German Medical Data Sciences 2024
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI240853

De-Identifying GRASCCO – A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus

Christina LOHR[ad1], Franz MATTHIES[ad], Jakob FALLER[bd], Luise MODERSOHN[cd], Andrea RIEDEL[bd], Udo HAHN[ad], Rebekka KISER[c], Martin BOEKER[cd], and Frank MEINEKE[ad]
[a] Institute for Medical Informatics, Statistics, and Epidemiology, Leipzig University, Germany, [b] Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, [c]Institute of Artificial Intelligence and Informatics in Medicine, Medical Center rechts der Isar, Technical University Munich, Germany,
[d] GeMTeX Consortium of the German Medical Informatics Initiative
ORCiD-ID: Christina Lohr https://orgid.org/0000-0001-9889-162X

Abstract. Introduction: The German Medical Text Project (GeMTeX) is one of the largest infrastructure efforts targeting German-language clinical documents. We here introduce the architecture of the de-identification pipeline of GeMTeX. **Methods:** This pipeline comprises the export of raw clinical documents from the local hospital information system, the import into the annotation platform INCEpTION, fully automatic pre-tagging with protected health information (PHI) items by the Averbis Health Discovery pipeline, a manual curation step of these pre-annotated data, and, finally, the automatic replacement of PHI items with type-conformant substitutes. This design was implemented in a pilot study involving six annotators and two curators each at the Data Integration Centers of the University Hospitals Leipzig and Erlangen. **Results:** As a proof of concept, the publicly available Graz Synthetic Text Clinical Corpus (GRASSCO) was enhanced with PHI annotations in an annotation campaign for which reasonable inter-annotator agreement values of Krippendorff's α ≈ 0.97 can be reported. **Conclusion:** These curated 1.4 K PHI annotations are released as open-source data constituting the first publicly available German clinical language text corpus with PHI metadata.

Keywords. Natural Language Processing, De-Identification, Patient Data Privacy

1. Introduction

The *Medical Informatics Initiative* (MII) [1] is the largest research effort up until now to process clinical patient data in Germany. Whereas structured data (e.g., diagnostic codes, laboratory data, administered medications) have long been the predominant focus of data integration efforts, the MII project *GeMTeX (German Medical Text Project)* is concerned with unstructured clinical free text and solutions for the public accessibility of German clinical documents [2]. More concrete, the aim of GeMTeX is to set up a collection of

[1] Corresponding author: Christina Lohr (christina.lohr@imise.uni-leipzig.de), Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Härtelstraße 16-18, D-04107 Leipzig

clinical patient histories and to assemble a composite text corpus from these histories at multiple clinical sites in Munich (TU), Leipzig, Dresden, Berlin (Charité), Essen and Erlangen in order to train (large) language models [3,4]. To grant accessibility of these data beyond local clinical walls (based on contractual Data Use Agreements (DUA)), a confident level of data protection must be guaranteed.

GeMTeX's approach to safe data accessibility is based on hiding patients' identity signals dispersed over clinical documents by de-identification, i.e., neutralizing privacy-sensitive Protected Health Information (PHI) text elements such as name, patient ID or address. Prior to running a large-scale de-identification campaign on the whole data set of GeMTeX, we conducted a pilot study on GRASSCO, a synthetic German-language clinical data set that is already publicly available [5] to test the design of GeMTeX's de-identification pipeline under prototype conditions. These data (1.4k PHI annotations published on ZENODO)[2] provide added value on their own, since they constitute the first German-language clinical text corpus containing publicly available PHI metadata and, therefore, are easily accessible for clinic-external NLP researchers. Moreover, these GRASSCO metadata will be used in the future as a common ground for comparison to investigate the validity of the local annotations provided by all six GeMTeX sites.

