Developing the German Medical Text Corpus (GeMTeX): Legal Compliance and Semantic Enrichment

Justin Hofenbitzer[*], Christina Lohr[†], Andrea Riedel[‡,§], Rebekka Kiser[*], Aliaksandra Shutsko[¶,α], Abanoub Abdelmalak[¶,α], Peter Klügl[‖], Jutta Romberg[β,γ], Sarah Riepenhausen[ϵ], Miriam Schechner[ζ], Jakob Faller[‡,§], Frank Meineke[†], Luise Modersohn[*], Markus Löffler[†], Juliane Fluck[¶,α], Udo Hahn[†], Stefan Schulz[‖,η], Martin Boeker[*]

[*]Technical University of Munich, Germany, [†]Leipzig University, Germany, [‡]FAU Erlangen-Nürnberg, Germany, [§]Uniklinikum Erlangen, Germany, [¶]ZB MED, Germany, [‖]Averbis GmbH, Germany, [α]University of Bonn, Germany, [β]Charité-Universitätsmedizin Berlin, Germany, [γ]Berlin Institute of Health @ Charité, Germany, [ϵ]University of Münster, Germany, [ζ]LMU University Hospital Munich, Germany, [η]MedUni Graz, Austria

justin.hofenbitzer@tum.de, martin.boeker@tum.de

Abstract

GeMTeX is a large-scale German Medical Text Corpus project with the goal to publish a clinical national reference corpus. The resource is currently under construction and comprises, as of February 2026, more than 15k clinical documents (20M tokens) from six German university hospitals. When building GeMTeX, attention was paid to comply with European regulatory requirements. In phase I, patients were asked to allow reuse of their clinical documents based on the legal foundation of an "informed consent". In phase II, consented documents from six major clinical sites in Germany underwent a thorough de-identification process. In phase III, we currently enrich this unlocked dataset with semantic information from the clinical domain. This annotation process is guided by Snomed CT, which supports to directly ground expressions within clinical documents in a worldwide shared medical documentation and ontology standard. The resource is currently under active development and is accessible upon request under controlled access conditions. We refer interested researchers to visit https://kiinformatik.mri.tum.de/en/gemtex or reach out via gemtex.mi@mh.tum.de.

Keywords: GeMTeX, clinical text corpus, data privacy, de-identification, semantic annotation, SNOMED CT

1. Introduction

Creating new clinical text corpora for natural language processing (NLP) is fundamentally constrained by complex organizational and legal regulations. These barriers have led to a well-documented scarcity of clinical data, an issue that is especially evident in non-English contexts (Névéol et al., 2018; Hahn, 2025).

To address this gap, we present the development of the new *German Medical Text Corpus* (GeMTeX), a large-scale German-language clinical text resource[1] (Meineke et al., 2023; Faller et al., 2025). The corpus compilation is embedded into the *Medical Informatics Initiative* (MII), a nationwide German medical informatics consortium (Semler et al., 2018). The GeMTeX project brings together 18 partners, of which six are university hospitals responsible for data acquisition and annotation: *TUM Klinikum Munich*, *Universitätsklinikum Leipzig*, *Uniklinikum Erlangen*, *Charité Berlin*, *Universitätsklinikum Carl Gustav Carus Dresden*, and *Universitätsklinikum Essen*. This multi-site approach helps minimize the influence of local jargons often found in single-site clinical document collections (Rodriguez-Esteban, 2009) and ensures a more representative linguistic and medical diversity across institutions and clinical domains. We expect a total size of about 18k real-world clinical documents from diverse medical departments, enriched with medically meaningful semantic annotations grounded in Snomed CT[2]. Beyond its research value, GeMTeX also aims to serve as a blueprint for legally compliant, seman-

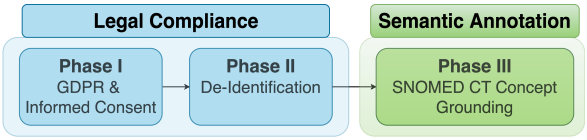

Figure 1: The three phases of legally compliant and semantically enriched clinical corpus creation according to GeMTeX's workflow.

[1] https://www.smith.care/en/gemtex_mii/about-gemtex/

[2] https://www.snomed.org/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1571–1584
11–16 May 2026.

tically interoperable clinical corpus creation under European data-protection frameworks.

In this paper, we describe three key phases during the corpus compilation (Figure 1): The first part of the paper describes phase I, the legal compliance process (Section 3.1), and phase II, its practical realization in the form of a thorough de-identification workflow (Section 3.2). Both phases are based on the legal requirements of the European *General Data Protection Regulation* (GDPR)[3].

In the second part of this paper, we then turn to phase III, the semantic annotation of such an unlocked clinical data resource (Section 4). This part lays out the medical annotation scheme of the GeMTeX annotation campaign. We opted for SNOMED CT as the conceptual backbone of the annotation process, as this widely adopted terminology lays the foundation for language-independent interoperability of annotated clinical corpora since medical concepts are given unique identifiers, irrespective of the language in which the documents are written.

Finally, we overview the current status of the corpus (Section 5) and detail the latest accessibility options (Section 6).

2. Related Work

We begin by providing a detailed overview of existing clinical corpora for English and German (Section 2.1), before we then survey corpora that contain semantic annotations grounded in SNOMED CT (Section 2.2).

2.1. Existing Clinical Corpora

Clinical NLP for English offers a reasonable, though limited, number of corpus resources. Under the HIPAA Safe Harbor rules[4], US privacy law defines 18 types of *personally identifiable information* (PII) that must be removed before public release. Once de-identified, such data may be shared under comparatively liberal *data use agreements* (DUA). The most prominent example is MIMIC (Johnson et al., 2023b), whose MIMIC-IV version contains de-identified discharge summaries of almost 332k patients from the *Beth Israel Deaconess Medical* Center (Johnson et al., 2023a). In addition, several English-language clinical NLP challenges, such as I2B2, N2C2 (Kumar et al., 2015; Stubbs et al., 2019), and tasks within SEMEVAL (Pradhan et al., 2014; Bethard et al., 2015; Elhadad et al., 2015; Bethard et al., 2016, 2017), have yielded corpora with diverse annotation layers, e.g., for de-identification (Stubbs and Uzuner, 2015; Stubbs et al., 2017), named entities (Uzuner et al., 2011; Pradhan et al., 2014; Elhadad et al., 2015; Henry et al., 2020a,b; Mahajan et al., 2023), and semantic relations (Uzuner et al., 2011; Henry et al., 2020b; Mahajan et al., 2023).

