

# De-Identifying GRASCCo

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

Christina LOHR<sup>1</sup>, Franz MATTHIES<sup>1</sup>, Jakob FALLER<sup>2</sup>, Luise MODERSOHN<sup>3</sup>,  
Andrea RIEDEL<sup>2</sup>, Udo HAHN<sup>1</sup>, Rebekka KISER<sup>3</sup>, Martin BOEKER<sup>3</sup> and Frank MEINEKE<sup>1</sup>

1

**imise.**



UNIVERSITÄT  
LEIPZIG  
Medizinische Fakultät

Institute for Medical Informatics, Statistics and Epidemiology,  
Leipzig University

2

Universitätsklinikum  
Erlangen



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

3

**TUM** **ARI**  
Klinikum rechts der Isar  
Technische Universität München

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

09/09/2024 – Dresden

GESUNDHEIT **GEMEINSAM**

**gmds**  
Deutsche Gesellschaft für  
Medizinische Informatik,  
Biometrie und  
Epidemiologie e.V.

SPONSORED BY THE



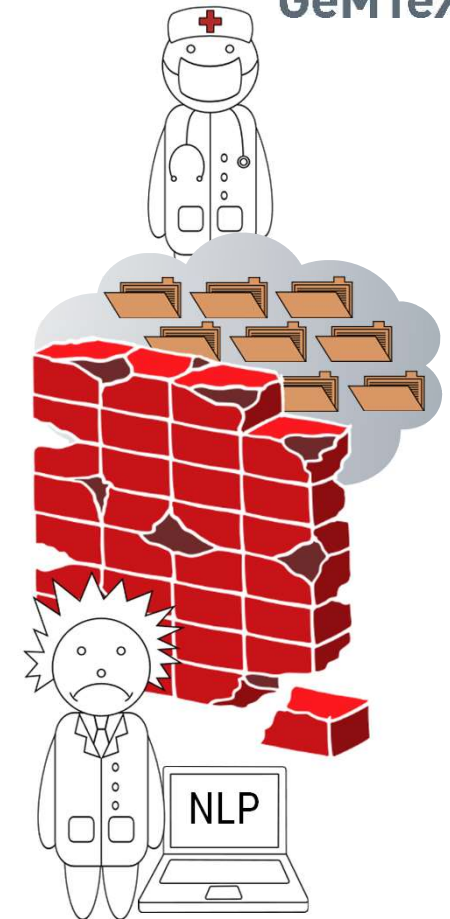
# Barriers for Clinical Natural Language Processing

## 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 Deld 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



© Christina Lohr

# GeMTeX: **German Medical Text 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?

our plan

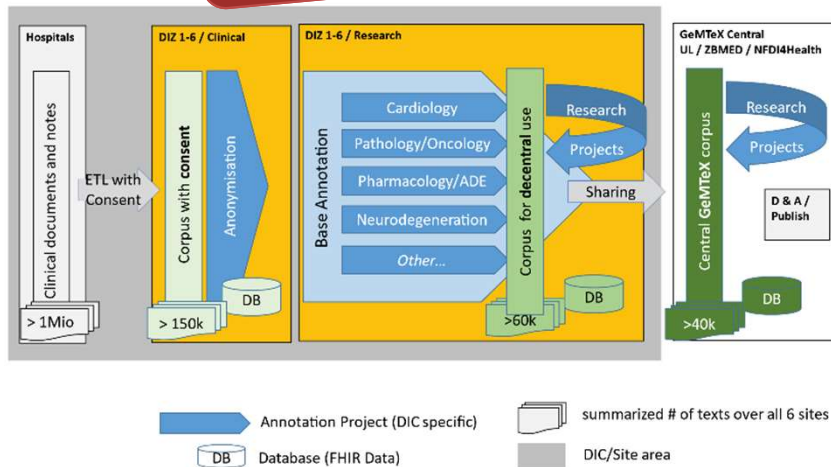


Pilot study to answer

## 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  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 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]  
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 | <ul style="list-style-type: none"><li>• Patient</li><li>• Relative</li><li>• Doctor</li><li>• External</li><li>• Name titles</li><li>• User names</li></ul> |
| 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

Summary of all date information, e.g.