2. Related Work

During the first funding phase of the MII, the "3000PA" text corpus was assembled, which is mainly composed of discharge letters of some three thousand patients. It was created exploiting the Electronic Health Records from deceased patients treated in three German university hospitals – Jena, Aachen, and Leipzig (the founding members of the MII SMITH consortium [6]) – comprising roughly 7 M tokens altogether (for a detailed description of 3000PA, see [7,8]). Selected parts of this corpus were the basis for manual annotation campaigns with focus on medications, de-identification, section heading segmentation, and crucial clinical named entity types, such as diagnoses, findings, and symptoms [7,9–11], respectively. 3000PA was assembled in 2016, a long time before the Broad Consent policy of the MII was established as a reaction to patient involvement based on individual consent for AI-related research purposes [12]. Due to the lack of patient consent, 3000PA could only be processed in the hospitals' local premises or by associated project partners. Public access (e.g., via DUA or even free download) was prohibited both for anonymized [9] and pseudonymized [13] versions of 3000PA.

In this brief discussion, we adopt a data release perspective on German clinical text corpora (for a more detailed overview of German corpora, see [14]).

The first release of a publicly accessible German language clinical text corpus was achieved in 2021 with BRONCO comprising 150 oncological discharge summaries (90 K tokens) [15]. The data set is available under a DUA regime. BRONCO comes with arbitrarily shuffled sentences (for increased data protection) which break the linear structure of the original discharge summary and anonymous placeholders for patient names and other privacy-sensitive text items. In 2023, CARDIO:DE was published containing 500 discharge summaries (993 K tokens) from a cardiology department [16] (also accessible under a DUA policy), yet preserving the coherent structure of a typical discharge summary. Based on an anonymization strategy described in [17] the released corpus contains placeholders to de-identify names. As an alternative to those real clinical

[2] https://doi.org/10.5281/zenodo.11502329

data sets, GRASSCO was manually created as a synthetic clinical corpus with fictitious, expert-style clinical reports [5]. It contains 63 discharge summaries (44 K tokens) with invented patient stories not referring to real individuals. Hence, de-identification is not an issue here and access is allowed for NLP researchers and developers without any restrictions. Alternatively, one might exploit non-clinical German language models for the automatic creation of synthetic clinical statements (isolated sentences only, yet not documents; see [18]). Automatic translations of non-German (typically English) clinical documents (see [19]) should be used with caution because (besides reliability and validity issues with automatic translations) some medical information is country-specific.

3. Methods

The lawful processing of health data requires the existence of a sufficient legal basis (Art. 6 and Art. 9 General Data Protection Regulation (GDPR)).[3] The processing of this so-called "special category of personal data" is usually based on the informed consent of the data subject (patient) or a specific legal basis. For clarification, it has to be pointed out that there must be a corresponding legal basis for each processing step. With regard to the creation of the GeMTeX corpus, these are *(1)* the local processing of the directly identifiable clinical text documents by the participating healthcare institutions such that the contents of the texts be de-identified in accordance with the *"GeMTeX-DeID-Guidelines,"* and *(2)* the processing of pseudonymized data within the framework of future research projects based on the GeMTeX corpus (as outlined, e.g., in [20]). Even if the contents of the medical texts undergo an anonymization/de-identification process, the local trustee offices connected to the participating health care providers, are in possession of the pseudonymization key. Therefore, they have the option of matching the medical document to the concrete individual the document is about. That is why, in the context of the GeMTeX project, a sufficient basis is needed for the potential transmission of pseudonymized data for research purposes to a third party.

While step *(1)*, the local creation of the GeMTeX corpus, is covered by various legal provisions, e.g., state hospital laws and since March 2024 also the nation-wide applicable *"Gesundheitsdatennutzungsgesetz"* (GDNG), there is no sufficient legal basis for step *(2)*, the use of pseudonymized personal health data in the context of a myriad of research projects. Therefore, only clinical documents are incorporated in the GeMTeX corpus from patients who have signed the MII Broad Consent [12] and thus agreed to further use of their health care data in various research projects. The GeMTeX project thereby reflects the current legal situation, which requires the consent of the data subject for the processing of pseudonymized health data in the context of scientific projects (see also § 6 (3) GDNG). Currently there is an appeal pending at the European Court of Justice (EuGH) regarding a judgment of the General Court of the European Union from April 26, 2023, in which the court commented on the personal reference of pseudonymized data. The court ruled that pseudonymized data transmitted to a third party is not to be considered personal data if the recipient does not have the means to re-identify the data subjects (Case T-557/20).[4] The EuGH's decision will have a significant influence on the legal basis for the use of pseudonymised health data by third parties and the question

[3] https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=celex%3A32016R0679 (accessed: June 25, 2024)
[4] https://curia.europa.eu/juris/liste.jsf?num=T-557/20 (accessed: June 25, 2024)

whether the GDPR is applicable for data processing carried out by this third party (see recital 26 GDPR).