Névéol et al. (2018) highlighted the scarcity of clinical corpora for non-English languages, a gap that persists, especially for German. In addition to two distributable German corpora noted by Hahn (2025), we identified two more, resulting in a total of only four accessible real-world clinical text resources: BRONCO (Kittner et al., 2021) contains 150 discharge summaries (75k tokens) from *Charité Berlin* and *Universitätsklinikum Tübingen*, annotated with ICD-10[5], OPS[6], and ATC[7]. However, its sentences were arbitrarily shuffled, limiting realism. CARDIO:DE's (Richter-Pechanski et al., 2023) creation was based on patients' informed consent and comprises 500 cardiology reports (993k tokens) from *Universitätsklinikum Heidelberg*, annotated for section headings and medications but without terminological grounding. Lenz et al. (2025) describe a third corpus of 150 oncological note snippets from *Universitätsmedizin Mainz* (31k tokens) annotated with ICD-10 codes, while Dewald et al. (2023) report a consent-based set of 1.9k short radiology reports from *Hanover Medical School* (estimated 200k tokens) without annotations. The latter two are freely available, whereas BRONCO and CARDIO:DE require a formal DUA.

2.2. Resources with SNOMED CT Concept Annotations

Annotations of clinical corpora grounded in SNOMED CT become increasingly available. Most prominent among them is the English MCN corpus (Luo et al., 2019), which arose from the fourth I2B2/VA task (Uzuner et al., 2011) and uses RXNORM[8], and SNOMED CT to ground medications and other clinical facts, respectively (13.7k annotations of 3.7k concepts). Suominen et al. (2013) and Pradhan et al. (2015) leveraged SNOMED CT for the annotation of disorders in the MIMIC-II[9] data (298 documents, 11k disorder annotations) within the SHARE/CLEF EHEALTH EVALUATION LAB 2013, the SEMEVAL-2014

[3] https://eur-lex.europa.eu/eli/reg/2016/679

[4] https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html

[5] https://www.who.int/standards/classifications/classification-of-diseases

[6] https://www.bfarm.de/EN/Code-systems/Classifications/OPS-ICHI/OPS/_node.html

[7] https://www.who.int/tools/atc-ddd-toolkit/atc-classification

[8] https://www.nlm.nih.gov/research/umls/rxnorm/index.html

[9] https://physionet.org/content/mimic-ii/2.6.0/

Task 7 (Pradhan et al., 2014), and the SemEval-2015 Task 14 (Elhadad et al., 2015). Also recently, Park et al. (2024) introduced the Camir corpus that consists of 609 English radiology reports annotated with 87 distinct Snomed CT concepts. The largest dataset for Snomed CT based term grounding was created during the Snomed CT entity linking challenge (Davidson et al., 2025) (272 documents, 74.8k annotations spanning 6.6k unique concepts).

Non-English clinical corpora using Snomed CT for concept grounding are reported by Skeppstedt et al. (2012) using Swedish clinical notes (26k tokens, 2.3k annotations) for three semantic classes, namely Disorder, Finding, and Body Structure. González-Agirre et al. (2019) deal with 1k Spanish case reports (400k tokens, 7.6k annotations) and Miranda-Escalada et al. (2022) introduce the multilingual DisTEMIST corpus comprising 1k clinical texts as part of the BioASQ challenge 2022. Borchert et al. (2022) annotated the *German Guideline Program in Oncology* corpus (GGPOnc) with three Snomed CT rooted top-level hierarchies including findings, substances, and procedures.

Uma and Moens (2024) discuss cross-lingual disease and image classification in X-ray reports using a lightweight Snomed CT grounded graph-based embedding method designed specifically for radiology reports. This study uses the conceptual structure of Snomed CT to create sense embeddings, instead of drawing them from an annotated corpus. For recent surveys on the incorporation of Snomed CT into medical large language models (LLMs) and the generation of knowledge graph embedding models grounded in the ontology, see Agarwal et al. (2019), Chang et al. (2020), and Chang and Sung (2024).

3. Legal Framework for GeMTeX

The goal of the GeMTeX project is to create a novel, large-scale, and multi-centric clinical text corpus, comprising different document types from several medical domains. As a prerequisite to make this resource distributable, legal compliance with European data privacy laws is required, and our approach is discussed in more depth in Section 3.1. Implementing law and regulations, i.e., the reliable de-identification of sensitive personal data in the text documents, is tackled in Section 3.2.

3.1. Phase I: European Data Privacy Requirements

Medical information, thus also clinical texts, fall under the so-called *special categories of personal data* according to Art. 9(1) GDPR. This type of data is subject to increased requirements for data processing. A lawful treatment of health data generally requires a specific legal provision or informed consent of the data subject, i.e., the patient (Art. 9(2) GDPR). Article 89 GDPR further stipulates that appropriate safeguards must be provided for the rights and freedoms of the data subject if it is scientifically processed. These safeguards include, in particular, technical and organizational measures as well as compliance with the principle of data minimization, e.g., through pseudonymization or anonymization. Pseudonymization enables the healthcare institutions that contribute data to trace the data back to its source in individual cases, with the advantage that data quality can be verified. This is a specific requirement of Art. 10 EU AI Act[10] for developing high-risk artificial intelligence (AI) systems, which generally include AI-based medical devices.

Article 9(2)(j) GDPR allows member states to create specific legal bases that enable processing health data for scientific purposes. While there are various such legal bases in Germany for research usage of health data, e.g., within healthcare facilities, there is no legal basis for sharing pseudonymized health data with third parties. Therefore, a nationwide standardized form of the informed consent, the MII Broad Consent (MBC) (Bild et al., 2020; Zenker et al., 2024a,b), was used to enable a legally robust foundation for researchers to work with sensitive clinical data. In particular, by signing the MBC, patients agree to the use of their data in diverse health-related scientific studies if the data undergo the aforementioned minimization strategies and meet the high data security and regulatory standards established within the MII.

In the GeMTeX project, all partnering sites are committed to only process such clinical documents, of which the respective patient has signed the bespoke MBC. Furthermore, the university hospitals agreed to store patient data in a decentralized, pseudonymized form. Based on this legal framework, the acquisition of clinical documents for the GeMTeX corpus additionally required that all participating sites obtain a valid ethics approval for processing and compiling the data under the MBC. To that end, the *TUM Klinikum* obtained approval from its local institutional review board (IRB), which was subsequently acknowledged by the IRBs of the other participating sites, following the *1 study, 1 vote* principle of the *Bundesärztekammer* [German Medical Association][11] and the *Arbeitskreis Medizinischer Ethik-Kommissionen in der Bundesrepublik*

[10] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

[11] https://www.bundesaerztekammer.de/presse/aktuelles/detail/eine-studie-ein-votum

Deutschland e.V. [*Working Group of Medical Ethics Committees in the Federal Republic of Germany*][12].

3.2. Phase II: Implementation of European Privacy Regulations - De-Identification

To meet the European data privacy regulations previously described, the GeMTeX project implements a detailed data minimization process. The clinical document files are stored in a decentralized and pseudonymized form in the participating healthcare facilities, and the text itself undergoes a defined de-identification process, respecting the HIPAA Safe Harbor Principles (Lohr et al., 2024, 2025).

According to our de-identification scheme, Table 1 summarizes the high-level PII entity types that must be removed from the clinical text documents. Notably, we included a category *other*, which allows annotators to mark information not falling under the predefined PII categories. A typical example is a mention indicating that a patient is known as, or related to, a public figure, such as a politician, celebrity, or their relatives or friends.[13]

PII	Example
Name	*We report on **Mary Jones**.*
Date	*She passed away on **December 1, 2024**.*
Age	*At the age of **92**, she received a pacemaker.*
Location	*She was found unconscious in front of her flat in **Munich**.*
ID	*She was brought to room **3.10**.*
Contact	*Her mother's phone number is **4439**.*
Profession	*She works as a **baker**.*
Other	*The patient is **married to the German chancellor**.*

Table 1: Overview of the GeMTeX de-identification annotation scheme with eight main categories.

