Gefördert durch:

Bundesministerium
für Forschung, Technologie
und Raumfahrt

MEDIZIN
INFORMATIK
INITIATIVE

GeMTeX

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

2. Version (07. September 2025)

Autoren und Autorinnen:

Christina Lohr, Frank Meineke, Franz Matthies, Udo Hahn (Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig); **Steffen Franke, Oksana Galusch** (Datenintegrationszentrum Universitätsklinikum Leipzig);
Peter Klügl (Averbis GmbH);
Jakob Faller (Medizinisches Zentrum für Informations- und Kommunikationstechnik (MIK) Datenintegrationszentrum (DIZ), Universitätsklinikum Erlangen); **Andrea Riedel** (Medizinisches Zentrum für Informations- und Kommunikationstechnik (MIK) Datenintegrationszentrum (DIZ), Universitätsklinikum Erlangen; Friedrich-Alexander-Universität Erlangen-Nürnberg);
Luise Modersohn, Martin Boeker, Justin Hofenbitzer, Raffael Bild (Institut für Künstliche Intelligenz (KI) und Informatik in der Medizin, TUM School of Medicine and Health, TU München);
Fabian Prasser, Jutta Romberg (Berlin Institute of Health (BIH) at Charité, Universitätsmedizin Berlin);
Jazia Omeirat (Central IT Department, Data Integration Center,
Institute for Artificial Intelligence in Medicine, Universitätsklinikum Essen);
Aliaksandra Shutsko, Abanoub Abdelmalak (ZB MED - Informationszentrum Lebenswissenschaften);
Markus Wolfien (Institut für Medizinische Informatik und Biometrie (IMB), Medizinische Fakultät Carl Gustav Carus der TU Dresden), Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI); **Hung Manh Nguyen** (Institut für Medizinische Informatik und Biometrie (IMB), Medizinische Fakultät Carl Gustav Carus der TU Dresden);
Raziye Sari, Phillip Richter-Pechanski, Marvin Seiferling, Christoph Dieterich
(Klaus-Tschira-Institut für Computerkardiologie, Universitätsklinikum Heidelberg)

https://doi.org/10.5281/zenodo.15747389

Inhalt

https://doi.org/10.5281/zenodo.15747389

1. Einleitung

Dieses Dokument beschreibt die zweite 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 (Meineke et al. 2023). 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 (Modersohn et al. 2022) und eine Weiterentwicklung zu (Lohr et al. 2024b) von (Lohr et al. 2024a) und eine Ergänzung zur Publikation (Lohr et al. 2025):

Lohr C, Faller J, Riedel A, Nguyen HM, Wolfien M, Hofenbitzer J, Modersohn L, Romberg J, Prasser F, Omeirat J, Wen Y, Galusch O, Hahn U, Seiferling M, Dieterich C, Klügl P, Matthies F, Kind J, Boeker M, Löffler M, Meineke F. GeMTeX's De-Identification in Action: Lessons Learned & Devil's Details. Stud Health Technol Inform. 2025 Sep 3;331:274-282. doi: 10.3233/SHTI251406. PMID: 40899551.

Für die Definition der zu betrachtenden Strukturen sind Elemente der Kriterien der *Protected Health Information (PHI)* des *Health Insurance Portabilty and Accountability Acts (HIPAA)*[1] sowie den Vorarbeiten von (Richter-Pechanski et al. 2018; Richter-Pechanski et al. 2019; Kolditz et al. 2019; Tobias Kolditz et al. 2023) 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
[2] https://averbis.com/de/health-discovery/

2. Begriffe

- **Textkorpus:** Sammlung von maschinenlesbaren Texten, in GeMTeX sind dies deutschsprachige klinische Dokumente (z.B. Arztbriefe).
- **Token**: Kleinster Bestandteil eines Textes, der für die Annotation relevant ist, z.B. einzelne Wörter, Zahlen und Interpunktionszeichen voneinander getrennt.
- **Entität:** Eindeutig zu bestimmende, medizinisch relevante (hier: de-identifikations-relevante) Textpassage (Token(folge)), die PII-Informationen enthält (etwa ein konkreter Personenname, ein spezifisches Geburtsdatum, eine bestimmte E-Mail-Adresse).
- **Annotationstyp:** Verallgemeinerung einer Entität zu einer begrifflichen Klasse (z.B. Personenname, Geburtsdatum, E-Mail-Adresse).
- **Typensystem**: die Menge aller Annotationstypen.
- **Annotat**: Inhaltlich markierte Textspanne (eine Tokenfolge, ein einzelnes Token oder ein Teil eines Tokens) in einem Text, hier in einem Arztbrief.
- **PHI:** „Protected Health Information" - zu de-identifizierende Informationseinheit in einem Text (Token(folge)).
- **PII: „**Personally Identifiable Information"
- **IDAT**: „Identifizierende Daten" - identifizierende Daten bezüglich betroffener Personen, im Datenschutzkonzept der Medizininformatik-Initiative definiert
- **De-Identifikation:** Erkennung und Entfernung identifizierender Informationen
- **Anonymisierung**: (GeMTeX-Antrag entnommen)
  - „Umwandlung beliebiger ursprünglicher PHI-Daten in eine beliebige, meist künstliche, semantikfreie Zeichenfolge;
  - Unkenntlichmachung einer identifizierenden Information."
- **Pseudonymisierung:**
  - Bezeichnet den Prozess, identifizierende Daten aus Texten zu entfernen und 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 (Drepper 2015) um Pseudonymität zu erreichen;
  - Veränderung von personenbezogenen Daten mittels einer Zuordnungsvorschrift, indem Einzelangaben ohne Kenntnis oder Nutzung der Zuordnungsvorschrift nicht mehr einer natürlichen Person zugeordnet werden können (Drepper 2015).
  - Im Kontext dieses Dokuments und dem GeMTeX-Antrag entnommen: „Umwandlung eines ursprünglichen PHI-Elements in ein künstliches (synthetisches), aber typ-konformes und natürlich wirkendes Surrogat, das die Grundbedeutung der ursprünglichen Zeichenfolge beibehält, aber eine andere Bezeichnung als Referenz verwendet."