Obtaining the MII Broad Consent has officially been integrated in the clinical admission workflow in all participating university hospitals.

3.1. Process

The goal of de-identification of clinical documents is to preclude the re-identification of individual persons involved in clinical procedures. Hence, the de-identification task can be phrased as the reliable recognition of text stretches that carry the potential to identify human individuals and mask these text pieces for subsequent processing. GeMTeX's de-identification process consists of the following steps:

- Automatic recognition of PHI (by a dedicated component of the commercial software *Averbis Health Discovery (AHD)*; Averbis is a GeMTeX partner.
- Independent checks of the PHI pre-tagging resulting from AHD are carried out by two trained human annotators using the INCEpTION annotation platform [21]. This step leads to the correction of unrecognized PHI (false negatives) and incorrectly tagged types of PHI (false positives), if necessary. The full plain text and all tagging decisions of AHD are provided at this stage. In addition, (obeying to data protection regulations in Germany) a sample of the de-identified texts has to be picked by an annotation curator for cross-checking to ensure the correctness of the de-identification process and to guarantee (almost) zero misses.
- Automatic replacement of PHI stretches with type-preserving entity names (e.g., person name, address, phone number) that change the original PHI items in such a way that it will no longer be possible to recover individual persons.

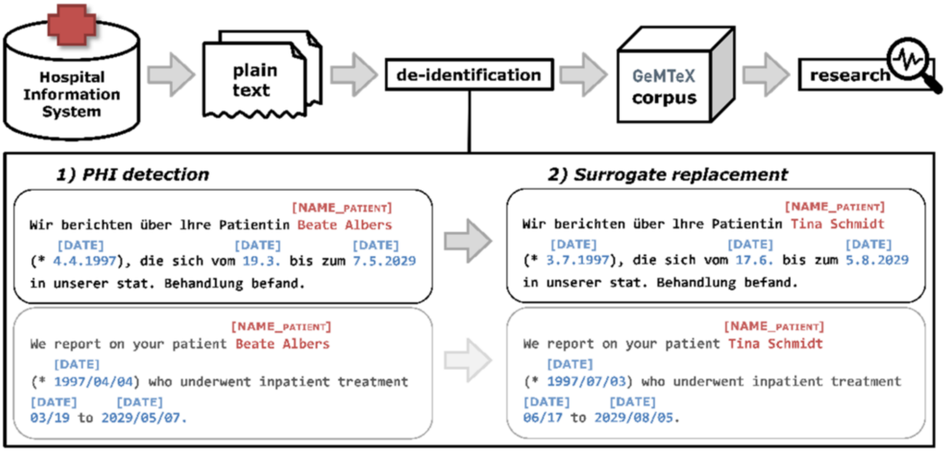
Hospital Information System
plain text
de-identification
GeMTeX corpus
research
1) PHI detection
2) Surrogate replacement
[NAME_PATIENT]
Wir berichten über Ihre Patientin Beate Albers
[DATE] [DATE] [DATE]
(* 4.4.1997), die sich vom 19.3. bis zum 7.5.2029
in unserer stat. Behandlung befand.
[NAME_PATIENT]
Wir berichten über Ihre Patientin Tina Schmidt
[DATE] [DATE] [DATE]
(* 3.7.1997), die sich vom 17.6. bis zum 5.8.2029
in unserer stat. Behandlung befand.
[NAME_PATIENT]
We report on your patient Beate Albers
[DATE]
(* 1997/04/04) who underwent inpatient treatment
[DATE] [DATE]
03/19 to 2029/05/07.
[NAME_PATIENT]
We report on your patient Tina Schmidt
[DATE]
(* 1997/07/03) who underwent inpatient treatment
[DATE] [DATE]
06/17 to 2029/08/05.