The de-identification process is organized in a semi-automated fashion, using the INCEpTION annotation platform (Klie et al., 2018; Eckart De Castilho et al., 2024) at all contributing sites: Two independent annotators mark all instances of PIIs with the corresponding labels throughout every document. The human annotators are supported by INCEpTION's string matcher, which suggests annotations to text spans, which have previously been annotated. In addition, we leveraged a recommendation system powered by the commercial product Health Discovery[14] by Averbis that produces on-the-fly suggestions as the annotators process the documents. All annotators are medical students in at least their second year. They underwent a training process to reach a well-defined quality check point, a site-level Krippendorff's α (Krippendorff, 1970) of at least 0.9 on the synthetic GraSCCo corpus (Modersohn et al., 2022).[15] All annotators receive fair compensation for their work, i.e., they are employed as student assistants. In addition to the annotators, an independently installed third person curates the annotation results, resolves disagreements, and provides feedback about error patterns to the annotators.

Once the annotation of a document is complete, all annotated spans are replaced by type-sensitive placeholders. Example (1) illustrates how an original sentence (1-a) is transformed into one with a placeholder (1-b). We decided to exclude documents containing de-identification annotations of the type *other* from GeMTeX because we think these cases represent a disproportionately high risk of re-identification. At the same time, they are hard to replace without harming the structure of a document.[16]

(1) a. Bericht zu Fr. **Maier**.
*Report about Ms. **Maier**.*
b. Bericht zu Fr. **[[**NAME P23**]]**.
*Report about Ms. **[[**NAME P23**]]**.*

4. Phase III: Semantic Annotation with Snomed CT

Phases I and II in the creation of GeMTeX tackled the legal foundation for the acquisition of clinical documents within the project and the detailed three-pass de-identification process. On this foundation, we will turn to phase III, the semantic annotation of the de-identified clinical documents using the medical ontology-based terminology Snomed CT. Crucially, the annotation scheme does not focus on high-level categories of named entities like in traditional annotation campaigns, but applies the direct annotation of concepts contained in Snomed CT, similar to the work done by the AIDAVA project (Schulz et al., 2023) or the requirements of the Snomed CT challenge (Davidson et al., 2025).

Section 4.1 introduces Snomed CT and its conceptual framework, followed by the outline of the concept-oriented semantic annotation scheme de-

[12] https://www.akek.de/sonstige-studien/

[13] Access the GeMTeX de-identification guidelines for a comprehensive overview: https://doi.org/10.5281/zenodo.11502328.

[14] https://averbis.com/health-discovery/

[15] Access the gold-standard de-identification layer for the GraSCCo corpus: https://doi.org/10.5281/zenodo.11502328.

[16] We released our surrogation tool Surrogator for public usage: https://github.com/medizininformatik-initiative/GeMTeX/tree/main/surrogator.

veloped in the GeMTeX project in Section 4.2. Section 4.3 sheds light on a set of annotation maxims developed to streamline the cognitively demanding annotation process, described in Section 4.4, along with preliminary reports about inter-annotator agreements.

4.1. SNOMED CT

The international standard SNOMED CT is an ontology-based clinical terminology designed for detailed clinical documentation, comprising about 370k units of meaning, named SNOMED CT concepts. Unlike classification systems such as ICD, which are primarily used for billing and population statistics, SNOMED CT is more granular and compositional, suited for representing nearly all clinical facts of interest in electronic health records, such as clinical conditions, procedures, lab values, medications, organisms, chemicals, devices, but also administrative healthcare processes (cf. Benson and Grieve, 2016; Bird, 2025).

Each SNOMED CT concept has a unique *identifier*, a unique English label, called *fully-specified name* (FSN), and a *semantic tag*. The identifier is displayed as code, e.g., *233604007*. The FSN is the human-readable concept description, e.g., *Pneumonia*. Moreover, the FSN is linked to synonyms and clinical terms in other languages and is, thus, an ideal target for NLP tasks such as concept normalization. The semantic tag is always shown in parentheses after the FSN and expresses a concept's categorization, e.g., *disorder*, *finding*, or *body structure*. For instance, the concept *'67079006 |Glucose (substance)|'* is a taxonomic descendant of the concept *'105590001 |Substance (substance)|'*, which is indicated by the semantic tag in parentheses.

SNOMED CT's backbone is 19 multiple inheritance taxonomies, enhanced by formal definitions that use the description logic EL++ (Baader et al., 2005)[17]. An example of this poly-hierarchical structure is that all instances of the concept *'75570004 |Viral pneumonia (disorder)|'* are instances of both the concept *'233604007 |Pneumonia (disorder)|'* and the concept *'312134000 |Viral lower respiratory infection (disorder)|'*). Compositionality is an atomic component of SNOMED CT: Some concepts in the terminology come already *pre-coordinated*, i.e., their meaning is defined by explicit links to other existing SNOMED CT concepts. In addition, all concepts can be subject to *post-coordination*, i.e., the post-hoc composition of existing concepts using certain pre-defined relations to arrive at a more expressive denotation.

[17]Nearly all SNOMED CT concepts correspond to OWL classes, some few to metadata annotations and to object properties, i.e., binary relations.

4.2. Different Roles of SNOMED CT Concepts in Semantic Annotation

The semantic annotation scheme of the GeMTeX project aims at high medical accuracy while keeping the cognitive load for the annotators as low as possible (Hofenbitzer et al., 2025).[18] Therefore, we decided to define three major roles of SNOMED CT concepts guiding the semantic annotation: *Core Concepts* (Section 4.2.1), *Modifier Concepts* (Section 4.2.2), and *Qualifier Concepts* (Section 4.2.3). Table 2 illustrates how we assigned SNOMED CT concepts to these three high-level roles by their semantic tag. To simplify the annotation even further, we decided to exclude entire branches of SNOMED CT from the annotation process (Table 5). For the GeMTeX corpus we use the SNOMED CT international version from April 2024.

Concept Role	Semantic Tag(s)
Core Concepts	*disorder*, *event*, *finding*, *clinical drug*, *medicinal product*, *medicinal product form*, *substance*, *observable entity*, *product*, *procedure*, *regime/therapy*
Modifier Concepts	*body structure*, *cell*, *cell structure*, *specimen*, *morphologic abnormality*, *organism*, *physical object*
Qualifier Concepts	*qualifier value*, *disposition*, *administration method*, *basic dose form*, *role*, *dose form*, *intended site*, *number*, *product name*, *release characteristic*, *state of matter*, *transformation*, *supplier*, *unit of presentation*

Table 2: Overview of the assignment of semantic tags to the three roles of SNOMED CT concepts defined by the GeMTeX annotation scheme.

