



GeMTeX

De-Identifikation deutschsprachiger klinischer Texte in GeMTeX – Annotationsleitlinien für identifizierende Merkmale

Christina Lohr, Frank Meineke, Franz Matthies, Udo Hahn

(IMISE Universität Leipzig);

Steffen Franke, Oksana Galusch (DIZ Universitätsklinikum Leipzig);

Peter Klügl (Averbis);

Jakob Faller, Andrea Riedel (Universitätsklinikum Erlangen);

Luise Modersohn, Martin Boeker (TU München);

Fabian Prasser (BIH@Charité);

Raziye Sari, Phillip Richter-Pechanski (Universitätsklinikum Heidelberg)

Erste Version – <https://doi.org/10.5281/zenodo.11502329>

9. September 2024

Inhaltsverzeichnis

1	Einleitung	2
2	Begriffe	3
3	Einordnung der Prozesse im GeMTeX-Projekt	3
4	Typen-System für De-Identifikation (PHI-Elemente)	4
4.1	Namen von Personen [NAME]	4
4.2	Datumsangaben [DATE]	5
4.3	Altersangaben [AGE]	5
4.4	Adressangaben [LOCATION]	6
4.5	Kennung (Identifier) [ID]	7
4.6	Kontaktdaten [CONTACT]	7
4.7	Beruf [PROFESSION]	7
4.8	Other [OTHER]	7
4.9	Weitere Hinweise zur manuellen Annotation	8
4.9.1	Was bei der Annotation beachtet werden soll	8
4.9.2	Was bei der Annotation nicht beachtet werden soll	8
	Referenzen	9

1 Einleitung

Dieses Dokument beschreibt die erste öffentliche Version *De-Identifikation deutschsprachiger klinischer Texte in GeMTeX – Annotationsleitlinien für identifizierende Merkmale*.

Das Ziel des Projekts *GeMTeX (German Medical Text Corpus)* besteht darin, eine Sammlung von klinischen Arztbriefen bzw. ein Textkorpus zu erstellen (1). Dieses Korpus soll zum Ende des Projekts für das Training von Modellen der künstlichen Intelligenz bereitgestellt werden. Für eine vollumfängliche Nutzung dieser Ressource muss ein hohes Maß an Datenschutz gewährleistet werden.

Dieses Dokument ist im Rahmen von Annotationsarbeiten des GRASSCO-Textkorpus entstanden (2) und eine Ergänzung zur Veröffentlichung (3):

Lohr C, Matthies F, Faller J, Modersohn L, Riedel A, Hahn U, Kiser R, Boeker M, Meineke F. De-Identifying GRASCCO - A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus. Stud Health Technol Inform. 2024 Aug 30;317:171-179. doi: 10.3233/SHTI240853 . PMID: 39234720 .

Für die Definition der zu betrachtenden Strukturen sind Elemente der Kriterien der *Protected Health Information (PHI)* des *Health Insurance Portability and Accountability Acts (HIPAA)*¹ sowie den Vorarbeiten von (4; 5; 6; 7) und der Komponente zur De-Identifikation der Software *Averbis Health Discovery*² entstanden. Zu Beginn wird sowohl in Begriffe als auch in den Ablauf der Arbeiten eingeführt. Es folgt eine Auflistung der zu de-identifizierenden Merkmale, die im Rahmen des gesamten Prozesses der De-Identifikation unkenntlich gemacht werden.

¹<https://www.hhs.gov/hipaa/index.html> (Letzter Zugriff: 9. September 2024)

²<https://averbis.com/de/health-discovery/> (Letzter Zugriff: 9. September 2024)