Figure 1: De-Identification as part of the Architecture in GeMTeX

3.2. Data Extraction

The raw texts are fetched from the clinical site in various technical formats from very different clinical information systems, yet textual content will be input to the pipeline as plain UTF-8-compatible text. The actual process of replacing any potentially identifying occurrence of names, dates or some such in each text is realized by two services – AHD and INCEpTION. Whereas the former is a proprietary healthcare text mining and machine learning platform for analyzing large amounts of patient data,[5] the latter is publicly available and an open-source Web-based platform to facilitate the task of (mainly semantically) annotating text corpora under various scenarios.[6] We employ INCEpTION as the main hub, upload the documents to it and query AHD during the annotation process where it functions as a tag recommender system and each instance of a potentially identifying text stretch found by the latter can be accepted or rejected by the annotators via the former. Both services utilize the UIMA framework[7] and communicate effectively by transferring CAS files between them, an XMI format that stores both text and accompanying annotations.

3.3. Annotation of the GeMTeX Gold Standard

At all GeMTeX annotation sites, we plan to work with a group of around five medical students who will review and curate automatic pre-annotations from AHD. It is mandatory that all students involved have passed the first medical exam and are studying at least in the 5th semester. The annotation groups are supervised by 1-2 (or more) study assistants who are supervised by a co-lead of senior staff from Munich and Leipzig.

3.4. Protected Health Information – Category System and Annotation Guideline

The de-identification targets that apply to GeMTeX are based on the Protected Health Information (PHI) category system of the US *Health Insurance Portability and Accountability Act* (HIPAA)[8] and on adaptations to German clinical reporting habits and legal requirements [9,17,22,23]. These will be implemented in the de-identification component of the AHD. Table 1 contains the names of the Type System for de-identification we employ.

4. Results of the Pilot Study

4.1. Annotation Campaign

Data: We ran the annotation campaign on the entire GRASSCO text corpus. The data was pre-annotated by AHD (v6.23.0). **Staff**: In Leipzig and Erlangen, for the manual tasks a team of six annotators (students of medicine) and two annotation curators (in Leipzig: one with background in linguistics, the other with background in clinical coding; in Erlangen: one physician and one scientific researcher) collaborated. The group was supervised by a person with a scientific background of clinical NLP. **Process**: Annotators and curators got acquainted with the PHI annotation guideline and the annotation tool INCEpTION (v30.2). In Leipzig, all annotators annotated all 63 GRASSCO documents based on a preliminary version of the annotation guide (without `Age`, `Profession`, `Name Title` and simplified name roles). During the annotation process, questions

[5] https://averbis.com/health-discovery/ (accessed: June 25, 2024)

[6] https://inception-project.github.io/ (accessed: June 25, 2024)

[7] https://uima.apache.org/ (accessed: June 25, 2024)

[8] https://www.hhs.gov/hipaa/ (accessed: June 25, 2024)

were collected and answered by the curators and the manager during meetings. With these answers in mind the annotation guideline was refined as the basis for the second round. Questions and open issues were also collected for an update of the annotation guideline. Erlangen used this updated guideline and started an annotation champaign with two iterations where the 63 documents were split in half for each iteration. **Agreement**: In Leipzig, the first round on the whole of GRASSCO with the simplified category system yielded an Inter-Annotator Agreement (IAA) of (Krippendorff's) $\alpha = 0.97$, the second one with the final category system resulted in $\alpha = 0.97$ [24]. In Erlangen, an IAA of $\alpha = 0.95$ was measured for the first and $\alpha = 0.97$ for the second iteration (both on the final category system), respectively. These results coincide align wellwith other PHI annotation campaigns [9,25,26]. (Mostly, divergences could be traced to confusions about the `Name Title` type.) **Curation**: After the annotation cycles, the data set was curated by the curators and delivered as the final version of the corpus.