4.2.1. Core Concepts

The primary characteristic of core concept annotations is that they do not require a combined interpretation with other concept annotations, i.e., a core concept alone already denotes a clinical fact. The concepts belonging to *clinical procedures*, *conditions*, or *observables* meet that requirement, because instances of these concepts are always *true* or *present* for a patient given the context of a clinical letter. For instance, in utterance (2), it is not necessary to add other concept annotations, because the interpretation of the SNOMED CT

[18]Access the GeMTeX Semantic Annotation Guidelines: https://doi.org/10.5281/zenodo.15689930.

concept displayed in (2-a) is enough to understand that the patient suffers from anarthria.

Moreover, core concepts include *medications* and *substances* in the sense that their mention denotes the actual administration of a pharmacological product or a substance. For example, an isolated annotation with *'387458008 |Aspirin (substance)|'* is understood such that the patient has actively taken aspirin.

(2) Symptomatik: **Anarthrie (a.)**.
Symptoms: ***Anarthria (a.)***
a. 48257004 |Anarthria (finding)|

4.2.2. Modifier Concepts

The role of modifier concepts does not typically express the central clinical information in a statement. Instead, they provide essential structural, physiological, or biological context found frequently in clinical narratives. This is why modifier concepts should always accompany a core concept in clinical statements.

For example, *body structures* describe normal and abnormal anatomical structures, including pathological specimens. *Physical objects* include human-produced entities such as medical devices or implants, and the *organisms* hierarchy covers all living structures, from bacteria to mammals. Example (3) illustrates that a core concept (3-a) is required to interpret the clinical meaning of the whole sentence, i.e., that the patient suffers from a left-hand fracture. The modifier concept (3-b) alone would not yield a clinically complete interpretation.

(3) Patient hat **Fraktur (a.)** an **linker Hand (b.)**.
Patient has ***fracture (a.)*** *of* ***left hand (b.)***.
a. 125605004 |Fracture of bone (disorder)|
b. 85151006 |Structure of left hand (body structure)|

4.2.3. Qualifier Concepts

Qualifier concepts form a broad concept role in our annotation scheme, and allow the representation of information that contextualizes clinical statements, e.g., by adding factuality or temporality properties to core concept annotations.

General qualifiers include units, like *cm*, or adjectival modifiers, like *normal*, (cf. example (6-d)). To capture temporal information in GEMTEX, all absolute dates receive a specific concept annotation, as exemplified in (4), while relative dates are annotated compositionally: Since numerals are not represented as concepts in SNOMED CT, they are only normalized as decimal numbers as shown in (5-c). The respective time unit receives its corresponding concept, as displayed in (5-d). Expressions that signal event order are annotated using specific concepts for *before* (*'272113006 |Before values (qualifier value)'*), *during* (*'272114000 |During values (qualifier value)'*), and *after* (*'288563008 | After values (qualifier value)'*). Those concepts are, however, only used if two events stand in explicit temporal relation to each other. Furthermore, planned events are marked with the concept *'97943006 | Planned (qualifier value)|'*. At the same time, trigger words expressing that an event is over are annotated with *'410513005 |In the past (qualifier value)|'* as shown in (5-b) and (6-a).

Factualities, such as *negation*, *presence*, and *uncertainty*, are important features of clinical narratives, e.g., it usually makes a difference whether a patient *has* or *does not have* a specific diagnosis. To account for this special role of factualities, our approach pays special attention to this decision problem. By default, all statements are interpreted as *true* and *certain* unless they contain a specific factuality trigger. For example, the expression "history of" is considered a trigger of posteriority, which makes it receive the concept annotation shown in (6-a). At the same time, "history of" triggers high certainty and therefore receives the SNOMED CT code depicted in (6-b) when expressed with affirmative polarity, as so-called stacked annotation. On the other hand, "unlikely" is an example of a negative-polarity trigger expressing lower certainty, which receives the concept annotation *'410593006 |Probably not present (qualifier value)|'*.

Table 3 provides a non-exhaustive overview of the factuality types, their polarity, and representative trigger words, together with the SNOMED CT concepts that annotators are instructed to use.

(4) ED: **17.10.2025 (a.)**.
Initial Diagnosis: ***17.10.2025 (a.)***.
a. 410672004 |Date property (qualifier value)|

(5) **OP (a.)** war **vor (b.) zwei (c.) Wochen (d.)**.
Surgery (a.) *was* ***two (c.) weeks (d.) ago (b.)***.
a. 387713003 |Surgical procedure (procedure)|
b. 410513005 |In the past (qualifier value)|
c. 2.0
d. 258705008 |week (qualifier value)|

(6) **Z.n. (a., b.) COVID-19 (c.) normal(d.)**.
History of (a., b.) COVID-19 (c.) normal (d.).
a. 410513005 |In the past (qualifier value)|
b. 410605003 |Confirmed present (qualifier value)|
c. 840539006 |Disease caused by severe acute respiratory syndrome coronavirus 2 (disorder)|
d. 17621005 |Normal (qualifier value)|

Factuality	Polarity	SNOMED CT	Triggers
Higher Certainty	+	410605003 |Confirmed present (qualifier value)|	*history of, confirmed, ...*
	-	410516002 |Known absent (qualifier value)|	*exclusion of, refuted, ...*
Certainty	+	∅	-
	-	410516002 |Known absent (qualifier value)|	*not, never, ...*
Lower Certainty	+	410592001 |Probably present (qualifier value)|	*probably, suspected, ...*
	-	410593006 |Probably not present (qualifier value)|	*unlikely, unexpected, ...*

Table 3: Overview about certainty factuality triggers coupled with the polarity of statements, their corresponding SNOMED CT concepts, and example trigger words from the GEMTEX guidelines.

4.3. Annotation Maxims

In addition to the three high-level concept roles previously defined, we developed a set of annotation maxims, which guide the annotators through the annotation process. The first two maxims, *(I) Concept before Structure* and *(II) Better Together*, remind the annotators that, as a consequence of the poly-hierarchical structure and the compositionality of SNOMED CT, the concepts contained in SNOMED CT have a differing level of detail. This directly impacts our annotation scheme, as the delineation of the annotations is not always clear. For example, one could add the two differing annotations depicted in (7) and (8) to the same sentence. In the first example, a longer-span annotation was chosen to represent the span *"linker Hand"* (*left hand*) with the appropriate concept (7-b), precisely expressing the meaning of the span. On the other hand, the second example shows that annotators may also choose to split that span into (8-b) and (8-c), corresponding to *"linker"* (*left*) and *"Hand"* (*hand*), respectively.

At the same time, to prevent annotators from overthinking such annotations, meaningful shorter-span annotations are considered equally valid compared to more complete longer-span annotations. Thus, the longer-span annotation and the shorter-span annotation of the same utterance displayed in (7) and (8) would be considered equally correct. This is a core difference to comparable SNOMED CT-based annotation campaigns (e.g. Schulz et al., 2023; Davidson et al., 2025), but due to SNOMED CT's key characteristic, the compositionality, we do not see any problems related to loss of information in this maxim.