3. Ablauf: Von identifizierenden Merkmalen zum Pseudonym und Einordnung der manuellen Schritte

(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[3], dabei ggf. Korrektur nicht erkannter PHI[4] (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.

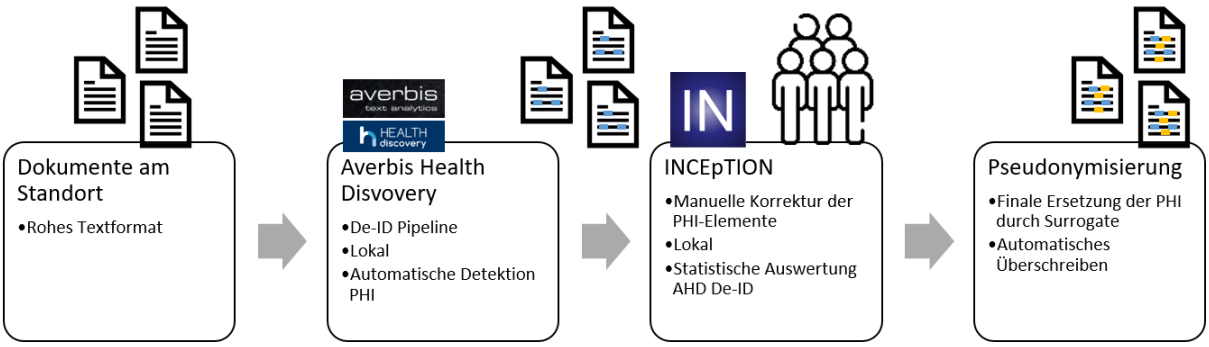

Abb. 1. - Prozess der De-Identifizierung im GeMTeX-Projekt

[3] https://inception-project.github.io/
[4] Deidentifizierende Merkmale sind im GeMTeX-Antrag als PHI bezeichnet, werden hier jedoch weiterhin als PII benannt.

4. Typen-System für die De-Identifikation (PII-Elemente)

1. Namen von Personen [NAME]

Der Typ NAME umfasst alle Arten von Namen von Personen. Vor- und Familiennamen werden in einer Spanne zusammen annotiert. Bezeichnungen von Titeln werden gesondert behandelt. Die sozialen und funktionalen Rollen, durch die Namen weiter charakterisiert sind, werden wie folgt unterschieden und sind bei der Annotation zu beachten:

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

Entitäten-Typ	**Beschreibung der Rolle**	**Beispiele**
NAME_PATIENT	Patient / Patientin (auch Ärzte oder Personen mit akadem. Titeln können Patienten sein)	*Wir berichten über [Max Mustermann]*$^{NAME\_PATIENT}$ *Patientin: [Mustermann, Maxima]*$^{NAME\_PATIENT}$ *… Frau [Schmidt]*$^{NAME\_PATIENT}$ *zeigte …,* *… Herr [H. Sudeck]*$^{NAME\_PATIENT}$ *kam in die Sprechstunde …*
NAME_RELATIVE	Angehörige von Patienten und Patientinnen	Die Tochter des Pat., Frau *[B. Schulze]*$^{NAME\_RELATIVE}$, wurde benachrichtigt. Frau *[M. Müller]*$^{NAME\_RELATIVE}$, die Tochter der Pat., wurde informiert.
NAME_DOCTOR	internes und externes Personal, das am Patienten behandelt sowie Personal im medizinischen Kontext	Chefarzt *[Prof. Dr.]*$^{NAME\_TITLE}$ *[Moritz Schmidt]*$^{NAME\_DOCTOR}$ *Die abschließende Stellungnahme von Frau [PD Dr.]*$^{NAME\_TITLE}$ *[Schmidt]*$^{NAME\_DOCTOR}$*,* *Referenzpathologie [Charite Berlin]*$^{LOCATION\_HOSPITAL}$
NAME_EXT	*abstrakte Personengruppen* Personen, die nicht am Patienten behandeln, Bsp. von Ämtern, Verwaltungen, Versicherungen, gesetzl. Betreuende, Klinik-Vorstände	Frau *[Meyer]*$^{NAME\_EXT}$ vom Jugendamt Medizinischer Vorstand: *[Max Müller]*$^{NAME\_EXT}$ Betreuer: *[E. Schmitt]*$^{NAME\_EXT}$
NAME_USERNAME	Username, Sekretariatskürzel, Kunstnamen für Systemlogins	Unser Zeichen: *[a.jd]*$^{NAME\_USERNAME}$*, [lohrc]*$^{NAME\_USERNAME}$
NAME_TITLE	akademische Titel-Bezeichnungen	*Chefarzt [Prof. Dr.]*$^{NAME\_TITLE}$ [Moritz Schmidt]$^{NAME\_DOCTOR}$ *Assistenzarzt [Dr. med.]*$^{NAME\_TITLE}$ [M. Muster]$^{NAME\_DOCTOR}$ *… Herr* [Dr. rer.-nat.]$^{NAME\_TITLE}$ *[H. Sudeck]*$^{NAME\_PATIENT}$ *kam in die Sprechstunde …* *[Dr. medic (IM Temesch)]*$^{NAME\_TITLE}$ *[Adrian Popescu]*$^{NAME\_DOCTOR}$