3. AGE

- Birth date
- Admission date

4. LOCATION

5. IDENTIFIERS

6. CONTACT

7. PROFESSION

8. OTHER



# 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

- Phone
- Email
- Fax
- URL

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

# 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



- annotation tool:

INCEpTION





# Pilot Study – Example Annotation with INCEpTION

INCEpTION Projects Dashboard

Read-only fritschh GeMTeX: De-Identification (GraSCCo raw) - UKL final Dupuytren.txt

1 **LOCATION\_HOSPITAL**  
Universitätsklinikum Klein Haasbeck

2 Universitätsklinik für Dermatologie und Venerologie

3

4 **NAME\_TITLE** **NAME\_DOCTOR**  
Klinikvorstand: Prof.Dr. med. Wieland Wagner

5 **LOCATION\_ZIP** **LOCATION\_CITY** **LOCATION\_STREET** **CONTACT\_PHONE**  
20223 Klein Haasbeck, Sauerbruchplatz 8, Tel.: 02216/325-15423,

6

7

8

9 **NAME\_PATIENT**  
Claudia Dupuytren

10 **LOCATION\_STREET**  
Am Hasenstall

11 **LOCATION\_ZIP** **LOCATION\_CITY**  
20223 Klein Haasbeck

12

13 Klinische Abteilung für Allgemeine Dermatologie

14 Leiter:

15 **NAME\_TITLE** **NAME\_DOCTOR**  
Prof. Dr. med. Jürgen W. von Wetterstein

1 **NAME\_TITLE** **NAME\_DOCTOR**  
Sehr geehrter Herr Dr. Albers ,

2

3 **NAME\_PATIENT** **DATE** **LOCATION\_ZIP** **LOCATION\_CITY**  
wir berichten über unseren gemeinsamen Patienten Herrn Klaus Neubauer, \* 23.11.1999, wohnhaft 73333 Gingen

4 **DATE** **DATE**  
der Zeit vom 21.02.2024 bis 25.02.2024 in der stationärer Behandlung in unserer Klinik befand.

5

6 Diagnosen (ICD 10):

7 Psychische und Verhaltensstörungen durch Opioide und Opiate: Schädlicher Gebrauch (F11.1)

8 Anamn. Psychische und Verhaltensstörungen durch Alkohol: schädlicher Gebrauch (F10.2) , (F10.1)

9 sonstige abnorme Gewohnheiten und Störungen der Impulskontrolle

10

11 Aufnahme Modus / aktuelle Anamnese:

12 **NAME\_DOCTOR**  
Patient kommt zur Aufnahme nach Telefonischer Ankündigung durch PIA. Patient in Behandlung bei Fr. OÄ Schönfeld in PIA.

13 Patient berichtet im Aufnahme Gespräch dass, ihm geht in der letzten Zeit schlecht, könne nicht schlafen, habe versucht

14 mit der Seroquel-Dosis von alleine hoch zu gehen, nahm gestern 2250 mg Seroquel Prolong, verschriebene Dosis sei 50 mg

15 Seroquel Prolong.

16 Er gibt an, seit Ende Januar keine Drogen mehr zu nehmen, stattdessen rauche er ab und zu Mal Cannabis, beklagt

17 aggressive Impulse, innere Unruhe, Opiat Verlangen und wolle von Opiate Substituiert werden.

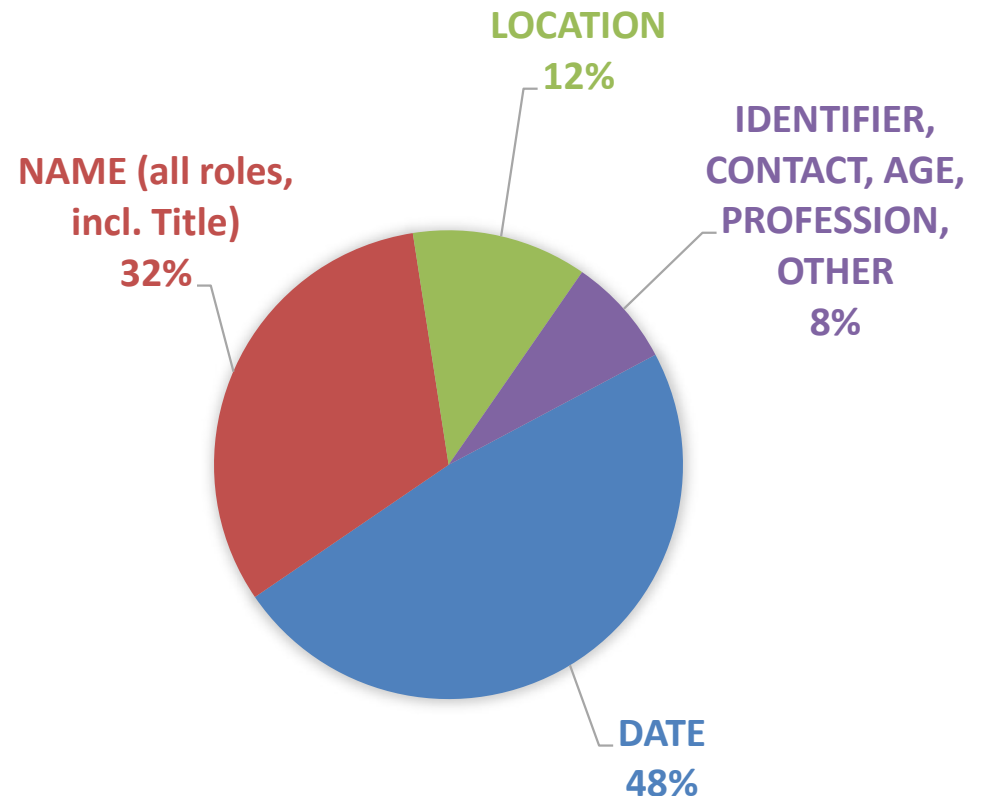
18 **ID** **ID** **DATE** **DATE**  
Pat. ist im Haus bekannt, mehrere Aufenthalte auf PSY13, zuvor auf KJPP-2, zuletzt vom 26.10.2019 bis 20.12.2019.