Moreover, the maxim *(III) Identify the Meaning* instructs annotators to annotate the intended meaning of a span given its context. For example, anaphora should always receive the same SNOMED CT concept annotation as their antecedent. Ambiguous text spans should be annotated with all meaningful SNOMED CT concept variants, leveraging stacked annotations. At the same time, *(IV) Stack with Care* restricts stacked annotations to only a few situations. For instance, we allow stacked or nested annotations whenever German phrasal verbs are split from their particle, as illustrated in Figure 2.

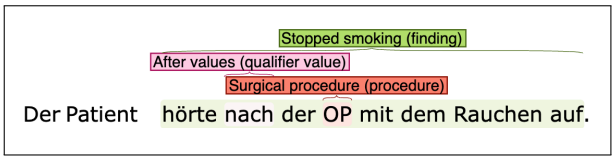

Figure 2: The German phrasal verb "auf$_{part}$hören$_{verb}$" (*stop*) allows the separation of particle and main verb. Such a grammatical construction triggers stacked annotations in the GEMTEX annotation scheme. English: *The patient stopped smoking after the surgery.*

The maxim *(V) In Doubt? Be Broad!* allows annotators to fall back to higher-level concepts if they are unable to identify the most precise SNOMED CT concept for a given text span. Such a fallback option has two advantages: the semantic tag of the annotation is preserved, while the space of options, as well as the time resources of the annotators, may be reduced. This is where the last maxim *(VI) Three Minutes Maximum* plugs in: Annotators must not search more than three minutes for a single concept. They should use the fallback strategy described previously and move on to the next relevant text span whenever this time period is exceeded.

(7) Patientin kommt mit **verletzter (a.) linker Hand (b.)**.
*The patient arrives with **injured (a.) left hand (b.)**.*
a. 417163006 |Traumatic or non-traumatic injury (disorder)|
b. 85151006 |Structure of left hand (body structure)|

(8) Patientin kommt mit **verletzter (a.) linker (b.) Hand (c.)**.
*The patient arrives with **injured (a.) left (b.) hand (c.)**.*
a. 417163006 |Traumatic or non-traumatic injury (disorder)|
b. 7771000 |Left (qualifier value)|
c. 85562004 |Hand structure (body structure)|

4.4. The Annotation Process and Preliminary Evaluation

Given the complexity and scale of the semantic annotation task, we adopted a canonical annotation workflow: The semantic annotation is performed by a single, trained annotator for each document using the INCEPTION platform. All annotators were trained on synthetic documents from the GRASCCO corpus before the productive annotation.

Annotators navigate SNOMED CT via the dedicated SNOMED CT BROWSER[19] to identify suitable concepts for each span. To increase efficiency and consistency, the process is supported by a pre-annotation step and a recommendation system. For pre-annotation, AVERBIS HEALTH DISCOVERY is used to detect candidate spans, followed by a dictionary-based lookup with the SNOMED CT GRAZ INTERFACE TERMINOLOGY (Hashemian Nik et al., 2019). The recommendation system, built on the ID LOGIK[20] terminology server by ID BERLIN, suggests alternative SNOMED CT concepts for each span, providing annotators with multiple options beyond the initial pre-annotation.

To assess the reliability of our annotation scheme, we conducted a pilot IAA study on a single document from the GRASCCO corpus.[21] The study involved eight annotators from *Technical University of Munich*, who collectively produced 3.7k annotations. Agreement was computed on a span-based *candidate universe* defined as the union of all spans annotated by any annotator; each unique span in this union constituted one agreement unit. For each unit and annotator, we encoded the assigned labels as an unordered, stack-aware *bag* to preserve stacked annotations. Units not annotated by a given annotator were explicitly marked, yielding a dense unit-by-annotator matrix. We quantified reliability using Krippendorff's α on these categorical bag values, and estimated 95 % confidence intervals by bootstrapping agreement units with replacement (1k resamples). We report agreement at two granularities: (i) the SNOMED CT concept and (ii) the semantic tag. Using conservative exact-span unitization, we obtained α = 0.65 (95% CI: [0.63, 0.68]) at the concept level and α = 0.71 (95% CI: [0.69, 0.74]) at the semantic-tag level.[22] These numbers suggest solid reliability of our semantic annotation scheme, given the relatively high number of independent annotators and the complexity of the concept-oriented annotation.

[19] https://browser.ihtsdotools.org/?

[20] https://www.id-berlin.de/produkte/nlp-forschung/id-logik/

[21] Access the document with gold standard annotations: https://doi.org/10.5281/zenodo.18861607.

[22] Access our agreement computation scripts: https://github.com/Jhofenbitzer/gemtex-semantic-agreement.

Overall, the GEMTEX project developed in phase III a full-stack, medically relevant semantic annotation scheme and workflow based on the direct grounding of clinical narratives with SNOMED CT. The process involves a general semantic annotation, and we introduced thorough guidelines to navigate annotators through this complex process. Moreover, a pilot inter-annotator agreement study underlined their viability.

5. Current Status of GEMTEX

Currently, the GEMTEX project members are working in parallel in phases II and III, i.e., de-identification and semantic annotation. Thus, we report an interim status regarding the size and composition of the text corpus in Table 4: As of February 2026, the resource contains over 15k distinct, de-identified documents, corresponding to more than 20M tokens. The semantic annotation campaign started only recently, and 382 documents ($\approx$ 791k tokens) have already been annotated. The annotators across the six partner hospitals have created more than 189k annotations. This corresponds to about 239 annotations per 1k tokens and shows how densely information is distributed in clinical notes. Notably, this value corresponds to raw annotation counts, including diverging annotations from multiple annotators.

	De-ID	Semantic Annotation
# Documents	15,463	382
# Tokens	20,346,527	791,464
# Annotations	682,047	189,362

Table 4: GEMTEX corpus statistics for de-identification and semantic annotation subsets (February 2026).

Figure 3 shows the six SNOMED CT concepts, which were annotated most often in the corpus, grouped by semantic tag and the concept role they belong to. Almost half of the annotations (43.8%), i.e., more than 83k, carry the semantic tag *qualifier value*. The subsequent five semantic tags belong to the core concepts, i.e., *procedures* with 22k (12.0%), *findings* with 17k (9.3%), *disorders* with 14k (7.4%), and *observable entities* with 12k annotations (6.3%). The only semantic tag among the top six belonging to the modifier concept group is *body structure* with 11k annotations (5.9%). Moreover, Figure 3 illustrates that the six semantic tags with the highest distribution make almost 85% of all annotations, with the other groups contributing only 15.2% to the overall count.

Especially the latter finding suggests a long-tail distribution of annotated SNOMED CT concepts. It highlights the role of the core concepts in our an-

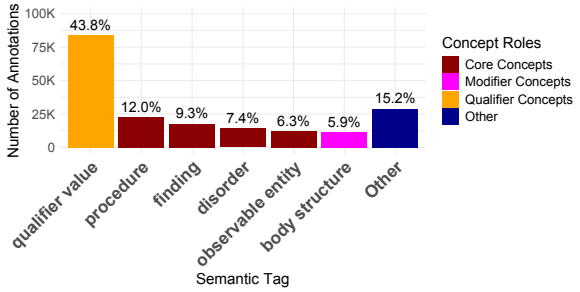

Figure 3: Raw count and percentage distribution of annotated concepts, grouped by their semantic tag.

notation scheme: Four out of six semantic tags from the top-ranked annotations belong to this role of concepts, which underlines their importance in clinical documents. Similarly, it is unsurprising to see a high proportion of qualifier values in this distribution, as qualifier values are not only an extensive group of SNOMED CT concepts, but also a very important one. Whenever a clinical statement needs a contextualized representation, annotators will likely choose a concept belonging to the qualifier values. Additionally, all dates are annotated using a qualifier concept.

