

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. Schapranow¹

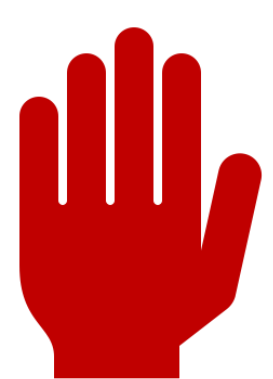
¹ HPI Digital Health Center, Hasso Plattner Institute, Potsdam

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

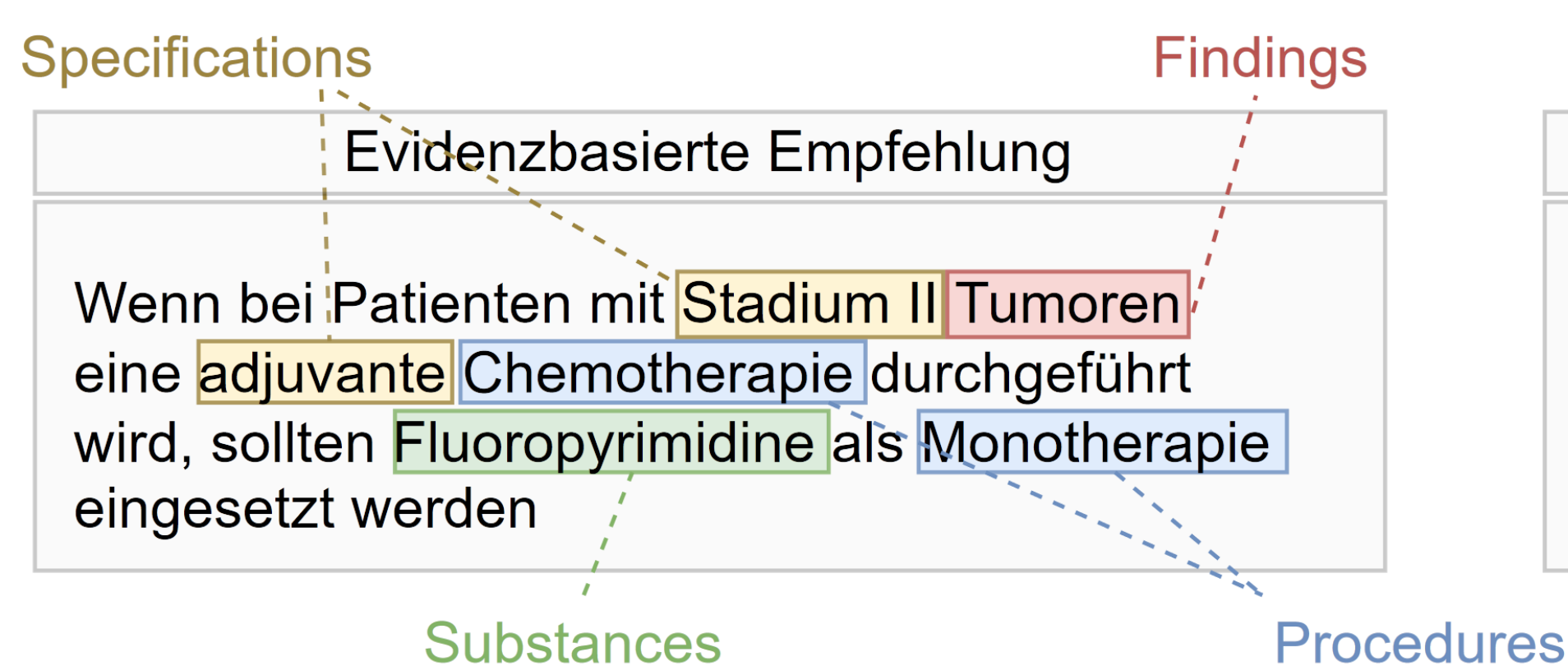


Most medical texts contain sensitive data, inaccessible for NLP researchers



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

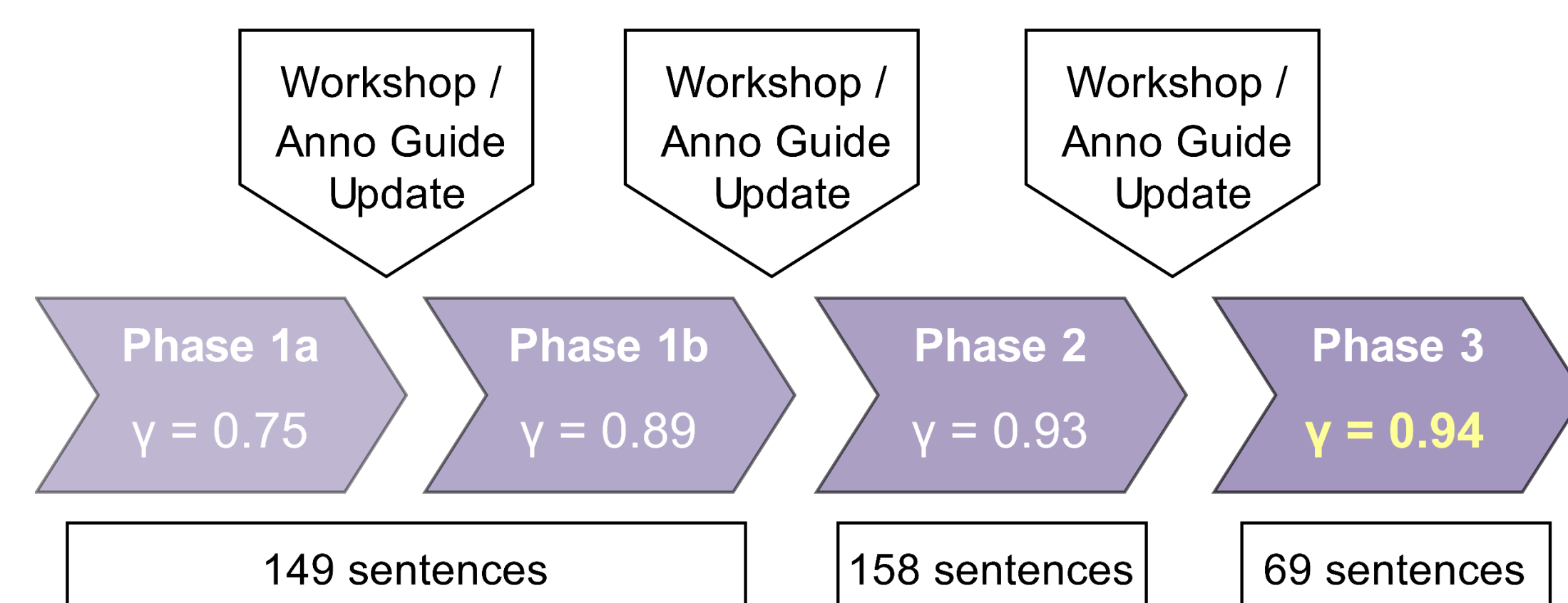
Accessibility

- 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 ($\gamma = 0.94$)

SNOMED-CT-oriented Annotation Scheme

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

Four Dataset Configurations

Annotation	bei:zytologischem	Verdacht	auf:CIN	1
Level	Length			
Coarse	Short	☐	☐	☐
Fine	Short	☐	☐	☐
Coarse	Long	☐	☐	☐
Fine	Long	☐	☐	☐

Entity Boundaries

- 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



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



Digital Engineering · Universität Potsdam