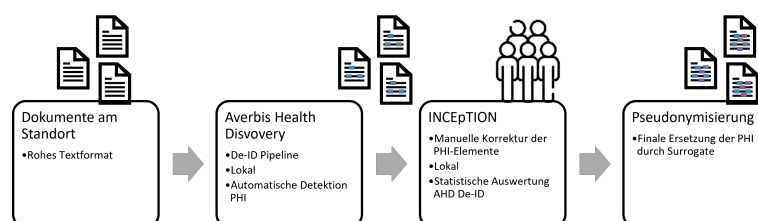
2 Begriffe

- **Textkorpus:** Sammlung von Texten, hier klinische Dokumente.
- **Token:** Kleinster Bestandteil eines Textes, der für die Annotation relevant ist, z.B. einzelne Wörter und Satzzeichen voneinander getrennt.
- **Annotat:** Inhaltlich markiertes Textstück (ein einzelnes Token bzw. eine Spanne von Token) in einem Text, zum Beispiel in einem Arztbrief.
- **Entität:** Eindeutig zu bestimmendes Objekt, worüber Informationen gespeichert wird
- **Typ:** Verallgemeinerung einer Entität
- **PHI:** Protected Health Information - zu anonymisierende Information Anonymisierung
- **De-Identifikation:** Erkennung und Entfernung identifizierender Informationen
- **Pseudonymisierung:**
 - Bezeichnet im Kontext dieses Dokuments den Prozess, identifizierende Daten aus Texten zu entfernen und es mit einem nicht-sprechenden Pseudonym zu assoziieren, mit dem der Rückbezug zum Patienten wiederhergestellt werden kann, entsprechend der Mechanismen des Leitfadens zum Datenschutz in medizinischen Forschungsprojekten der TMF (8) um Pseudonymität zu erreichen.
 - Veränderung von personenbezogenen Daten mittels einer Zuordnungsvorschrift derart, dass Einzelangaben ohne Kenntnis oder Nutzung der Zuordnungsvorschrift nicht mehr einer natürlichen Person zugeordnet werden können. (8)

3 Einordnung der Prozesse im GeMTeX-Projekt

Aus dem Datenschutz-Konzept von GeMTeX entnommen.

1. Automatisierte Erkennung von PHI durch die Averbis Health Discovery (AHD)
2. Zweimalige unabhängige Überprüfung der PHI-Annotation durch geschulte menschliche Annotatoren („4-Augen-Prinzip“ unter Nutzung der unten dargestellten Annotations-Plattform INCEPTION (9)³, dabei ggf. Korrektur nicht erkannter PHI (falsch-negativ), fälschlich erkannter PHI (falsch-positive) und falsch typisierter PHI. Die speziell eingewiesenen Annotatoren „sehen“ dabei die noch nicht überschriebenen identifizierenden Merkmale in den Texten, da diese sonst nicht in entsprechender Qualität der PHI-Annotation überprüft werden können. Zusätzlich wird eine Stichprobe der de-identifizierten Texte durch die Annotationsleitung auf die Korrektheit der De-Identifikation überprüft, um eine hohe Qualität zu gewährleisten.
3. Automatisiertes Überschreiben der PHI mit typerhaltenden Termen, die die PHI so verändern, dass eine Zuordnung zu einzelnen Personen nicht mehr möglich ist, eine möglichst weitgehende Anwendung von NLP-Algorithmen aber weiterhin gegeben bleibt. In folgenden Prozessschritten - insbesondere der Annotation - sind die PHI also endgültig entfernt.



³<https://inception-project.github.io/> (Letzter Zugriff: 9. September 2024)

4 Typen-System für De-Identifikation (PHI-Elemente)

4.1 Namen von Personen [NAME]

Die Kategorie **NAME** umfasst alle Arten von Namen von Personen. Vor- und Familiennamen in einer Spanne werden zusammen annotiert. Bezeichnungen von Titeln werden gesondert behandelt. Die Rollen hinter den Namen werden wie folgt unterschieden und sind bei der Annotation zu beachten:

- Seite des Patienten
 - Patient / Patientin selbst
 - Angehörige des Patient / der Patientin
- Seite der Behandlung
 - intern (Arztpersonal, Pflegepersonal und weiteres medizinisches Personal)
 - extern (Nennung von Personen im Zshg. mit Ämtern, Versicherungen, etc.)