6. Access to GEMTEX

The GEMTEX resource is under active development at least until the end of 2026. As outlined in Section 3, GEMTEX is a distributed corpus. The clinical documents remain at the site of their origin and are only compiled into a unified resource for usage after successful application.[23] Therefore, the requirement to use GEMTEX for research is a formal DUA process, operationalized under the established regulations and governance of the MII[24]. This allows also for the retrieval of structured data, e.g., codes of diagnoses and procedures, or laboratory values of the patients which provided text to the corpus, enabling a joint analysis of textual and structured data. The standardized procedure for a data use application consists of three steps: (1) Provision of a study protocol outlining the intended use of the corpus, (2) IRB vote for the intended study, and (3) application for data access by submitting the study protocol and the IRB vote to the *Forschungsdatenportal Gesundheit* [*German Portal for Medical Research Data*] (FDPG),[25], a nationally established platform to access clinical data.

We refer researchers interested in using GEMTEX to visit https://kiinformatik.mri.tum.de/de/mi/research: We are committed to keeping that page always up-to-date regarding data access options. If questions remain, we encourage to reach out via gemtex.mi@mh.tum.de.

7. Conclusion and Future Work

We have presented three key steps in creating the novel clinical language resource GEMTEX: (1) ensuring legal compliance with European privacy laws, (2) de-identifying real clinical documents to make them usable for research, and (3) semantically enriching medically relevant spans within the narratives using the international standard terminology SNOMED CT. Our process may serve as a model project for future clinical corpus creation processes. As shown in Section 2, such initiatives are necessary, as only four German real-world clinical corpora are currently available, each stemming from at most two sites and reaching a maximum of 1M tokens. GEMTEX surpasses these in all respects: it already contains more than 15k documents, about 20M tokens, and 189k annotations from multiple clinical domains in a multi-centric setting. These numbers are expected to increase significantly, since the resource is still under active development. In addition, we plan to add a second domain-specific semantic annotation layer for the clinical disciplines cardiology, neurology, oncology, and adverse drug events.

Future work will focus on a thorough evaluation of GEMTEX, combining corpus characterization, annotation quality assessment, and downstream benchmarking. We will analyze structural and semantic document composition across document types, clinical disciplines, and participating institutions. On the semantic layer, we will extend inter-annotator agreement and error analyses by explicitly disentangling disagreement sources such as span boundary variation, concept selection, and stacked annotations. We further aim to establish standardized task suites for German clinical NLP, including span- and concept-level named entity recognition, SNOMED CT–based normalization and coding with hierarchy-aware scoring. Owing to its controlled accessibility, GEMTEX is expected to remain resistant to large-scale scraping and thus largely absent from future LLM training corpora, positioning it as a contamination-robust benchmark for German clinical domain adaptation and model comparison.

[23] The GEMTEX project developed a core dataset module as part of a broad MII-based standard to store and deliver data for research, particular for texts from the corpus: https://www.medizininformatik-initiative.de/Kerndatensatz/KDS_Dokument/MIIIGModulDokument.html.

[24] https://www.medizininformatik-initiative.de/en/standardised-use-and-access-rules

[25] https://forschen-fuer-gesundheit.de/

Limitations

As this work describes a resource currently under development, the primary limitation is that GEMTEX is not yet complete. The processing of the sensitive clinical documents is ongoing. Consequently, while we provided a quantitative evaluation of the annotation quality, including measures of inter-annotator agreement, only within a pilot experiment using a synthetic GRASCCO document, a comprehensive, site-level evaluation has not yet been conducted. We plan to perform these evaluations as the annotation process progresses.

Ethics Statement

All annotators were employed as student assistants and received fair compensation for their efforts. The GEMTEX corpus compilation received ethics approval from the IRB of TUM Klinikum under the identifier 2024-180-S-SB, which was acknowledged by the partnering sites. Leipzig, Essen and Dresden assigned this acknowledged vote another identifier: 196/24-lk (Leipzig), 24-12044-BO (Essen), and BO-EK-201052024 (Dresden). Berlin adopted the ethics vote of TU Munich based on the regulations in the professional code of the Berlin Chamber of Physicians. We did not consider any ethnic or gender criteria by selecting the documents.

Use of AI Assistants

The authors acknowledge the use of AI agents for grammatical and stylistic improvements of the submitted manuscript.

Acknowledgments

We would like to thank Serwar Basch, Raffael Bild, Claudio Benzoni, Felicitas De La Cruz Rothenfußer, Richard Eckart De Castilho, Johanna Eicher, Carsten Eickhoff, Anna Fackler, Steffen Franke, Diego Frassinelli, Oksana Galusch, Thomas Ganslandt, Matthias Gietzelt, Viktoria Hartmann, Emma Hess, Karolin Hofmann, Leen Houri, Andrei-Albert Kiss, Adrian Kehl, Lisa Kellner, Janina Kind, Markus Kreuzthaler, Juliane Krieger, Nektarios Ladas, Jacqueline Lammert, Franz Matthies, Hung Manh Nguyen, Matthias Nüchter, Jazia Omeirat, Peter Pallaoro, Fabian Prasser, Frank Richter, Phillip Richter-Pechanski, Andreas Rohrbach, Lena Rollinger, André Sander, Eyal Schejter, Suteera Seeha, Helmut Spengler, Cosima Strantz, Elena Thias, Yutong Wen, Johannes Wendl, Thomas Wendt and Markus Wolfien such as all annotation students for their continuous support and vivid discussions. This work was funded by the German Federal Ministry of Research, Technology, and Space under grants 01ZZ2314A, 01ZZ2314B, 01ZZ2314E, 01ZZ2314G, 01ZZ2314K, 01ZZ2314L, 01ZZ2314O, 01ZZ2314P, and 01KX2121.

References

Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne R. Tamang, and Robert Rallo. 2019. SNOMED2VEC : random walk and Poincaré embeddings of a clinical knowledge base for healthcare analytics. In *DSHealth 2019 — Proceedings of the 2019 Workshop on Applied Data Science for Healthcare: Bridging the Gap between Data and Knowledge @ KDD 2019. Anchorage, Alaska, USA, August 5, 2019*, page #43, New York/NY. Association for Computing Machinery (ACM).

Franz Baader, Sebastian Brandt, and Carsten Lutz. 2005. Pushing the EL envelope. In *Proceedings of the 19th international joint conference on Artificial intelligence*, IJCAI'05, pages 364–369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tim Benson and Grahame Grieve. 2016. *Principles of Health Interoperability: SNOMED CT, HL7 and FHIR*. Springer International Publishing.