## Results: Final Corpus and Annotations

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

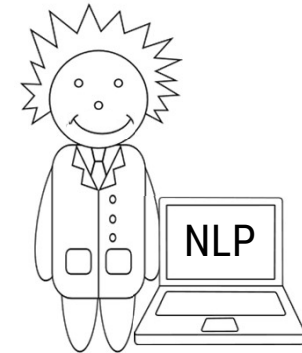


<https://doi.org/10.5281/zenodo.11502329>



# 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

---

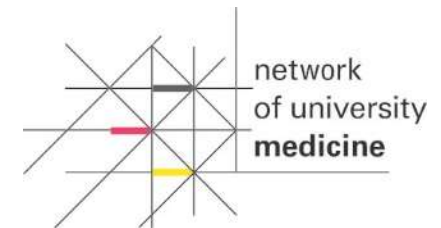
Christina LOHR, Franz MATTHIES, Jakob FALLER, Luise MODERSOHN,  
Andrea RIEDEL, Udo HAHN, Rebekka KISER, Martin BOEKER and Frank MEINEKE

» [christina.lohr@imise.uni-leipzig.de](mailto:christina.lohr@imise.uni-leipzig.de)  
» [https://www.smith.care/de/gemtex\\_mii](https://www.smith.care/de/gemtex_mii)

---



## GeMTeX



SPONSORED BY THE



Federal Ministry  
of Education  
and Research

**Thanks to**

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

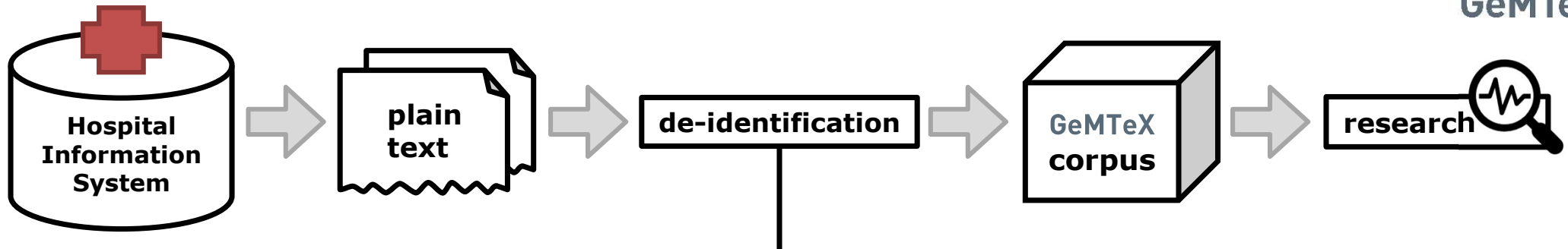
This work was supported by BMBF within the GeMTeX project under grant 01ZZ2314B (CL, FMei, FMat, UH) and 01ZZ2314A (MB, LM) as well as 01ZZ2314G (JF) and NUM 2.0 under grant 01KX2121 (AR).

Christina Lohr // 09/09/2024 // Dresden // De-Identifying GRASCCO – A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus

GESUNDHEIT **GEMEINSAM**

PHI Category	count	$\mu$	$\sigma$	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



## 1) PHI detection

Wir berichten über Ihre Patientin **[NAME\_PATIENT] Beate Albers**  
 (\* **[DATE] 4.4.1997**), die sich vom **[DATE] 19.3.** bis zum **[DATE] 7.5.2029**  
 in unserer stat. Behandlung befand.