Hinweise

- Titel / *NAME_TITLE*:
  - Assistenzarzt, Oberarzt, leitender Oberarzt, Chefarzt sind Funktionsbezeichnungen innerhalb der Krankenhaushierarchie (auch ausländische Bezeichnungen) für Aufgabenbereiche, Befugnisse und Vergütungen und sollen **nicht** annotiert werden[5] → siehe auch 5.3.) Bsp.:
    - Chefarzt *[Prof. Dr. med. habil. Dr. h.c. Dr. E.h.]*NAME_TITLE *[Max Müller]*NAME_DOCTOR
    - *[Dipl. med.]*NAME_TITLE und *[Dipl. med. (univ.)]*NAME_TITLE
    - OA *[Dr. med.]*NAME_TITLE *[M. Schmidt]*NAME_DOCTOR
  - Privatdozent / PD soll annotiert werden:
    - *[PD Dr.med]*NAME_TITLE *[M. Schmidt]*NAME_DOCTOR
  - Universitätsprofessor / Univ.-Prof. soll annotiert werden (da offizieller akademischer Titel):
    - *[Univ.-Prof. Dr. med.]*NAME_TITLE *[U. Stock]*NAME_DOCTOR
  - Akad. Titel ohne medizinischem Bezug von Personen (auch Patienten) sollen ebenfalls mit annotiert werden:
    - Patient *[Dr. rer-nat.]*NAME_TITLE *[Tim Schmidt]*NAME_PATIENT stellte sich vor.
- Titel-Bezeichnungen können auch hinter dem Namen stehen:
  - *[Dr. med.]*NAME_TITLE *[László Kanyuk]*NAME_PATIENT *[PhD]*NAME_TITLE
- gegenderte Titel-Bezeichnungen: *[Dr.in]*NAME_TITLE, Frau *[Dr.]*NAME_TITLE
- Namen von Klinik-Einrichtungen zählen hier nicht dazu, siehe LOCATION.
- Vornamen von Kindern (die selbst Patient sind) bzw. Vornamen im pädiatrischen Kontext sollen normal wie Vollnamen behandelt werden → *NAME_PATIENT*:
  - Die 18 Monate alte *[Hanna]*NAME_PATIENT wurde von ihren Eltern zur Impfung in der Sprechstunde vorgestellt.
- Von Tieren abgeleitete Namen von Personen aller (Entitäten-Typen), die durch Vorannotation nicht erkannt sind, sollen annotiert werden:
  - Patient *[Tony Fuchs]*NAME_PATIENT kam in die Sprechstunde.
  - gezeichnet *[Dr. med.]*NAME_TITLE *[Caroline Wolf]*NAME_PATIENT

- NAME_DOCTOR / NAME_EXT: bei absoluten Grenzfällen NAME_DOCTOR

- Ist ein Arzt bzw. eine Ärztin im Briefkopf genannt, z.B. Klinikvorstand, Klinikdirektoren, und in Unterschriftszeile als Teil der behandelnden Ärzte: Person in Rolle annotieren und ggf. mit unterschiedlichen Rollen annotieren

- Flektierte Namen mit Genitiv-S sollen als Token möglichst vollständig annotiert werden, z.B. *[Marijas]*, Frau *[Müllers]* Sohn …

- Nennung von Sekretariaten: „ ... Vereinbaren Sie bitte einen Termin zur Wiedervorstellung telefonisch bei *[Silke Müller]*NAME_EXT …"

[5]siehe Liste akademischer Titel: https://flexikon.doccheck.com/de/Liste_akademischer_Titel_in_der_Medizin

2. Datumsangaben [DATE]

Der Typ *DATE* deckt alle möglichen Arten von eindeutig identifizierenden Datumsangaben ab:

Das Geburtsdatum eines Patienten bzw. einer Patientin sowie das Sterbedatum (falls genannt) soll gesondert betrachtet werden. Alle weiteren Datumsangaben, zum Beispiel für Aufnahme, Verlegung und Entlassung und Daten vorhergehender Behandlungen. Es soll **nur** das Datum markiert werden und nicht der Bezug dazu.

Es sollen nur absolute Datumsangaben annotiert werden, die eindeutige Informationen über einen Tagesangabe liefern. Relationale Datumsangaben sollen mit Ausnahme von reinen Monats-Nennungen ignoriert werden (siehe auch Hinweise).

Entitäten-Typ	Beschreibung	Beispiele
DATE_BIRTH	Nennung des Geburtsdatums des Patienten bzw. der Patientin	Geburtsdatum: *[21.03.2024]*$^{DATE\_BIRTH}$
DATE_DEATH	Nennung des Sterbedatums des Patienten bzw. der Patientin	verst. am *[21. März 2024]*$^{DATE\_DEATH}$
DATE	alle weiteren Datumsangaben, hier nur das Datum betrachten und nicht den Bezug	Aufnahme: *[21/03/2024]*^{DATE}
	Zeitspannen möglichst nicht zusammenhängend annotieren, auch Nennung von Zeitspannen innerhalb eines Tokens (technisch gesehen), zwei Spannen bei Unterbrechung der numerischen Werte durch Worttoken	Fistel (*[01]*^{DATE}*-[12/64]*^{DATE}), am *[06.]*^{DATE}*/[07.11.2024]*^{DATE}, *[19.3.]*^{DATE} bis zum *[7.5.2029]*^{DATE}, *[März 2024]*^{DATE}, *[02]*^{DATE}*-[04/2021]*^{DATE}
	auch ungenaue Datumsangaben beachten (z.B. Jahresangaben zur Diagnosen)	Erstdiagnose *[2024]*^{DATE}
	einzelne Nennung von Monaten und Feiertagen, auch wenn keine Nennung von Jahresangabe	im *[Juni]*^{DATE} letzten Jahres, Diagnose: *[Juni 2024]*^{DATE} Unfall : *[Weihnachten 2023]*^{DATE} Ende *[Januar]*^{DATE} Wiedervorstellung im *[September]*^{DATE} gewünscht.

Hinweise

- Nennungen von Jahreszahlen von Studienprotokollen oder Leitlinien, die im Kontext mit Behandlungen genannt sind, sollen stehen bleiben und nicht annotiert werden, z.B. “UICC *2024”, “*NB*2004”,*
- Definitionen von Tumor-Stadieneinteilungen sind ausgeschlossen („*TNM 2017*”)

- **Zeitangaben sollen nicht annotiert werden:**
  - Tageszeitangaben (z.B. „*15:04"*, „*3.30"* Uhr)
  - Wochentage (z.B. „*jeden Freitag"*, „*jeden dritten Mittwoch"*)
- **Wochentage sollen nicht annotiert werden** (kommen in der Medikation vor).
- **Relationale Datumsangaben** sollen erhalten bleiben und **nicht** annotiert werden. Beispiele:
  1. Die Ehefrau berichtet von seit ca. 3 Jahren nachlassendem Gedächtnis.
  2. Bis vor etwa einem Jahr habe er selbst Bankgeschäfte sehr sorgfältig erledigt.
  3. Die Störung ist vor zwei Jahren aufgetreten.
  4. "Sturz im letzten Jahr" → letzten Jahr sollt **nicht** annotiert werden
  5. Montag, 15.07.2024 → Montag, *[15.07.2024]*^{DATE}
  6. Jahreswechsel
  7. seit Winter 2023 → nur Jahr annotieren → Winter [2023]
  8. Vorerkrankung seit Mitte 80er Jahre → stehen lassen
- Datumsangaben mit Platzhalter → siehe 5.2. was nicht annotiert werden soll

3. Altersangaben [AGE]

Mit dem Typ AGE werden alle Altersangaben ausschließlich von Patienten **in Jahren** annotiert. Es soll nur die Angabe der Zahl (numerisch oder alphabetisch) annotiert werden, nicht die Einheit in Jahren. Angaben zu vorhergehenden Ereignissen (z.B. Erkrankungen, die seit einem bestimmten Alter bereits bestehen) sollen nicht betrachtet werden.
(Das Ziel der Annotation mit AGE besteht darin, Patienten, die älter als 89 Jahre alt sind, zu identifizieren; solche eher seltenen Altersangaben würden eine Re-Identifikation von Personen ermöglichen.[6])
Es wird nur die Angabe der Jahreszahl annotiert, eine Angabe von Tagen und Monaten wird nicht annotiert.

Entitäten-Typ	Beschreibung	Beispiele
AGE	Altersangaben (Altersangaben von weiteren Personen sollen nicht betrachtet werden.)	*[87]*^{AGE} Jahre *[fünf]*^{AGE}jähriger *[Fünfzig]*^{AGE}jährige *[Fünfig]*^{AGE}jährige (auch Fehler beachten) *[49]*^{AGE}jährig *[50]*^{AGE} j. *[64]*^{AGE} Jahre und 2 Monate

Hinweise:
- Im Alter von 74 Jahren erkrankt jetzt [75]^{AGE} Jahre alt.
- Alter: [2 ½]^{AGE} Jahre…
- Alter: [2 1/2]^{AGE} Jahre…
- Die [16]^{AGE}jährige Patientin wurde von ihrer 7 jährigen Schwester bewusstlos aufgefunden.

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

4. Adressangaben [LOCATION]

Die Kategorie *LOCATION* umfasst alle Arten von Adressinformationen.

Diese Kategorie enthält Bezeichnungen von Einrichtungen, z.B. Standorte von Einrichtungen, spezifische (Teil-)Organisationen oder Einrichtungen sowie Organisationseinheiten.

Entitäten-Typ	Beschreibung	Beispiele
LOCATION_STREET	Straßenname und Hausnummer	*[Musterstraße 1]*$^{LOCATION\_STREET}$
LOCATION_CITY	Stadt, inkl Stadtteile, Gemeinde, Dörfer	*[Berlin]*$^{LOCATION\_CITY}$*, [Tegel]*$^{LOCATION\_CITY}$*, [Berlin-Tegel]*$^{LOCATION\_CITY}$*, [Wedding]*$^{LOCATION\_CITY}$*, [Markt Erlbach]*$^{LOCATION\_CITY}$
LOCATION_ZIP	Postleitzahl, Landeskennungen mit Postleitzahl → Annotation der gesamten Spanne einer PLZ	*[10115]*$^{LOCATION\_ZIP}$ *[Berlin]*$^{LOCATION\_CITY}$ *[A-1234]*$^{LOCATION\_ZIP}$ *[D-98765]*$^{LOCATION\_ZIP}$
LOCATION_COUNTRY	Nennung eines Landes	Urlaub in *[Frankreich]*$^{LOCATION\_COUNTRY}$
LOCATION_STATE	Nennung eines Staates oder Bundeslandes, Landkreis, (Regierungs-)Bezirk	Aufenthalt in *[Bayern]*$^{LOCATION\_STATE}$ *[Berchtesgadener Land]*$^{LOCATION\_STATE}$
LOCATION_HOSPITAL	Nennung einer Einrichtung mit klinischem Bezug und medizinischen Bezug in der Prozesskette der Behandlung eines Patienten *⇒ Bei Trennung durch Präpositionen wie 'in', allgemeinen Teil stehen lassen und den de-identifizierenden Teil so behandeln, wie er ist, z. B. hier Leipzig als City*	*[Universitätsklinikum Leipzig]*$^{LOCATION\_HOSPITAL}$*, Universitätsklinikum in [Leipzig]*$^{LOCATION\_CITY}$ *[Diakonissenkrankenhaus Berlin]*$^{LOCATION\_HOSPITAL}$*, [Diakonissenkrankenhaus Berlin]*$^{LOCATION\_HOSPITAL}$ *– Ophthalmologisches Zentrum,* *[ARCOS-KLINIK FLENSBURG Akademisches Lehrkrankenhaus der Otto-Waalkes-Universität Borkum]*$^{LOCATION\_HOSPITAL}$*,* *[ARCOS-KLINIK FLENSBURG Akademisches Lehrkrankenhaus der Otto-Waalkes-Universität]*$^{LOCATION\_HOSPITAL}$ *in [Borkum]*$^{LOCATION\_CITY}$ *[Paul-Langerhans-Station]*$^{LOCATION\_HOSPITAL}$ *[Praxis Dr. Mustermann]*$^{LOCATION\_HOSPITAL}$
LOCATION_ ORGANIZATION	Nennung einer Organisation ohne klinischen Bezug (u.a. Versicherungen)	*[BZH Heidelberg]*$^{LOCATION\_ORGANISATION}$ … arbeitet an *[Universität Leipzig]*$^{LOCATION\_ORGANISATION}$ Versicherung: *[AOK PLUS]*$^{LOCATION\_ORGANISATION}$
LOCATION_OTHER	Sonstige Adressen, eindeutig identifizierbare lokale Entitäten, Orte ohne klinische Funktionen	*[Rotes Haus]*$^{LOCATION\_OTHER}$ *(Gebäude auf dem Uniklinikgelände in Leipzig), [Villa Zeppelin]*$^{LOCATION\_OTHER}$*, [Rosensäle]*$^{LOCATION\_OTHER}$*, [Paulinum]*$^{LOCATION\_OTHER}$

Hinweise zur Annotation *LOCATION_HOSPITAL* und *LOCATION_ORGANIZATION*:

- Nennung für Station / Zimmer besteht aus Nummernbezeichnung, dann **keine** Annotation als *LOCATION_ORGANIZATION* oder *LOCATION_HOSPITAL* → "5. Kennung [ID]"
- Nennung für Station / Zimmer mit Namen von Personen oder Ortsangaben: *LOCATION_HOSPITAL*
- *Nennung von Praxen mit Arzt-Namen als LOCATION_HOSPITAL betrachten*
    - *[Orthopädische Praxis Dr. Mustermann]*LOCATION_HOSPITAL
- Allgemeine Beschreibungen von klinischen Einrichtungen und anderen 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"
    - „Neurologische Ambulanz"
    - „Abteilung für Innere Medizin"
    - „Neurologie"
    - „HNO-Klinik''
    - „Klinische Abteilung für Onkologie"
    - „Ophthalmologisches Zentrum"
    - „Institut für Pathologie"

- Es gibt spezifische Bezeichnungen von Kliniken, Zentren und Abteilungen, die durch die Kombination von mindestens zwei standardisierten medizinischen Fachabteilungen eine hohe Standortpräzision aufweisen. Aufgrund ihrer eindeutigen Identifizierbarkeit werden diese mit Location_Hospital annotiert:
    - Essen: *[Klinik für Hämatologie und Stammzelltransplantation]*Location_Hospital
    - Neuss: *[Zentrum für Zytologie, Pathologie und Molekularpathologie]*Location_Hospital
    - Berlin: *[Zentrum für außerklinische Beatmung und Sauerstofftherapie (CABS)]*LOCATION_HOSPITAL
- Scheinbar allgemeine Beschreibungen, die trotzdem Rückschluss auf eine Einrichtung geben, sollen annotiert werden, z.B. *[Charité Zentrum für außerklinische Beatmung und Sauerstofftherapie (CABS)]*LOCATION_HOSPITAL

Weitere Hinweise:

- Rotes Haus: Eindeutig identifizierbare Lokalität
- Marketing-Slogan *[Medizin ist unsere Berufung!]*, der z.B. das Uniklinikum Leipzig identifiziert, sollte als Teil von *LOCATION_HOSPITAL* annotiert werden.
- Zusätzliche Bezeichnungen von Einrichtungen, z.B. *[AöR]* und *[GmbH]*, sollen mit annotiert werden und Teil der Annotation der Einrichtung sein (z.B. *„Universitätsklinikum Leipzig AöR" → LOCATION_HOSPITAL*).
- Grenzfälle HOSPITAL vs. ORGANIZATION:
    - Spezifische Eigenbezeichnung einer Laboreinrichtung oder einer Pflegeeinrichtung: ⇒ als HOSPITAL
    - allgemeine Begriffe wie Labor, Zentrallabor oder Aufnahmelabor, die keine spezifische Benennung tragen (z.B. „*Labor"*, „*Zentrallabor"*, „*Aufnahmelabor", "Pflegeeinrichtung"*)
      ⇒ nicht annotieren

5. Kennung [ID]

Der Typ ID umfasst alle Nennungen von Identifikatoren in Gestalt einer Abfolge von Zeichen, einschließlich Nummern, Ziffern bzw. alphanumerische Kombinationen aus Zahlen und Buchstaben. Dazu gehören Patienten-IDs und Fall-IDs, ebenso IDs aus medizinischen Subsystemen, Versicherungsnummern und Kontonummern.