Steven J. Bethard, Leon R. A. Derczynski, Guergana K. Savova, James D. Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 Task 6: Clinical TempEval. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015. Denver, Colorado, USA, June 4-5, 2015*, pages 806–814, Red Hook/NY. Association for Computational Linguistics (ACL), Curran Associates.

Steven J. Bethard, Guergana K. Savova, Wei-Te Chen, Leon R. A. Derczynski, James D. Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 Task 12: Clinical TempEval. In *SemEval 2016 — Proceedings of the 10th International Workshop on Semantic Evaluation @ NAACL-HLT 2016. San Diego, California, USA, June 16-17, 2016*, pages 1052–1062, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Steven J. Bethard, Guergana K. Savova, Martha Palmer, and James D. Pustejovsky. 2017. SemEval-2017 Task 12: Clinical TempEval. In *SemEval 2017 — Proceedings of the 11th International Workshop on Semantic Evaluation @ ACL 2017. Vancouver, British Columbia, Canada,*

August 3-4, 2017, pages 565–572, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Raffael Bild, Martin Bialke, Karoline Buckow, Thomas Ganslandt, Kristina Ihrig, Roland Jahns, Angela Merzweiler, Sybille Roschka, Björn Schreiweis, Sebastian Stäubert, Sven Zenker, and Fabian Prasser. 2020. Towards a comprehensive and interoperable representation of consent-based data usage permissions in the German medical informatics initiative. *BMC Medical Informatics and Decision Making*, 20(1):103.

Linda Bird. 2025. *The Essential Guide to SNOMED CT®*. Springer Nature Switzerland.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P. Schapranow. 2022. GGPONC 2.0 — The *German Clinical Guideline Corpus for Oncology*: curation workflow, annotation policy, baseline NER taggers. In *LREC 2022 — Proceedings of the 13th International Conference on Language Resources and Evaluation. Marseille, France, June 20-25, 2022*, pages 3650-–3660, Paris. European Language Resources Association (ELRA).

David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia A. Brandt, and Richard Andrew Taylor. 2020. Benchmark and best practices for biomedical knowledge graph embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Natural Language Processing (BioNLP 2020) at ACL 2020, July 9, 2020, Virtual Event*, pages 167–176.

Eunsuk Chang and Sumi Sung. 2024. Use of SNOMED CT in large language models: scoping review. *JMIR Medical Informatics*, 12:1–20, article no. e62924.

Rory Davidson, Will Hardman, Guy Amit, Yonatan Bilu, Vincenzo Della Mea, Aleksandr Galaida, Irena Girshovitz, Mikhail Kulyabin, Mihai Horia Popescu, Kevin Roitero, Gleb Sokolov, and Chen Yanover. 2025. SNOMED CT entity linking challenge. *Journal of the American Medical Informatics Association*, 32(9):1397–1406.

Cornelia L. A. Dewald, Alina Balandis, Lena S. Becker, Jan B. Hinrichs, Christian von Falck, Frank K. Wacker, Hans Laser, Svetlana Gerbel, Hinrich B. Winther, and Johanna Apfel-Starke. 2023. Automated classification of free-text radiology reports: using different feature extraction methods to identify fractures of the *distal fibula*. *RöFo –Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 195(8):713–719.

Richard Eckart De Castilho, Jan-Christoph Klie, and Iryna Gurevych. 2024. Integrating INCEpTION into larger annotation processes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 110–121, Miami, Florida, USA. Association for Computational Linguistics.

Noémie Elhadad, Sameer S. Pradhan, Sharon Lipsky Gorman, Suresh Manandhar, Wendy W. Chapman, and Guergana K. Savova. 2015. SemEval-2015 Task 14: Analysis of Clinical Text. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015. Denver, Colorado, USA, June 4-5, 2015*, pages 303–310, Red Hook/NY. Association for Computational Linguistics (ACL), Curran Associates.

Jakob Faller, Christina Lohr, Martin Boeker, and Frank Meineke. 2025. Building the Infrastructure for the German Medical Text Corpus Project (GeMTeX). *Studies in Health Technology and Informatics*, 327:894–895.

Aitor González-Agirre, Montserrat Marimón, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PHARMACONER : Pharmacological substances, Compounds and proteins Named Entity Recognition Track. In *BioNLP-OST 2019 — Proceedings of the 5th Workshop on BioNLP Open Shared Tasks @ EMNLP-IJCNLP 2019. Hong Kong, China, November 4, 2019*, pages 1–10, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Udo Hahn. 2025. Clinical document corpora—real ones, translated and synthetic substitutes, and assorted domain proxies: a survey of diversity in corpus design, with focus on German text data. *JAMIA Open*, 8(3).

David Hashemian Nik, Zdenko Kasáč, Zsófia Goda, Anita Semlitsch, and Stefan Schulz. 2019. Building an Experimental German User Interface Terminology Linked to SNOMED CT. *Studies in Health Technology and Informatics*, 264:153–157.

Sam Henry, Yanshan Wang, Feichen Shen, and Özlem Uzuner. 2020a. The 2019 National Natural language processing (NLP) Clinical Challenges (n2c2)/Open Health NLP (OHNLP) Shared Task on Clinical Concept Normalization for Clinical Records. *Journal of the American Medical Informatics Association*, 27(10):1529-1537.

Samuel Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Özlem Uzuner. 2020b. 2018

n2c2 Shared Task on Adverse Drug Events and Medication Extraction in Electronic Health Records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Justin Hofenbitzer, Stefan Schulz, Martin Boeker, Peter Klügl, Sarah Riepenhausen, Christina Lohr, Jacqueline Lammert, Andrea Riedel, and Luise Modersohn. 2025. Introducing Medical Semantic Annotation Guidelines for German Clinical Documentation with SNOMED CT. *70. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS)*.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. MIMIC-IV-Note: Deidentified free-text clinical notes. Type: Dataset.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023b. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, Marit Zettwitz, Till de Bortoli, Leonie Ostermann, Jurica Ševa, Johannes Starlinger, Oliver Kohlbacher, Nisar P Malek, Ulrich Keilholz, and Ulf Leser. 2021. Annotation and initial evaluation of a large annotated German oncological corpus. *JAMIA Open*, 4(2).

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

Klaus Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(1):61–70.

Vishesh Kumar, Amber Stubbs, Stanley Shaw, and Özlem Uzuner. 2015. Creation of a new longitudinal corpus of clinical narratives. *Journal of Biomedical Informatics*, 58:6–10. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

Stefan Lenz, Arsenij Ustjanzew, Marco Jeray, Meike Ressing, and Torsten Panholzer. 2025. Can open source large language models be used for tumor documentation in Germany? An evaluation on urological doctors' notes. *BioData Mining*, 18:1–25, article no. 48.

Christina Lohr, Jakob Faller, Andrea Riedel, Hung Manh Nguyen, Markus Wolfien, Justin Hofenbitzer, Luise Modersohn, Jutta Romberg, Fabian Prasser, Jazia Omeirat, Yutong Wen, Oksana Galusch, Udo Hahn, Marvin Seiferling, Christoph Dieterich, Peter Klügl, Franz Matthies, Janina Kind, Martin Boeker, Markus Löffler, and Frank Meineke. 2025. GeMTeX's De-Identification in Action: Lessons Learned & Devil's Details. In *German Medical Data Sciences 2025: GMDS Illuminates Health*, pages 274–282. IOS Press.

Christina Lohr, Franz Matthies, Jakob Faller, Luise Modersohn, Andrea Riedel, Udo Hahn, Rebekka Kiser, Martin Boeker, and Frank Meineke. 2024. *De-Identifying GRASCCO - A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus*, volume 317, pages 171–179. IOS Press.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. MCN : a comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, 92:103–132.

Diwakar Mahajan, Jennifer J. Liang, Ching-Huei Tsou, and Özlem Uzuner. 2023. Overview of the 2022 n2c2 Shared Task on Contextualized Medication Event Extraction in Clinical Notes. *Journal of Biomedical Informatics*, 144(Special Issue on Clinical Natural Language Processing for Secondary Use Applications; ed. by Meliha Yetisgen, Özlem Uzuner, Yanjun Gao, Diwakar Mahajan):1–10, article no. 104432.

Frank Meineke, Luise Modersohn, Markus Loeffler, and Martin Boeker. 2023. *Announcement of the German Medical Text Corpus Project (GeMTeX)*. IOS Press.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources.

Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. GRASCCO - The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. *Studies in Health Technology and Informatics*, 296:66–72.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana K. Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9:1–13, article no. 12.

Namu Park, Kevin Lybarger, Giridhar Kaushik Ramachandran, Spencer Lewis, Aashka Damani, Özlem Uzuner, Martin L. Gunn, and Meliha Yetisgen. 2024. A novel corpus of annotated medical imaging reports and information extraction results using BERT-based language models. In *LREC-COLING 2024 — Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. Torino, Italia, 20-25 May, 2024 (Hybrid Event)*, pages 1280–1292, Paris. European Language Resources Association (ELRA) & International Committee on Computational Linguistics (ICCL), European Language Resources Association (ELRA).

Sameer S. Pradhan, Noémie Elhadad, Wendy W. Chapman, Suresh Manandhar, and Guergana K. Savova. 2014. SemEval-2014 Task 7: Analysis of Clinical Text. In *SemEval 2014 — Proceedings of the 8th International Workshop on Semantic Evaluation @ COLING 2014. Dublin, Ireland, August 23-24, 2014*, pages 54–62. Association for Computational Linguistics (ACL) Special Interest Group on the Lexicon (SIGLEX).

Sameer S. Pradhan, Noémie Elhadad, Brett R. South, David Martínez, Lee M. Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana K. Savova. 2015. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154.

Phillip Richter-Pechanski, Philipp Wiesenbach, Dominic M. Schwab, Christina Kiriakou, Mingyang He, Michael M. Allers, Anna S. Tiefenbacher, Nicola Kunz, Anna Martynova, Noemie Spiller, Julian Mierisch, Florian Borchert, Charlotte Schwind, Norbert Frey, Christoph Dieterich, and Nicolas A. Geis. 2023. A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters. *Scientific Data*, 10(1):207.

Raul Rodriguez-Esteban. 2009. Biomedical text mining and its applications. *PLoS Computational Biology*, 5(12):1–5, article no. e1000597.

Stefan Schulz, Warren Del-Pinto, Lifeng Han, Markus Kreuzthaler, Sareh Aghaei Dinani, and Goran Nenadic. 2023. Towards Principles of Ontology-Based Annotation of Clinical Narratives. In *Proceedings of the International Conference on Biomedical Ontologies 2023 together with the Workshop on Ontologies for Infectious and Immune-Mediated Disease Data Science (OIIDDS 2023) and the FAIR Ontology Harmonization and TRUST Data Interoperability Workshop (FOHTI 2023)*.

Sebastian C Semler, Frank Wissing, and Ralf Heyder. 2018. German medical informatics initiative. *Methods of Information in Medicine*, 57(S 01):e50–e56.

Maria Skeppstedt, Maria Kvist, and Hercules Dalianis. 2012. Rule-based entity recognition and coverage of SNOMED CT in Swedish clinical text. In *LREC 2012 — Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, May 21-27, 2012*, pages 1250–1257. European Language Resources Association (ELRA).

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. 2019. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: overview of 2016 CEGS NGRID Shared Tasks Track 1. *Journal of Biomedical Informatics*, 75(Supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry):S4–S18.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58(Supplement):S20–S29.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana K. Savova, Noémie Elhadad, Sameer S. Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martínez, and Guido Zuccon. 2013. Overview of the SHARE/CLEF EHEALTH EVALUATION LAB 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization. CLEF 2013 — Proceedings of the 4th International Conference of the CLEF Initiative. Valencia, Spain, September 23-26, 2013*, number 8138 in Lecture Notes in Computer Science (LNCS), pages 212–231, Berlin, Heidelberg. Springer.

Kanimozhi Uma and Marie-Francine Moens. 2024. Unraveling clinical insights: a lightweight and interpretable approach for multimodal and multilingual knowledge integration. In *CL4Health 2024 — Proceedings of the 1st Workshop on Patient-Oriented Language Processing @ LREC-COLING 2024. Torino, Italia, 20 May, 2024*, pages 197–203, Paris. European Language Resources Association (ELRA) & International Committee on Computational Linguistics (ICCL), European Language Resources Association (ELRA).

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Sven Zenker, Daniel Strech, Kristina Ihrig, Roland Jahns, Gabriele Müller, Christoph Schickhardt, Georg Schmidt, Roland Speer, Eva Winkler, Sebastian Graf von Kielmansegg, and Johannes Drepper. 2024a. Der Broad Consent der Medizininformatik-Initiative (MII). Rationale, Entwicklung und Erläuterungen.

Sven Zenker, Daniel Strech, Roland Jahns, Gabriele Müller, Fabian Prasser, Christoph Schickhardt, Georg Schmidt, Sebastian C. Semler, Eva Winkler, and Johannes Drepper. 2024b. National standardisierter Broad Consent in der Praxis: erste Erfahrungen, aktuelle Entwicklungen und kritische Betrachtungen. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 67(6):637–647.

Appendix

A. Excluded SNOMED CT Concepts

To account for the cognitively demanding SNOMED CT-based semantic annotation of clinical documents, we decided to exclude a group of SNOMED CT concepts, as depicted in Table 5.

Taxonomic Predecessor	**Semantic Tags**
Environment or geographical location	*environment, geographic location*
Qualifier value	*action, overlapping sites*
Record artifact	*record artifact*
Situation with explicit context	*situation*
SNOMED CT Model Component	*metadata, attribute, core metadata concept, foundation metadata concept, link assertion, linkage concept, namespace concept, OWL metadata concept*
Social context	*social concept, ethnic group, life style, occupation, racial group, religion/philosophy*
Special concept	*special concept, inactive concept, navigational concept*
Staging and scales	*staging scale, assessment scale, tumor staging*

Table 5: List of excluded SNOMED CT concepts from the GEMTEX semantic annotation, grouped by a taxonomic predecessor and semantic tags.