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

Christina LOHR[1], Franz MATTHIES[1], Jakob FALLER[2], Luise MODERSOHN[3],
Andrea RIEDEL[2], Udo HAHN[1], Rebekka KISER[3], Martin BOEKER[3] and Frank MEINEKE[1]

1 Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University

2 Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg

3 Institute of Artificial Intelligence and Informatics in Medicine, Medical Center rechts der Isar, Technical University Munich, Germany

09/09/2024 – Dresden

GESUNDHEIT GEMEINSAM

SPONSORED BY THE Federal Ministry of Education and Research

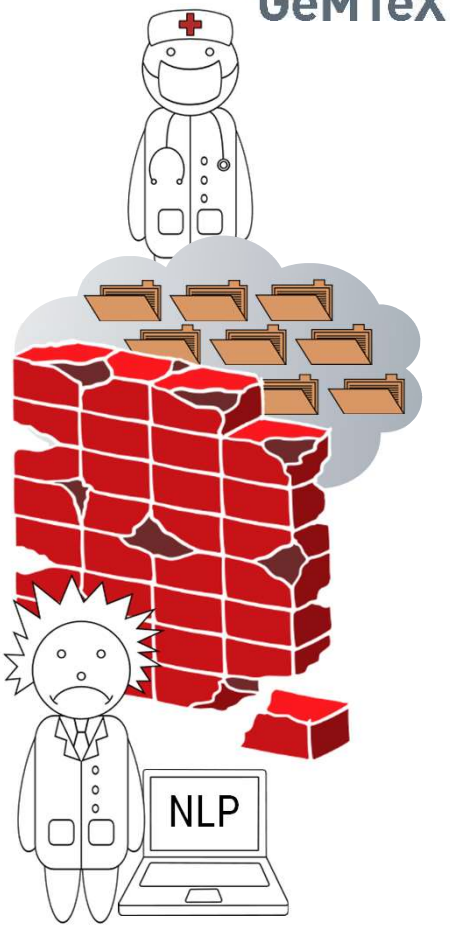
Barriers for Clinical Natural Language Processing

Legal protection

- processing of health data (in general) only possible on the basis of a consent or a specific legal basis (Art. 6, 9 GDPR)
- protection of identities and privacy-sensitive information

Current status

- up until now no distributable annotations
- no free downloadable German language corpus with real clinical text without distortion (BRONCO, Kittner et al. 2021) or placeholders (CARDIO:DE, Richter-Pechanski et al. 2023) for privacy-sensitive text items
- 3000PA text corpus and DeId annotation non-publishable (Kolditz et al. 2019)
- MEDBERT.DE (Bressem et al. 2023): language model downloadable
- GERNERMED (Kramer & Frei 2022): based on translated English language corpora
- GRASCCO (Modersohn et al. 2022): free downloadable synthetic discharge summaries, no distributable annotations

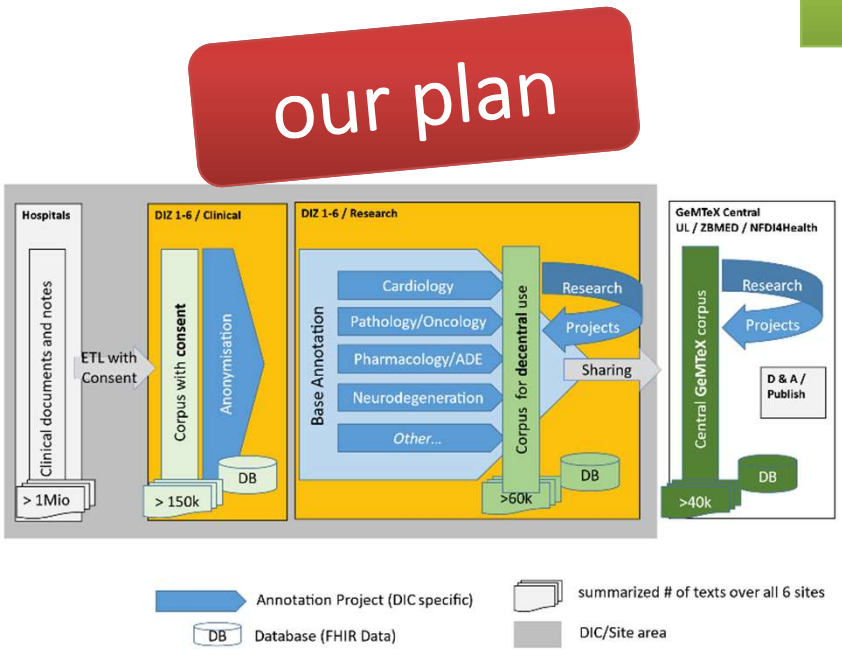


GeMTeX: **Ger**man **M**edical **Tex**t Corpus

- large infrastructure effort targeting German clinical language
  - part of the German Medical Informatic Initiative (MII)
  - 17 partners, including leading companies
  - Data integration centers of 6 University Hospitals
  - All patients consented – MII Broad Consent
  - 6 * 10.000 clinical text documents

- main goal: publicly available clinical text corpus for research
  - without privacy-sensitive information
  - semantic annotations
  - meta-data structure

Why GRASSCO?



Pilot study to answer

1. Feasibility…?
   - annotation tools
   - annotation workflow
   - management structures for local annotation teams
   - cross-hospital annotation requirements
   - (shared) annotation guidelines
   - appropriate ?
2. Competitive inter-annotator agreements scores?

GRASCCO

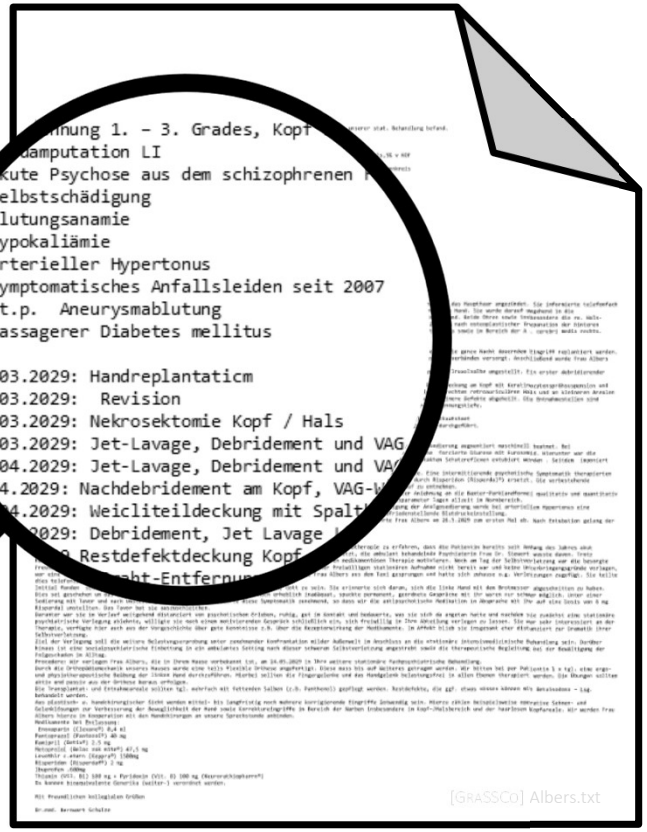
Graz Synthetic Clinical Text Corpus

63	synthetic discharge summaries
5,430	sentences
43.667	tokens

licence CC0 1.0 Universal („No copyright“)

download https://doi.org/10.5281/zenodo.6539131

more details Luise Modersohn, et al. „GRASCCO - The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus“. Stud Health Technol Inform. GMDS 2022, 2022 Aug 17;296:66-72.



Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE
4. LOCATION
5. IDE
6. CONTACT
7. PROFESSION
8. OTHER

PHI category system adapted from US law (HIPAA)

[NAME_PATIENT]
Wir berichten über lhre 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]
We report on your patient Beate Albers
[DATE]
(* 1997/04/04) who underwent inpatient treatment
[DATE] [DATE]
03/19 to 2029/05/07.

Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE
4. LOCATION
5. IDENTIFIERS
6. CONTACT
7. PROFESSION
8. OTHER

- Patient
- Relative
- Doctor
- External
- Name titles
- User names

Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE
4. LOCATION
5. IDENTIFIERS
6. CONTACT
7. PROFESSION
8. OTHER

Summary of all date information, e.g.
- Birth date
- Admission date

Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE — Is the person over 89 years old?
4. LOCATION
5. IDENTIFIERS
6. CONTACT
7. PROFESSION
8. OTHER

Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE
4. LOCATION
5. IDENTIFIERS
6. CONTACT
7. PROFESSION
8. OTHER

- Street
- City
- ZIP
- Country
- Hospital
- Organisation
- Other

Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE
4. LOCATION
5. IDENTIFIERS
6. CONTACT
7. PROFESSION
8. OTHER

Summary of all identifiers, e.g.:

- Insurance numbers
- Medical record numbers
- Account numbers

Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE
4. LOCATION
5. IDENTIFIERS
6. CONTACT
7. PROFESSION
8. OTHER

- Phone
- Email
- Fax
- URL

Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE
4. LOCATION
5. IDENTIFIERS
6. CONTACT
7. PROFESSION
8. OTHER

Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE
4. LOCATION
5. IDENTIFIERS
6. CONTACT
7. PROFESSION
8. OTHER

Protected Health Information (PHI) in GeMTeX

1. Person NAMES
2. DATE
3. AGE
4. LOCATION
5. IDENTIFIERS
6. CONTACT
7. PROFESSION
8. OTHER

deleted from US PHI/HIPAA list

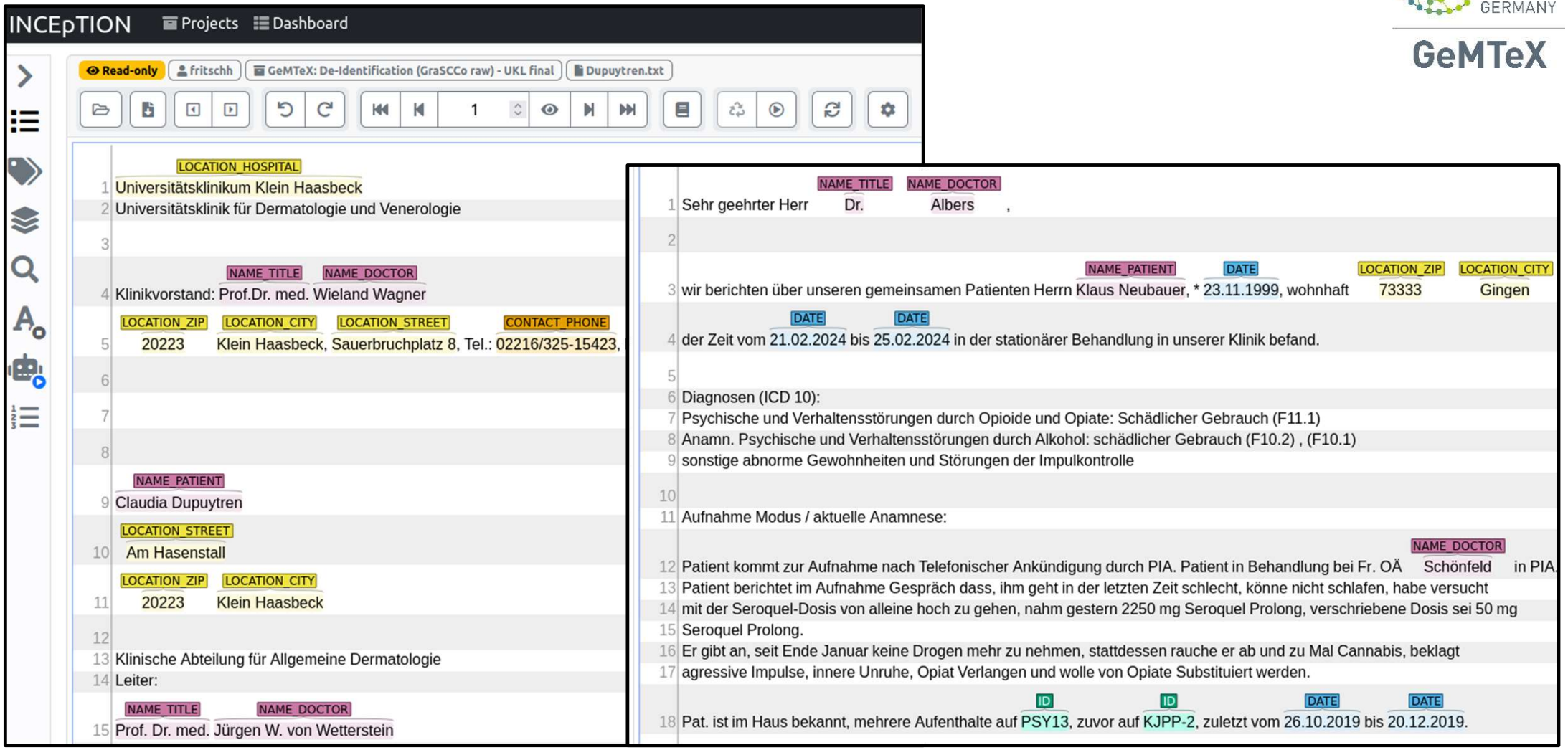
- Biometric Identifiers
- Images

Pilot Study – Organizational Setup

- 2 teams in Leipzig and Erlangen
  - 6 annotators
    - students of medicine
    - at least third year / first exam
  - 2 curators
- supervised by 1 NLP researcher
- supported by GeMTeX project coordinators

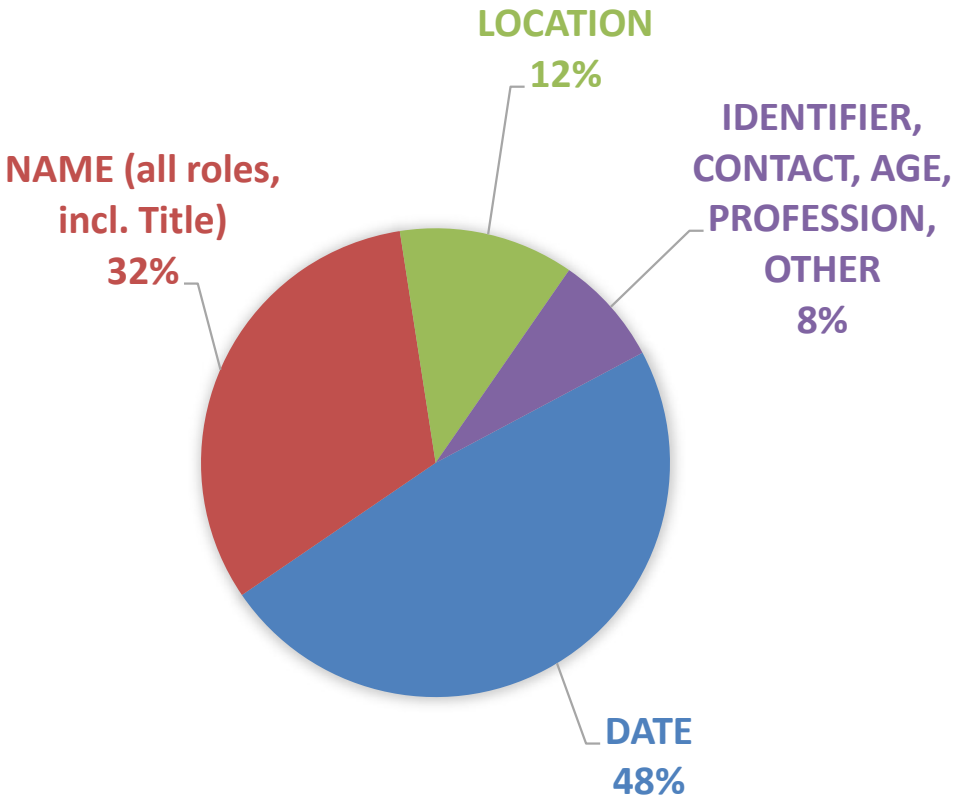
- pre-annotation: averbis text analytics; HEALTH discovery
- annotation tool: INCEpTION

Pilot Study – Example Annotation with INCEpTION



Results: Final Corpus and Annotations

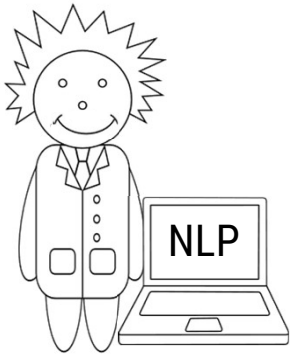
- inter annotator agreement
  - Krippendorff's $\alpha \approx 0.97$
- curated corpus
  - 1438 PHI annotations
  - ≈ 3 % of tokens
- curated files
  - provided as UIMA XMI



https://doi.org/10.5281/zenodo.11502329

Conclusion

- proof of concept of a large-scaled annotation campaign
- checked and adapted
  - workflows
  - tools & settings
  - inter annotator agreement ($\alpha \approx 0.97$) comparable with English-language corpora (i2b2 & n2c2)



- release
  - annotations: first publicly available clinical text corpus with de-identification annotations for German language without restrictions
  - annotation guideline
  - annotation metadata

https://doi.org/10.5281/zenodo.11502329

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

Christina LOHR, Franz MATTHIES, Jakob FALLER, Luise MODERSOHN,
Andrea RIEDEL, Udo HAHN, Rebekka KISER, Martin BOEKER and Frank MEINEKE

» christina.lohr@imise.uni-leipzig.de
» https://www.smith.care/de/gemtex_mii

Thanks to
- all annotators & curators
- DIZ & GeMTeX staff
- INCEpTION team and Averbis

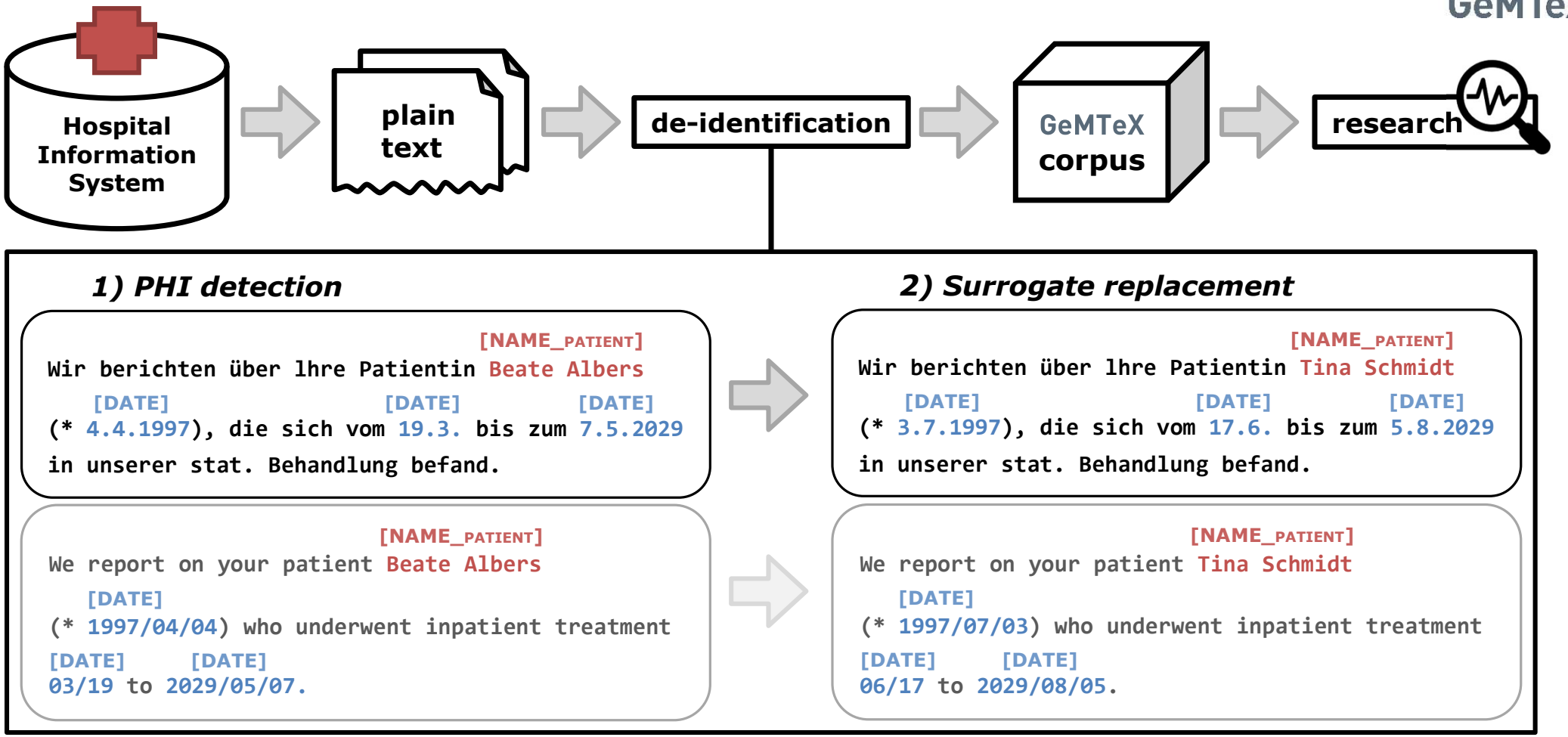
SPONSORED BY THE

Federal Ministry of Education and Research

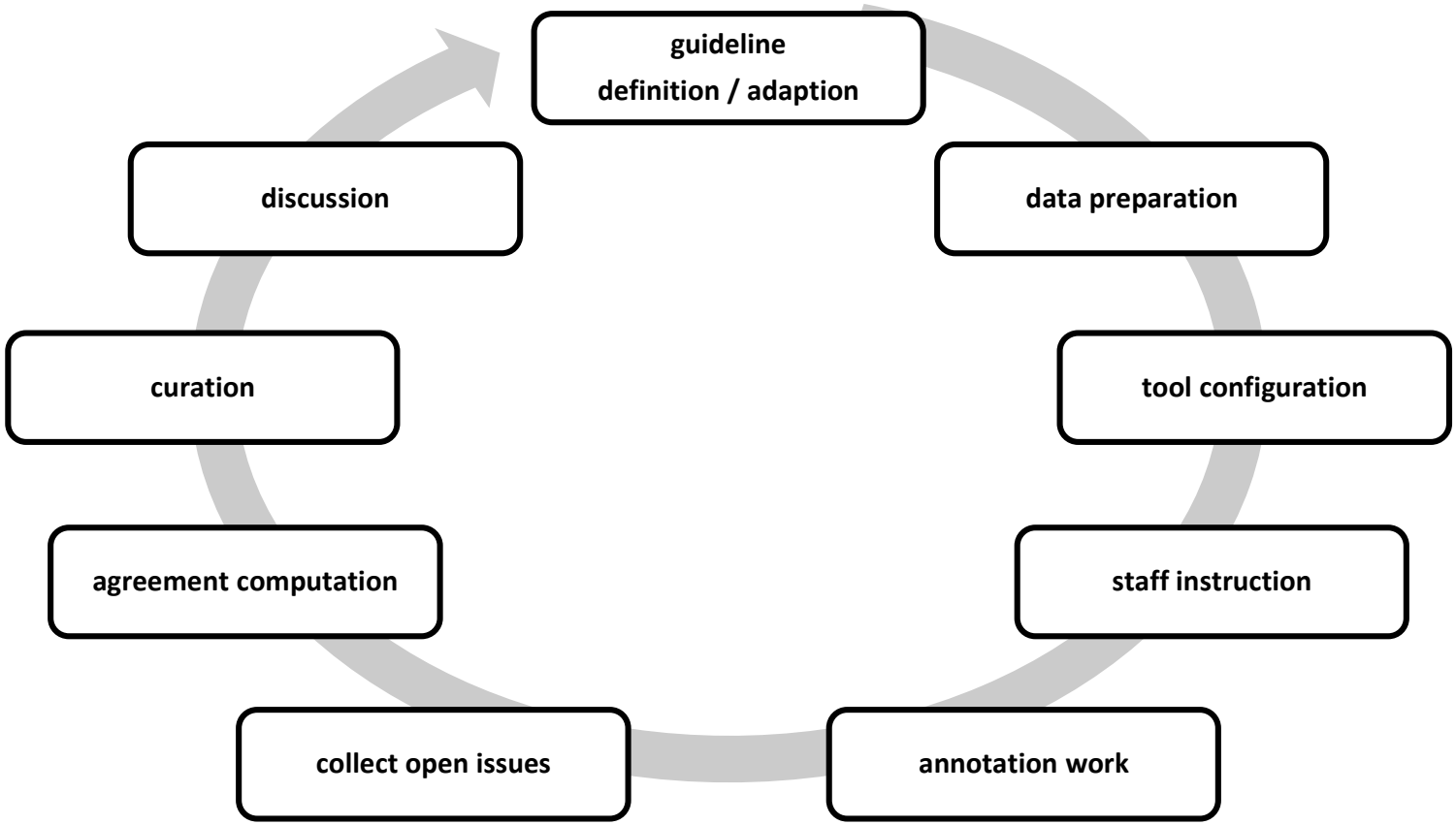
This work was supported by BMBF within the GeMTeX project under grant 01ZZ2314B (CL, FMei, FMat, UH) and 01ZZ2314A (MB, LM) as well as 01ZZ2314G (JF) and NUM 2.0 under grant 01KX2121 (AR).

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%

GeMTeX – De-Identification



Pilot Study – Process



Krippendorff‘s α

- Measure of the observed disagreement taking into account the number of categories and a random disagreement

- K. Krippendorff, 2013, Content Analysis: An Introduction to Its Methodology, 3rd ed. Thousand Oaks, CA, USA: Sage, PP. 221–250

Discussion

- PHI category system adapted from US law (HIPAA)

- rocessed data based on Broad Consent of MII, clarification included
- precise re-identification of a person possible in some rare cases
- deleting of document with to high potential of re-identification

- GDPR does not require that identification of a person is completely ruled out
- recital 26 of the GDPR – 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

Pilot Study – Process

Leipzig

- **first run:** preliminary version of annotation guide (without AGE, PROFESSION, NAME TITLE and simplified name roles) on full GRASCCO
- updated annotation guideline
- **second run** on full GRASCCO
- Agreement: **(1) 0.97 / (2) 0.97**

Erlangen

- used updated guideline
- annotation campaign with **two iterations** on GRASCCO (30 / 63 docs)
- Agreement: **(1) 0.95 / (2) 0.97**