Entitäten-Typ	Beschreibung	Beispiele
ID	Patienten-IDs, Fall-IDs, Zimmer- und Stationsnummern (nur die Nummern), Sonstige Identifikationsnummern, IBAN-Nummern, Studienpseudonyme bzw. Namen von Studien im Rahmen einer Kohorte, in der Pat. teilnimmt (aber nicht als Empfehlung der Behandlungsanweisung), Akkreditierungsnummern	Pat.Nr.: *[312654356]^{ID}*, Fall.Nr.: *[71543356]^{ID}*, *Haus [5]^{ID},* *Postfach [1521]^{ID}* Zimmer-Nr. *[312]^{ID}, Station [123]^{ID},* *mehrere Aufenthalte auf der [PSY13]^{ID}, zuvor auf [KJPP-2]^{ID}* Block-Nr.: *[H213578-6]^{ID},* Konto: *[DE02120300000000202051]^{ID},* im Rahmen der Darmkrebsstudie nahm der Patient als *[S05NL-ZKA06]^{ID}* teil, Patient der *[LIFE]^{ID}*-Studie, *[3 Süd]^{ID}*

Hinweise

- *Haus [5]^{ID}, Postfach [1521]^{ID}*: keine lokale Einordnung (im Vergleich zu PLZ)
- Versionsnummern von Geräten (z.B. Ultraschall) sollen nicht annotiert werden
- Namen von Studien als Teil der Behandlung nicht annotieren: „UICC *2024”,* „NB*2004”*
- Abgekürzte Kennungen von Ärzten, Ärztinnen, etc. → *NAME_USER*

Hinweis ID und LOCATION

- Überprüfung, ob die Abkürzung generisch ist (z.B. “ZNA” = “Zentrale Notaufnahme” oder “Nephrologie” allein im Text)
  → Es soll **nichts annotiert** werden.
- Enthält die Abkürzung eine Identifikationsnummer oder -abkürzung, als ID annotieren
  - *„… verlegt von der [Süd 3] …” → ID*
  - *„… verlegt von der [Med 4] …” → ID → alles annotiert als ID*
    { Med 4 == “Medizinische Klinik 4” (Nephrologie in Erlangen) }
  - *„... Medizinische Klinik 4 …” → nur “4” als ID*

Hinweis ID und STUDIE

- „Der Patient ist in eine *[CLEARANCE]-Studie (Randomized Comparison of interventional closure of the left atrial appendage using a LAA closure device versus oral Anticoagulation)* eingeschlossen.”
  - Die Studie wird als Akronym angegeben und in Klammern wird das gesamte Acronym beschrieben.
  - Ausgeschriebene Studienbezeichnungen wie ID betrachten und annotieren (technische Hintergründe).

6. Kontaktdaten [CONTACT]

Der Typ CONTACT deckt Kontaktinformationen ab. Diese umfassen Telefonnummern, E-Mail-Adressen, Fax-Nummern, Internet-Adressen.

Entitäten-Typ	Beschreibung	Beispiele
CONTACT_PHONE	Telefonnummer (mit allen Bestandteilen Vorwahl und Durchwahl, auch Pager-Nummern)	Tel. *[02345-12345]*$^{CONTACT\_PHONE}$ Tel. *[+43 (453) 14-12345]*$^{CONTACT\_PHONE}$
CONTACT_EMAIL	E-Mail-Adresse	Mail: *[max.muster@beispiel.de]*$^{CONTACT\_EMAIL}$
CONTACT_FAX	Faxnummer	Fax. *[04324-65432]*$^{CONTACT\_FAX}$
CONTACT_URL	URL einer Homepage	Internet: *[www.beispiel.de]*$^{CONTACT\_URL}$

7. Beruf [PROFESSION]

Der Typ PROFESSION deckt Informationen zum Beruf eines Patienten bzw. einer Patientin ab. Das Ziel besteht darin, in einem Folgeschritt zu klären, ob dadurch mit der Berufsbezeichnung eine Information angegeben ist, die Rückschluss auf einen Patienten gibt.

Entitäten-Typ	Beschreibung	Beispiele
PROFESSION	Bezeichnungen von Berufen im Zshg. mit dem Patienten, ohne Nennung eines Status (z.B. arbeitslos, berentet, Teilzeit)	*[Verkäuferin]*^{PROFESSION}, *[Pfleger]*^{PROFESSION}, *[Bergbauarbeiter]*^{PROFESSION}, *[Monteur]*^{PROFESSION}, *[Frisörin]*^{PROFESSION} *[arbeitet im Büro]*^{PROFESSION}, *[Mitarbeiter im Schlachthof]*^{PROFESSION} *berenteter [Büroarbeiter];* *[Schwesternhelferinnenausbildung]*^{PROFESSION} *absolviert. Später hätte sie bei einer Lebensmittelkette gearbeitet.;* *[Krankenschwester]*^{PROFESSION} und *[arbeitet auf 450 Euro Basis in einem Hospiz]*^{PROFESSION}*;* *[Arbeit in der Gastronomie]*^{PROFESSION}*;* *da sie [in einem Bundesverband der Sicherheits und Verteidigungsindustrie arbeite]*^{PROFESSION}*;* *[arbeitet in Verwaltung]*^{PROFESSION} *in einer evangelischen Kirche;* *berufstätig (12 Jahre lang) in einer [Werkstatt]*^{PROFESSION}*, wobei sie Kleberdämpfen und Schleifstäuben ausgesetzt war;* *nach Ausbildung zum [medizinisch-technischen Angestellten]*^{PROFESSION} *berufstätig in [leitender Position in einem Labor]*^{PROFESSION}*;* *arbeite selbstständig als [Künstlerin]*^{PROFESSION} *und [Trainerin (Thaiboxen)]*^{PROFESSION}*;* *Sohn betreibe die Nachfolgefirma, der von dem Patienten über die letzten 4 Jahre abgewickelten [Firma für Medizintechnik (Vertrieb und Service)]*^{PROFESSION}*;*

Hinweis:

- Annotation der Berufsbezeichnung soll so einfach wie möglich gehalten sein. Im Rahmen eines Kurationsschritts soll entschieden werden, ob der Beruf mit einem Surrogat ersetzt wird oder ob im Zweifel (bei sehr seltenen Berufen, die einfach Re-Identifizierung erlauben) der gesamte Text aus dem GeMTeX-Korpus herausgenommen wird.
- „studiert *[Pharmazie]*": als Beruf annotieren → „studiert *[Pharmazie]*^{PROFESSION}"
- Status im Sinne der Ausbildung wie "Schüler", „Student" soll nicht annotiert werden.
- Alles, was einen Hinweis zu einem Beruf gibt, soll annotiert werden. Ist eine Aussage der Beschreibung für eine Berufsausübung nur über eine Einrichtung ersichtlich, so ist entweder nur die (rückverfolgbare und identifizierbare) Einrichtung als *LOCATION_OTHER* zu annotieren oder die (möglichst allgemein gehaltene Einrichtung als Teil der Beschreibung mit der Tätigkeitsausübung zu annotieren.
- „Inv-Renter" (Invaliden-Rentner) nicht annotieren.
- Sollte die Aussage der Beschreibung für eine Berufsausübung nur über eine Einrichtung bestehen, so ist die Einrichtung mit den entsprechenden Kategorien zu annotieren
- Berufe von Angehörigen sollen nach Möglichkeit nicht annotiert werden.
    - Sollte solch eine Nennung über eine Einrichtung genannt sein, ist die Einrichtung als entsprechende Kategorie zu annotieren.
    - Sollte solch eine Formulierung genannt sein und Rückschluss auf eine Re-Identifizierung geben, bitte OTHER verwenden,

8. Other [OTHER]

Mit dem Typ OTHER werden alle schützenswerten Passagen annotiert, für die keine der obigen Typ-Definitionen zutrifft, die aber eine Person trotzdem eindeutig identifizieren können. Diese Kategorie sollte nur dann genutzt werden, wenn nach einer Prüfung aller anderen Kategorien keine Zuordnung möglich ist.

- Ein Beispiel dafür ist die Angabe zu einem Verwandtschaftsverhältnis einer Person öffentlichen Lebens (z.B. *„Tochter der bayerischen Gesundheitsministerin"*).
- Ein weiteres Beispiel sind einzelne Individuen referenzierende Nominalphrasen (z.B. *„der letzte Alpwirt im Stubaital"*).

5. Weitere Hinweise zur manuellen Annotation

5.1. Was bei der Annotation *beachtet* werden soll

- Im Zweifelsfall immer annotieren! Eine falsche Annotation ist wertvoller als eine übersehene Annotation.
- Behandlung von **Fehlern (Rechtschreibung, Tippfehler)**: so damit umgehen, als wäre es korrekt geschrieben.
- **Abkürzungen** so behandeln, als sei die Abkürzung ausgeschrieben, z.B. abgekürzte Namen („M. Muster") oder Punkt von Namenstiteln („Dr.").
- Es soll nur das Konzept verfolgt werden, aber nicht interpretiert werden. In vielen Fällen ist die Bedeutung eines PII offensichtlich. Beispiel: Eine Reihung von Zahlen und Buchstaben lässt eindeutig aus dem Kontext erkennen, dass der behandelnde Arzt gemeint ist. Hier soll trotzdem die Abkürzung nur als *[ID]* markiert werden und nicht [Name_Doctor] interpretiert werden.

5.2. Was ***nicht* annotiert** 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 Anfang oder Ende eines potentiellen Annotats).
    - **Ausnahme**: Namensbestandteile Teil des Annotats (z.B. *[von der Leyen]*)

- Kategorisierungen, Attribute und nähere Beschreibungen eines Individuums oder einer Ortsangabe (z.B. *„geboren am"*, *„wohnhaft in"*)

- Platzhalter im Text, die für etwas Identifizierendes gestanden hätten, wenn es ausgeschrieben worden wäre, z.B. :
    - *„ ... vom xxxx bis ERSETZEN …"*
    - *„ vom [17.07.2024]^{DATE} bis xxxx …"*
    - *[17]^{DATE}.XX.XXXX*
    - *__.__.[2023]^{DATE}*

- **Hinweise auf Covid-Standardprotokolle zu Testverfahren:**
    - *„Beurteilung Coronavirus SARS-CoV-2-RNA. In der Probe wurde mittels real-time PCR keine RNA des neuen Coronavirus Sars-CoV-2 nachgewiesen. Diese Untersuchung basiert auf dem im Januar 2020 veröffentlichten Protokoll des Konsortialabors für Coronaviren am Institut für Virologie der Charite Berlin.*

- **Referenzen auf Publikationen und Zugehöriger Arbeitsgruppe bzw. Arbeitsgruppenstandort**
    - *„Diese Entität wurde erst kürzlich von der Arbeitsgruppe um Frau Professor Müller vom Max-Müller-Forschungszentrum, New York, USA beschrieben (Müller, Agathe et al. Cancer is curable. Nature 2024)."*

- **Konsile, Boards** (keine festen medizinische Einheiten, die identifizierende Namen haben könnten)

- **Eponyme (Eigennamen) von Skalen, Diagnosen, Prozeduren, etc.** sollen nicht beachtet werden (können aber fälschlicherweise automatisiert markiert sein) → Bsp: *“Von-Willebrand-Faktor”, “Operation nach Hartmann”, “Wernicke-Enzephalopathie”, “Morbus Sudeck”, “Stadium nach Rutherford”, “Faktor V Leiden”, “Barthel-Score”*

- **Skaleneinstufungen** wie bei „bifrontalen Kopfschmerz (NRS 10/10)”, NRS ist hierbei eine numerische Rating-Skala mit welcher der Patient subjektiv empfundenen Schmerz einordnen kann

- **Studienprotokoll-Bezeichnungen** inkl. Jahreszahl (z.B. „NB2004”, „UICC 2004”)