Entitäten-Typ	Rolle des Namens	Beispiele
PATIENT	Patient / Patientin	Wir berichten über <i>[Max Mustermann]</i> _{PATIENT} Patientin: <i>[Mustermann, Maxima]</i> _{PATIENT}
RELATIVE	Angehörige	Die Tochter des Pat. Frau <i>[B. Schulze]</i> _{RELATIVE}
DOCTOR	Medizinischen Personal	<i>[Prof. Dr.]</i> _{TITLE} <i>[Moritz Schmidt]</i> _{DOCTOR}
EXT	Nennung von Personen im Zshg. mit Ämtern, Versicherungen, etc.	Frau <i>[Meyer]</i> vom Jugendamt
USERNAME	Username, Sekretariatskürzel	Unser Zeichen: <i>[a.jd]</i>
OTHER	Nennung weiterer Namen	
TITLE	Titel-Bezeichnungen	<i>[Prof. Dr.]</i> _{TITLE} <i>[Moritz Schmidt]</i> _{DOCTOR}

Hinweis zu Bezeichnungen von Titeln und Namen

- Assistenzarzt, Oberarzt, leitender Oberarzt, Chefarzt sind Funktionsbezeichnungen der Krankenhaushierarchie für Aufgabenbereiche, Befugnisse und Vergütungen implizieren und sollen nicht annotiert werden.⁴ Bsp.:
 - Chefarzt *[Prof. Dr. med habil. Dr. h.c. Dr. E.h.]*
 - OA *[Dr. med. M. Schmidt]*
- Titel ohne Bezug zu einem behandelnden Personen sollen ebenso annotiert werden.

⁴siehe auch https://flexikon.doccheck.com/de/Liste_akademischer_Titel_in_der_Medizin (Liste akademischer Titel) (Letzter Zugriff: 9. September 2024)

4.2 Datumsangaben [DATE]

Die Kategorie **DATE** deckt alle möglichen Arten von Datumsangaben ab: Geburtsdaten, Sterbedaten, Datumsangaben zur Aufnahme, Verlegung und Entlassung, ebenso vorhergehende Behandlungen. Es soll nur das Datum annotiert werden und nicht der Bezug dazu. Einzelne Nennung von Zeitangaben, Wochentagen und Monatsangaben sollen nicht annotiert werden.

Entitäten-Typ	Beschreibung	Beispiele
DATE	nur das Datum betrachten und nicht den Bezug	Geburtsdatum: [21.03.2024], Aufnahme: [21/03/2024], verst. am [21. März 2024], Fistel ([01-12/64])

4.3 Altersangaben [AGE]

Mit der Kategorie **AGE** werden alle Altersangaben von Patienten sowie deren Angehörigen annotiert. Es soll nur die Angabe der Zahl annotiert werden, nicht die Einheit in Jahren.

Entitäten-Typ	Rolle des Namens	Beispiele
AGE	Altersangaben	[87] Jahre

4.4 Adressangaben [LOCATION]

Die Kategorie **LOCATION** umfasst alle Arten von Adressinformationen. Alle nicht weiter zuordenbaren Adressangaben werden unter **LocationOther** annotiert, z.B. Postfächer. Diese Kategorie enthält Bezeichnungen von Einrichtungen, z.B. Standorte von Einrichtungen, spezifische (Teil-)Organisationen oder Einrichtungen sowie Organisationseinheiten.

