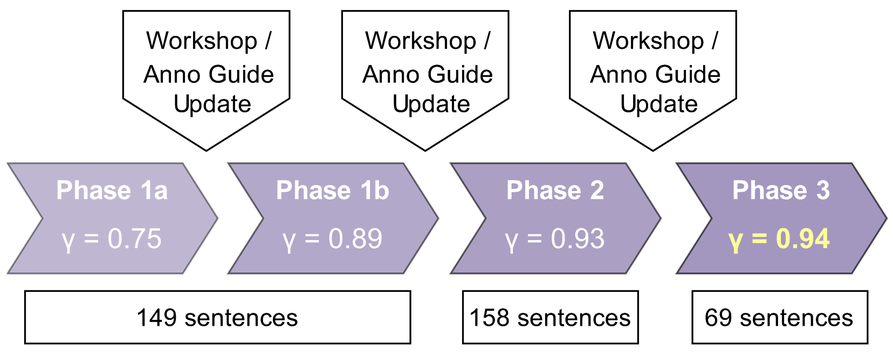
- Build upon 30 German oncology guidelines
- 1.8M+ tokens
- 250k+ entity annotations
- 5k+ elliptical coordinated noun phrases (CNPs) + resolution

Accessibility

- Free to use for research
- Get access here →

3. Annotation

Iterative Refinement of Annotation Guide

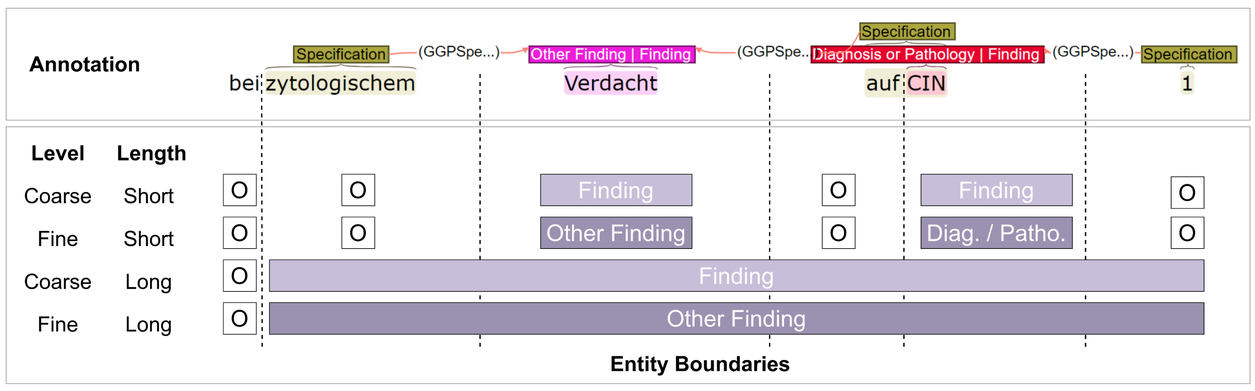


- Personnel:
  - 7 annotators (medical students)
  - 1 curator (medical doctor)
- Time: 6 months, approx. 1.2k hours
- Software: INCEpTION (open source)
- Very high inter-annotator agreement (γ = 0.94)

SNOMED-CT-oriented Annotation Scheme

	Finding	Procedure	Substance
Coarse			
Fine	▪ Diagnosis / Pathology ▪ Other Finding	▪ Therapeutic ▪ Diagnostic	▪ Clinical Drug ▪ Nutrient / Body Substance ▪ External Substance

Four Dataset Configurations



- Short spans = head word only
- Long spans = including all specifications / modifiers

4. Baseline Named Entity Taggers

- 1 HuggingFace token classification model per dataset configuration
- Initialized with GBERT weights (German BERT pre-trained on general-domain text)
- Additional spaCy SpanCategorizer model for long, overlapping spans

Classifier Performance

F1 Score	Coarse	Fine
Short	.89	.86
Long	.75	.72

Error Analysis

- Long vs. short span models: large performance difference
- Mostly due to boundary errors

SPONSORED BY THE
Federal Ministry of Education and Research

This work was generously supported by the German Federal Ministry of Research and Education (BMBF) under grants (01ZZ1802H, 01ZZ1803G)

Contact
Florian Borchert
Dr. Matthieu-P. Schapranow
Digital Health Center
Hasso Plattner Institute
University of Potsdam
Prof.-Dr.-Helmert-Str. 2-3
14482 Potsdam, Germany