- **Familienstand** (ledig, verheiratet, geschieden, verwitwet, Vater von 3 Kindern)

- **Religionszugehörigkeiten** (z.B. Zeuge Jehova, Muslim)

- **ICD-10-Codes, ISO-Normen** und andere Nennungen von medizinischen Terminologie-Systemen, wie Tumor- bzw. oder **TNM**-Klassifikationen

5.3. Beispiele für Titel-Beschreibungen

- Deutschland
    - Doktor der Medizin (Dr. med.)
    - Doktor der Zahnmedizin (Dr. med. dent.)
    - Doktor der Medizinischen Wissenschaften (Dr. rer. medic., Dr. rer. med., Dr. med. sci., Dr. rer. hum.)
    - Doktor der Humanwissenschaften (Dr. sc. hum.)
    - Doktor der naturwissenschaftlichen Medizin (Dr. nat. med.)
    - Doktor mit Lehrberechtigung (Dr. med. habil.)
    - Doctor honoris causa (Dr. h.c., Dr. E.h.)
    - MBA
    - M.A. (Master of Arts/Magister Artium)
- Historische Titel:
    - Diplommediziner (Dipl.-Med.) in der DDR
    - Diplom-Stomatologe (Dipl.-Stom.) in der DDR
    - Diplommedizinpädagoge (Dipl.-Med. Päd.) in der DDR
    - Doctor scientiae medicinae (Dr. sc. med.) in der DDR
- Inoffizielle Titel:
    - Studiosus medicinae (stud. med.)
    - Candidatus medicinae (cand. med.)
    - Candidatus medicinae dentariae (cand. med. dent.)
- Österreich
    - Bachelor of Science in Medical Sciences (BScMed)
    - Doktor der gesamten Heilkunde (Dr. med. univ.)
    - Doktor der Zahnmedizin (Dr. med. dent.)
    - Doktor der medizinischen Wissenschaft (Dr. scient. med.)
- Schweiz
    - Bachelor of Medicine (BMed)
    - Master of Medicine (MMed)
    - Doktor der Medizin (Dr. med.)
    - Bachelor of Dental Medicine (B Dent Med)
    - Master of Dental Medicine (M Dent Med)
    - Doktor der Zahnmedizin (Dr. med. dent.)
    - Master of Chiropractic Medicine (M Chiro Med)
    - Doktor der Chiropraktischen Medizin (Dr. med. chiro.)
    - Doktor der Medizinischen Wissenschaft (Dr. sc. med.)
    - Bachelor of Science in Nursing (BSN, BScN)
    - Master of Science in Nursing (MSN, MScN)
- Tschechien und Slowakei
    - Medicinae universae doctor (MUDr.)
    - Medicinae dentalis doctor (MDDr.)
    - Candidatus scientiarum (CSc)
    - Doctor scientiarum (DrSc, DSc)
- allgemein
    - PhD / Ph.D.

Literaturverzeichnis

Drepper, Johannes (2015): Leitfaden zum Datenschutz in medizinischen Forschungsprojekten. Generische Lösungen der TMF 2.0. Erscheinungsort nicht ermittelbar: MWV Medizinisch Wissenschaftliche Verlagsgesellschaft.

Kolditz, Tobias; Lohr, Christina; Hellrich, Johannes; Modersohn, Luise; Betz, Boris; Kiehntopf, Michael; Hahn, Udo (2019): Annotating German Clinical Documents for De-Identification. In: *Studies in health technology and informatics* 264, S. 203–207. DOI: 10.3233/SHTI190212.

Lohr, Christina; Faller, Jakob; Riedel, Andrea; Nguyen, Hung Manh; Wolfien, Markus; Hofenbitzer, Justin et al. (2025): GeMTeX's De-Identification in Action: Lessons Learned & Devil's Details. In: *Studies in health technology and informatics* 331, S. 274–282. DOI: 10.3233/SHTI251406.

Lohr, Christina; Matthies, Franz; Faller, Jakob; Modersohn, Luise; Riedel, Andrea; Hahn, Udo et al. (2024a): De-Identifying GRASCCO - A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus: IOS Press (Studies in Health Technology and Informatics, 317). In: *Studies in health technology and informatics*, S. 171–179.

Lohr, Christina; Matthies, Franz; Jakob, Faller; Modersohn, Luise; Riedel, Andrea; Hahn, Udo et al. (2024b): GraSCCo_PHI - Graz Synthetic Clinical text Corpus with Protected Health Information Annotations.

Meineke, Frank; Modersohn, Luise; Loeffler, Markus; Boeker, Martin (2023): Announcement of the German Medical Text Corpus Project (GeMTeX). In: *Studies in health technology and informatics* 302, S. 835–836. DOI: 10.3233/SHTI230283.

Modersohn, Luise; Schulz, Stefan; Lohr, Christina; Hahn, Udo (2022): GRASCCO - The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. In: *Studies in health technology and informatics* 296, S. 66–72. DOI: 10.3233/SHTI220805.

Richter-Pechanski, Phillip; Amr, Ali; Katus, Hugo A.; Dieterich, Christoph (2019): Deep Learning Approaches Outperform Conventional Strategies in De-Identification of German Medical Reports. In: *Studies in health technology and informatics* 267, S. 101–109. DOI: 10.3233/SHTI190813.

Richter-Pechanski, Phillip; Riezler, Stefan; Dieterich, Christoph (2018): De-Identification of German Medical Admission Notes. In: *Studies in health technology and informatics* 253, S. 165–169.

Tobias Kolditz; Christina Lohr; Luise Modersohn; Udo Hahn (2023): Annotationsleitlinien für deutschsprachige Medizintexte - Teil 2: Annotation von personenidentifizierenden PHI-Attributen.