Entitäten-Typ	Rolle des Namens	Beispiele
STREET	Straßenname und Hausnummer	<i>[Musterstraße 1]</i> _{STREET}
CITY	Stadt	<i>[Berlin]</i> _{CITY}
ZIP	Postleitzahl, Sind Landeskennungen Teil der Postleitzahl → Annotation der ges. Spanne einer PLZ	<i>[10115]</i> _{ZIP} <i>[Berlin]</i> _{CITY} <i>[A-1234]</i> _{ZIP} <i>[D-98765]</i> _{ZIP}
COUNTRY	Nennung eines Landes	Urlaub in <i>[Frankreich]</i> _{COUNTRY}
STATE	Nennung eines Staates oder Bundeslandes	Aufenthalt in <i>[Bayern]</i> _{STATE}
HOSPITAL	Nennung einer Einrichtung mit klinischem Bezug (Ort, wo man behandelt werden kann)	<i>[Universitätsklinikum Leipzig]</i> _{HOSPITAL} , <i>Universitätsklinikum in [Leipzig]</i> _{CITY} , <i>[Diakonissenkrankenhaus Berlin]</i> _{HOSPITAL} , <i>[Diakonissenkrankenhaus Berlin]</i> _{HOSPITAL} – Ophthalmologisches Zentrum, <i>[ARCOS-KLINIK FLENSBURG,</i> <i>Akademisches Lehrkrankenhaus der</i> <i>Otto-Waalkes-Universität, Borkum]</i> _{HOSPITAL} <i>[ARCOS-KLINIK FLENSBURG</i> <i>Akademisches Lehrkrankenhaus der</i> <i>Otto-Waalkes-Universität]</i> _{HOSPITAL} in <i>[Borkum]</i> _{CITY} , <i>[Paul-Langerhans-Station]</i> _{HOSPITAL}
ORGANIZATION	Nennung einer Organisation ohne klinischen Bezug (u.a. Versicherungen)	<i>[BZH Heidelberg]</i> _{ORGANIZATION} , ... arbeitet an <i>[Universität Leipzig]</i> _{ORGANIZATION} , Versicherung: <i>[AOK PLUS]</i> _{ORGANIZATION}
LOC_OTHER	Sonstige Adressen	<i>[Postfach 1521]</i> _{LOC_OTHER}

Hinweise zur Annotation PHIHospital und PHIOrganisation

- Nennung für Station / Zimmer besteht aus Nummernbezeichnung, dann keine Annotation als PHIOrganisation oder PHIHospital “5. Kennung [ID]”
- Allgemeine Beschreibungen von klinischen Einrichtungen ohne Nennung von Namen und Orten sollen nicht annotiert werden, z.B.
- “Universitätsklinik für Dermatologie und Venerologie”
 - “Technisches Gymnasium”
 - “Rettungsstelle unseres Zentrums”
 - “Intensivstation unseres Brandverletztenzentrums”

4.5 Kennung (Identifizier) [ID]

Die Kategorie **ID** (Identifizier) umfasst alle Nennungen von Identifikatoren mittels einer Abfolge von Zeichen, einschließlich Nummern. Dazu gehören Patienten-IDs und Fall-ID, ebenso ID's aus medizinischen Subsystemen und Versicherungsnummern. PHIID

Entitäten-Typ	Rolle des Namens	Beispiele
ID	Patienten-IDs, Fall-IDs, Zimmer- und Stationsnummern (nur die Nummern), sonstige Identifikationsnummern, IBAN-Nummern, Studienpseudonyme	Pat.Nr.: [312654356] _{ID} Fall.Nr.: [71543356] _{ID} , Zimmer-Nr. [312] _{ID} , Station [123] _{ID} , Block-Nr.: [H213578-6] _{ID} , Konto: [DE0212030000000202051] _{ID} , im Rahmen der Darmkrebsstudie nahm der Patient als [S05NL-ZKA06] _{ID} teil

4.6 Kontaktdaten [CONTACT]

Die Kategorie **CONTACT** deckt Informationen ab, die Kontaktinformationen darstellen. Dies umfasst Telefonnummern, E-Mail-Adressen, Fax-Nummern, Internetadressen sowie IP-Adressen.

Entitäten-Typ	Rolle des Namens	Beispiele
PHONE	Telefonnummer	Tel. [02345-12345] _{PHONE}
EMAIL	E-Mail-Adresse	Mail: [max.muster@beispiel.de] _{EMAIL}
FAX	Faxnummer	Fax. [04324-65432] _{FAX}
URL	URL einer Homepage	Internet: [www.beispiel.de] _{URL}

4.7 Beruf [PROFESSION]

Entitäten-Typ	Rolle des Namens	Beispiele
PROFESSION	Bezeichnungen von Berufen im Zusammenhang mit dem Patienten, ohne Nennung eines Status (z.B. arbeitslos, berentet, Teilzeit)	[Verkäuferin] _{PROFESSION} , [Pfleger] _{PROFESSION} , [Bergbauarbeiter] _{PROFESSION} , [Monteur] _{PROFESSION} , [Friseurin] _{PROFESSION}

4.8 Other [OTHER]

Mit der Kategorie **OTHER** werden alle schützenswerten Passagen annotiert, für die keine der obigen Kategorien zutrifft.