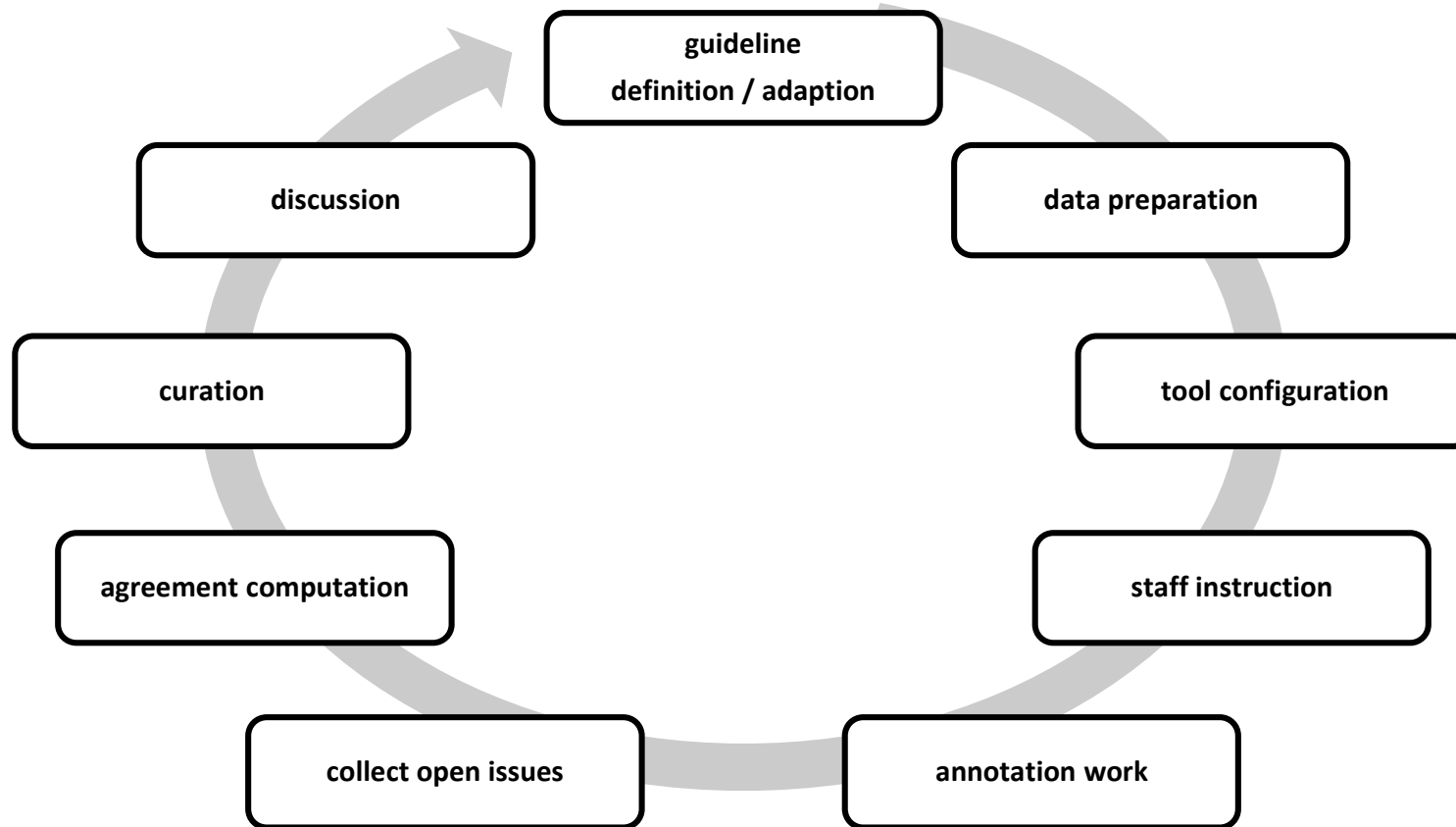
We report on your patient **[NAME\_PATIENT] Beate Albers**  
 (\* **[DATE] 1997/04/04**) who underwent inpatient treatment  
**[DATE] 03/19** to **[DATE] 2029/05/07**.

## 2) Surrogate replacement

Wir berichten über Ihre Patientin **[NAME\_PATIENT] Tina Schmidt**  
 (\* **[DATE] 3.7.1997**), die sich vom **[DATE] 17.6.** bis zum **[DATE] 5.8.2029**  
 in unserer stat. Behandlung befand.

We report on your patient **[NAME\_PATIENT] Tina Schmidt**  
 (\* **[DATE] 1997/07/03**) who underwent inpatient treatment  
**[DATE] 06/17** to **[DATE] 2029/08/05**.

# Pilot Study – Process



## Krippendorff's $\alpha$

- 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)
- processed data based on Broad Consent of MII, clarification included
- precise re-identification of a person possible in some rare cases
- deleting of document with too 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