4.2. Final Corpus

GRASSCO consists of 44 K raw text tokens, the curated version adds 1438 PHI annotations to this corpus (see Table 1). Roughly 3% of the tokens in GRASSCO refer to PHI mentions. About half of the annotations are `Date` annotations followed by `Name Patient` with around 12 %, `Name Doctor` about 11 %, and `Name Title` with slightly less than 10 % of all annotations. These values correspond to those reported in alternative annotation campaigns for clinical texts (e.g., 3000PA [9,22]).

Table 1. De-Identification Type System and Quantitative Breakdown of the Annotation Results for GRASSCO

PHI Category	count	μ	σ	min	max	av. ann.
NAME PATIENT	166	2.63	2.23	1	10	11.54%
NAME DOCTOR	154	2.57	1.82	1	8	10.71%
NAME RELATIVE	1	1.0	<0.01	1	1	0.07%
NAME USERNAME	1	1.0	<0.01	1	1	0.07%
NAME TITLE	139	2.4	1.59	1	8	9.67%
NAME EXTERN	1	1.0	<0.01	1	1	0.07%
DATE	694	11.02	9.70	2	55	48.26%
AGE	23	1.35	0.79	1	3	1.60%
LOCATION STREET	36	1.89	0.94	1	4	2.50%
LOCATION ZIP	59	1.97	1.07	1	4	4.10%
LOCATION CITY	38	1.73	0.94	1	4	2.64%
LOCATION COUNTRY	2	1.0	<0.01	1	1	0.14%
LOCATION HOSPITAL	36	1.2	0.55	1	3	2.50%
LOCATION ORGANIZATION	2	1.0	<0.01	1	1	0.14%
ID	58	1.93	1.14	1	5	4.03%
CONTACT PHONE	18	1.5	0.90	1	4	1.25%
CONTACT FAX	7	1.17	0.41	1	2	0.49%
CONTACT EMAIL	1	1.0	<0.01	1	1	0.07%
PROFESSION	2	1.0	<0.01	1	1	0.14%

5. Discussion

Neither the GDPR, nor any other applicable legal provisions stipulate minimum requirements for clinical documents in Germany which directly identifying personal information (PHI) must be removed as part of proper deidentification. Therefore, the de-identification approach we adhere to has its roots in the *Health Insurance Portability and Accountability Act* (HIPAA) from the US (for a survey, see [20]). The underlying PHI category system, which contains a list of 18 PHI types, has been adapted to German category demands. In addition, we added other potentially PHI-sensitive information, not included in HIPAA, such as profession, to our type system.

In some rare cases the semantical context allows a precise re-identification of a person. These cases refer mostly to persons of public interest where a fit between semantic information from the document and (commonsense) background knowledge is straightforward. To cope with such phenomena we introduced the PHI category ‘other’ for such special subjects. Each document tagged with ‘other’ should be screened in depth. We suggest, as a safety measure, deleting such a document from the corpus.

Based on the defined de-identification measures, we are confident that we meet the legal requirements with regard to anonymization. This is due to the fact that the GDPR does *not* require that the identification of a person is completely ruled out. According to recital 26 of the GDPR, in order to determine whether a natural person is identifiable, account should be taken of all the means likely to be used by the controller or another person to identify the natural person under scrutiny directly or indirectly. In determining whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the cost of identification and the time required, taking into account the technology available at the time of processing and technological developments. In closing, it should be noted that the MII Broad Consent also contains a clarification regarding the potential risk of identification (in particular due to criminal activities or on the basis of record linkage with publicly accessible information).

6. Conclusion

We introduced the architecture of the de-identification process for the GeMTeX corpus, which aims at identifying and replacing PHI items in real German-language clinical documents. This architecture was tested in a pilot study on the GraSSCo corpus, a synthetic text collection of German clinical documents.

The pilot study we conducted creates added value on two dimensions. First, the enhanced version of GraSSCo, GraSSCo$_{\text{PHI}}$, constitutes the first publicly available clinical text corpus for the German language that is with PHI metadata and is thus open for use by external NLP researchers and developers without any restrictions. Second, GraSSCo$_{\text{PHI}}$ will serve as a comparison standard for future PHI annotations of the GeMTeX corpus, generated at the six physically distributed annotation hubs of the GeMTeX project in Munich, Leipzig, Dresden, Berlin, Essen and Erlangen.

To manage such a large-scale project like GeMTeX, properly, it seems advisable to test the entire set-up of different annotation tools and processes in a smaller-scaled playground such as GraSSCo. Based on the experience gathered in the annotation project described above, the INCEpTION tool was already updated (release 33.0, 2024/06/11) and the PHI annotation guideline was adapted based on feedback from the annotation crew. The distribution of PHI types and IAA scores we measured on GRASSCO indicate that the synthetic approach from GRASSCO seems to align nicely with the real clinical data we already investigated in the 3000PA campaign within the SMITH project. All essential resources we developed are shared via ZENODO (see footnote 2):

- the PHI annotation guideline for GRASSCO$_{PHI}$ / GeMTeX,
- the annotation metadata, and
- the curated version of the annotation campaign of the GRASSCO$_{PHI}$ text corpus.

In accordance with most international publications concerning PHI, we suggest that the anonymization process cannot be accomplished by automatization only. Providing the highest standard of data privacy manual annotation steps of at least two annotators per text are inevitable. We propose that for statistical quality control at least an over-all α of [22] 0,9 as a quantitative requirement for IAA should be achieved.

Declarations

Ethics: For the synthetic GRASSCO data, no ethical considerations apply.

Conflict of interest: The authors declare no conflict of interest.

Contributions of the authors: CL: paper design, writing of the manuscript, experimental setup; CL, JF, AR: annotation groups management, evaluation & interpretation of experiments; LM, FMei, MB: paper design support; RK: support in legal questions regarding anonymization; UH, RK, FMat: contributions of individual (sub)sections; ALL: proofreading.

Acknowledgement: This work was supported by BMBF within the GeMTeX project under grant 01ZZ2314B (CL, FMei, FMat, UH) and 01ZZ2314A (MB, LM) as well as 01ZZ2314G (JF) and NUM 2.0 under grant 01KX2121 (AR).

We thank Serwar Basch and Richard Eckard de Castilho for their support related to the INCEpTION annotation platform and Peter Klügl for his support related to the Averbis Health Discovery system. Furthermore, we thank all annotation students, Steffen Franke and Oksana Galusch for the data curation in the Data Integration Center of the University Hospital Leipzig, as well as Janina Kind, Frank Richter, Thomas Wendt, Matthias Nüchter and Markus Löffler for their on-going support to organize the annotation team.

References

[1] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57(S 01):e50-e56. https://doi.org/10.3414/ME18-03-0003.

[2] Meineke F, Modersohn L, Loeffler M, Boeker M. Announcement of the German Medical Text Corpus Project (GeMTeX). Stud Health Technol Inform 2023;302:835–6. https://doi.org/10.3233/SHTI230283.

[3] Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O et al. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. ACM Comput. Surv. 2024;56(2):1–40. https://doi.org/10.1145/3605943.

[4] Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc 2023;30(2):340–7. https://doi.org/10.1093/jamia/ocac225.

[5] Modersohn L, Schulz S, Lohr C, Hahn U. GRASCCO - The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. Stud Health Technol Inform 2022;296:66–72. https://doi.org/10.3233/SHTI220805.

[6] Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V et al. Smart Medical Information Technology for Healthcare (SMITH). Methods Inf Med 2018;57(S 01):e92-e105. https://doi.org/10.3414/ME18-02-0004.

[7] Hahn U, Matthies F, Lohr C, Löffler M. 3000PA-Towards a National Reference Corpus of German Clinical Language. Stud Health Technol Inform 2018;247:26–30. https://doi.org/10.3233/978-1-61499-852-5-26.

[8] Hahn U, Modersohn L, Faller J, Lohr C. Final Report on the German Clinical Reference Corpus 3000PA. Stud Health Technol Inform 2024;310:599–603. https://doi.org/10.3233/SHTI231035.

[9] Kolditz T, Lohr C, Hellrich J, Modersohn L, Betz B, Kiehntopf M et al. Annotating German Clinical Documents for De-Identification. Stud Health Technol Inform 2019;264:203–7. https://doi.org/10.3233/SHTI190212.

[10] Lohr C, Luther S, Matthies F, Modersohn L, Ammon D, Saleh K et al. CDA-Compliant Section Annotation of German-Language Discharge Summaries: Guideline Development, Annotation Campaign, Section Classification. AMIA Annu Symp Proc 2018;2018:770–9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371337/.

[11] Lohr C, Modersohn L, Hellrich J, Kolditz T, Hahn U. An Evolutionary Approach to the Annotation of Discharge Summaries. Stud Health Technol Inform 2020; 270:28–32. https://doi.org/10.3233/SHTI200116.

[12] Zenker S, Strech D, Jahns R, Müller G, Prasser F, Schickhardt C et al. National standardisierter Broad Consent in der Praxis: erste Erfahrungen, aktuelle Entwicklungen und kritische Betrachtungen. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024. https://doi.org/10.1007/s00103-024-03878-6.

[13] Lohr C, Eder E, Hahn U. Pseudonymization of PHI Items in German Clinical Reports. Stud Health Technol Inform 2021;281:273–7. https://doi.org/10.3233/SHTI210163.

[14] Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, Matthieu-P. Schapranow. GGPONC 2.0 - The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers. Proceedings of the Thirteenth Language Resources and Evaluation Conference 2022:3650–60. https://aclanthology.org/2022.lrec-1.389/.

[15] Kittner M, Lamping M, Rieke DT, Götze J, Bajwa B, Jelas I et al. Annotation and initial evaluation of a large annotated German oncological corpus. JAMIA Open 2021;4(2):ooab025. https://doi.org/10.1093/jamiaopen/ooab025.

[16] Richter-Pechanski P, Wiesenbach P, Schwab DM, Kiriakou C, He M, Allers MM et al. A distributable German clinical corpus containing cardiovascular clinical routine doctor's letters. Sci Data 2023;10(1):207. https://doi.org/10.1038/s41597-023-02128-9.

[17] Richter-Pechanski P, Riezler S, Dieterich C. De-Identification of German Medical Admission Notes. Stud Health Technol Inform 2018;253:165–9. https://doi.org/10.3233/978-1-61499-896-9-165

[18] Frei J, Kramer F. Annotated dataset creation through large language models for non-english medical NLP. J Biomed Inform 2023;145:104478. https://doi.org/10.1016/j.jbi.2023.104478.

[19] Frei J, Kramer F. German Medical Named Entity Recognition Model and Data Set Creation Using Machine Translation and Word Alignment: Algorithm Development and Validation. JMIR Form Res 2023;7:e39077. https://doi.org/10.2196/39077.

[20] Negash B, Katz A, Neilson CJ, Moni M, Nesca M, Singer A et al. De-identification of free text data containing personal health information: a scoping review of reviews. Int J Popul Data Sci;8(1). https://doi.org/10.23889/ijpds.v8i1.2153.

[21] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations 2018:5–9. https://aclanthology.org/C18-2002/.

[22] Tobias Kolditz, Christina Lohr, Luise Modersohn, Udo Hahn. Annotationsleitlinien für deutschsprachige Medizintexte - Teil 2: Annotation von personenidentifizierenden PHI-Attributen. Zenodo; 2023. https://doi.org/10.5281/zenodo.7707882

[23] Richter-Pechanski P, Amr A, Katus HA, Dieterich C. Deep Learning Approaches Outperform Conventional Strategies in De-Identification of German Medical Reports. Stud Health Technol Inform 2019;267:101–9. https://doi.org/10.3233/SHTI190813.

[24] Krippendorff K. Estimating the Reliability, Systematic Error and Random Error of Interval Data. Educational and Psychological Measurement 1970;30(1):61–70. https://doi.org/10.1177/001316447003000105.

[25] Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015;58 Suppl(Suppl):S11-S19. https://doi.org/10.1016/j.jbi.2015.06.007.

[26] Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. J Biomed Inform 2017;75S:S4-S18. https://doi.org/10.1016/j.jbi.2017.06.011.