4.9 Weitere Hinweise zur manuellen Annotation

4.9.1 Was bei der Annotation beachtet werden soll

- Behandlung von **Fehlern (Rechtschreibung, vertippt)**: so damit umgehen, als wäre es korrekt.
- **Abkürzungen** so behandeln, als sei die Abkürzung ausgeschrieben, z.B. abgekürzte Namen ("M. Muster").
- Vornamen von Kindern bzw. Vornamen im pädiatrischen Kontext normal als Namen behandeln → PHIPatient.
- Von Tieren abgeleitete Namen, die durch Vorannotation nicht erkannt sind, trotzdem annotieren.

4.9.2 Was bei der Annotation nicht beachtet werden soll

- **Funktionswörter** wie Präpositionen (z.B. von, mit, aus, zu, in), Artikel (z.B. der, die, das, ein), Pronomen (z.B. unser, sein, sie), Konjunktionen etc. am Rand (am Anfang oder Ende einer Annotation)
 - **Ausnahme**: Namensbestandteile Teil des Annotats (z.B. [von der Leyen])
- Kategorisierungen, Attribute und nähere Beschreibungen eines Individuums oder einer Ortsangabe (z.B. "geboren am", "wohnhafte in")
- **Zeitangaben ohne Datum**
 - Tageszeitangaben (z.B. "15:04", "3.30" Uhr)
 - Wochentage (z.B. "jeden Freitag", "jeden Mittwoch")
- **Konsile, Boards** (keine festen medizinische Einheiten, die wiedererkennbaren Namen haben könnten) **Eponyme (Eigennamen) von Skalen, Diagnosen, Prozeduren, etc.** sollen nicht beachtet werden (aber fälschlicherweise automatisiert markiert sein können) → Beispiele: "Barthel-Score", "Von-Willebrand-Faktor", "Operation nach Hartmann" "Wernicke-Enzephalopathie", "Morbus Sudeck", "Stadium nach Rutherford", "Faktor V Leiden"
- Studienprotokoll-Bezeichnungen inkl. Jahreszahl (z.B. "NB2004")
- **Familienstand** (ledig, verheiratet, geschieden, verwitwet) soll erhalten bleiben.
- ICD-10-Codes und andere Nennungen von medizinischen Terminologie-Systemen sollen nicht beachtet werden.

Referenzen

- [1] Frank A. Meineke, Luise Modersohn, Markus Loeffler, and Martin Boeker. Announcement of the german medical text corpus project (gemtex). In Maria Hägglund, Madeleine Blusi, Stefano Bonacina, Lina Nilsson, Inge Cort Madsen, Sylvia Pelayo, Anne Moen, Arriel Benis, Lars Lindsköld, and Paris Gallos, editors, *Caring is Sharing - Exploiting the Value in Data for Health and Innovation - Proceedings of MIE 2023, Gothenburg, Sweden, 22 - 25 May 2023*, volume 302 of *Studies in Health Technology and Informatics*, pages 835–836. IOS Press, 2023.
- [2] Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. GRASCCO - the first publicly shareable, multiply-alienated german clinical text corpus. In Rainer Röhrig, Niels Grabe, Verena S. Hoffmann, Ursula Hübner, Jochem König, Ulrich Sax, Björn Schreiweis, and Martin Sedlmayr, editors, *German Medical Data Sciences 2022 - Future Medicine: More Precise, More Integrative, More Sustainable! - Proceedings of the Joint Conference of the 67th Annual Meeting of the German Association of Medical Informatics, Biometry, and Epidemiology e.V. (gmds) and the 14th Annual Meeting of the TMF - Technology, Methods, and Infrastructure for Networked Medical Research e.V. 2022 online in Kiel, Germany*, volume 296 of *Studies in Health Technology and Informatics*, pages 66–72. IOS Press, 2022.
- [3] Christina Lohr, Franz Matthies, Jakob Faller, Luise Modersohn, Andrea Riedel, Udo Hahn, Rebekka Kiser, Martin Boeker, and Frank Meineke. De-identifying grascoco—a pilot study for the de-identification of the german medical text project (gemtex) corpus. In Rainer Röhrig, Niels Grabe, Verena S. Hoffmann, Ursula Hübner, Klaus Jung, Ulrich Sax, and Antonia Zapf Carsten Oliver Schmidt, Martin Sedlmayr, editors, *Health – Thinking, Researching and Acting Together. Proceedings of the 69th Annual Meeting of the German Association of Medical Informatics, Biometry, and Epidemiology e.V. (gmds) 2024 in Dresden, Germany*, volume 317 of *Studies in Health Technology and Informatics*, pages 171–179. IOS Press, 2024.
- [4] Phillip Richter-Pechanski, Stefan Riezler, and Christoph Dieterich. De-identification of german medical admission notes. In Ursula Hübner, Ulrich Sax, Hans-Ulrich Prokosch, Bernhard Breil, Harald Binder, Antonia Zapf, Brigitte Strahwald, Tim Reißbarth, Niels Grabe, and Anke Schöler, editors, *German Medical Data Sciences: A Learning Healthcare System - Proceedings of the 63rd Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology (GMDS e.V.) 2018 in Osnabrück, Germany, 2-6 September 2018, GMDS 2018*, volume 253 of *Studies in Health Technology and Informatics*, pages 165–169. IOS Press, 2018.
- [5] Phillip Richter-Pechanski, Ali Amr, Hugo A. Katus, and Christoph Dieterich. Deep learning approaches outperform conventional strategies in de-identification of german medical reports. In Rainer Röhrig, Harald Binder, Hans-Ulrich Prokosch, Ulrich Sax, Irene Schmidtman, Susanne Stolpe, and Antonia Zapf, editors, *German Medical Data Sciences: Shaping Change - Creative Solutions for Innovative Medicine - Proceedings of the 64th Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology (gmds e.V.) 2019 in Dortmund, Germany - GMDS 2019*, volume 267 of *Studies in Health Technology and Informatics*, pages 101–109. IOS Press, 2019.
- [6] Tobias Kolditz, Christina Lohr, Johannes Hellrich, Luise Modersohn, Boris Betz, Michael Kiehntopf, and Udo Hahn. Annotating german clinical documents for de-identification. In

Lucila Ohno-Machado and Brigitte Séroussi, editors, *MEDINFO 2019: Health and Well-being e-Networks for All - Proceedings of the 17th World Congress on Medical and Health Informatics, Lyon, France, 25-30 August 2019*, volume 264 of *Studies in Health Technology and Informatics*, pages 203–207. IOS Press, 2019.

- [7] Christina Lohr, Elisabeth Eder, and Udo Hahn. Pseudonymization of PHI items in german clinical reports. In John Mantas, Lacramioara Stoicu-Tivadar, Catherine E. Chronaki, Arie Hasman, Patrick Weber, Paris Gallos, Mihaela Marcella Vida, Emmanouil Zoulias, and Oana Sorina Chirila, editors, *Public Health and Informatics - Proceedings of MIE 2021, Medical Informatics Europe, Virtual Event, May 29-31, 2021*, volume 281 of *Studies in Health Technology and Informatics*, pages 273–277. IOS Press, 2021.
- [8] Klaus Pommerening, Johannes Drepper, Krister Helbing, and Thomas Ganslandt. *Leitfaden zum Datenschutz in medizinischen Forschungsprojekten: Generische Lösungen der TMF 2.0*. Medizinisch Wissenschaftliche Verlagsgesellschaft, 2015.
- [9] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In Dongyan Zhao, editor, *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics.