Pseudonymization of PHI Items in German Clinical Reports

[1] **Christina LOHR**, [2] Elisabeth EDER and [1] Udo HAHN

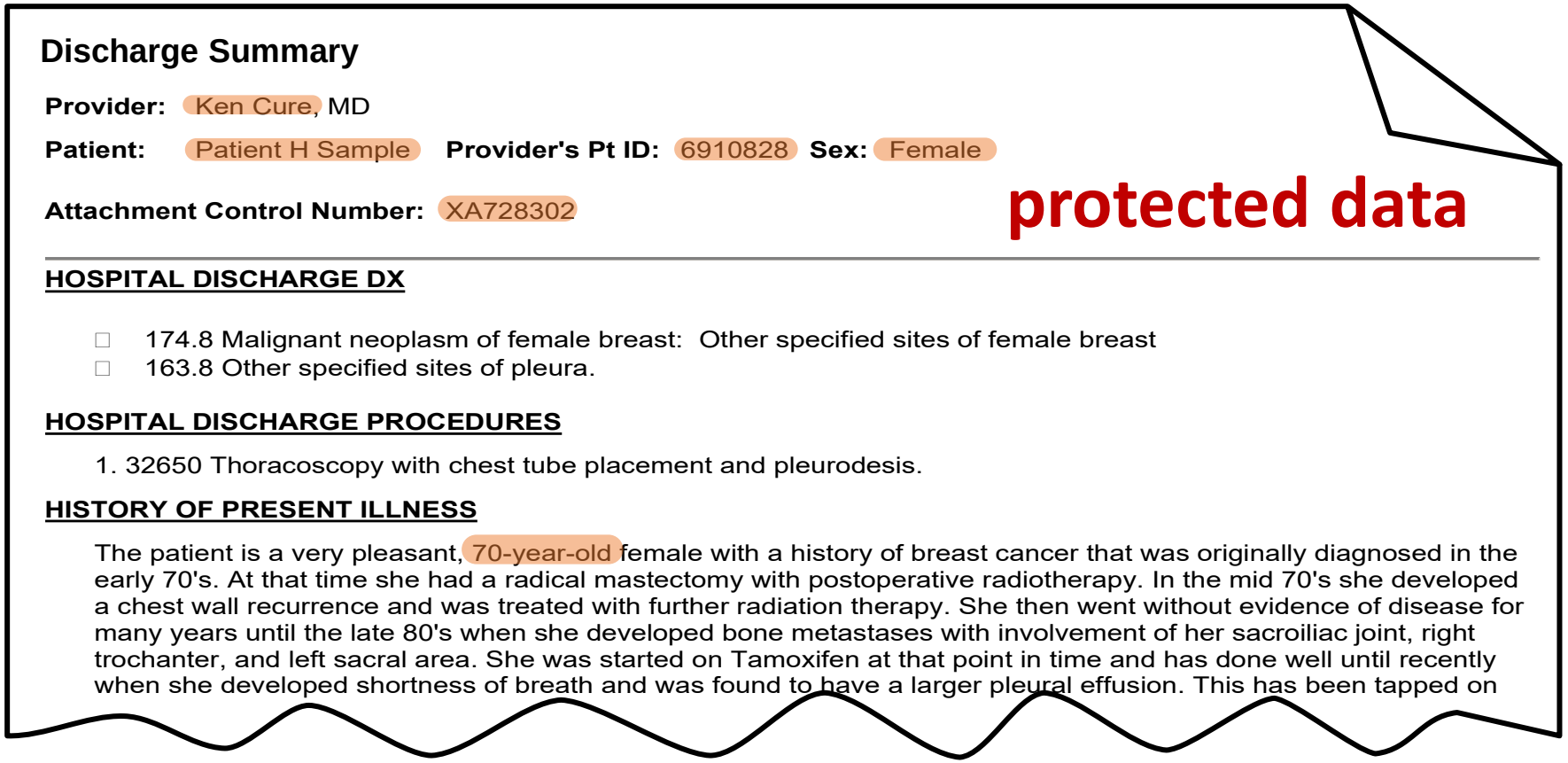
1 FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

2 UNIVERSITÄT KLAGENFURT

Medical Informatics Europe MIE 2021 EFMI
VIRTUAL CONFERENCE
PUBLIC HEALTH AND INFORMATICS
FROM THE COMFORT OF YOUR HOME OR OFFICE
29TH - 31ST MAY, 2021
GBHI

2021/05/31
Biomedical data, tools and methods - Natural language processing

Clinical text data – lots of personal information



Discharge Summary

Provider: Ken Cure, MD

Patient: Patient H Sample **Provider's Pt ID:** 6910828 **Sex:** Female

Attachment Control Number: XA728302

protected data

HOSPITAL DISCHARGE DX

- ☐ 174.8 Malignant neoplasm of female breast: Other specified sites of female breast
- ☐ 163.8 Other specified sites of pleura.

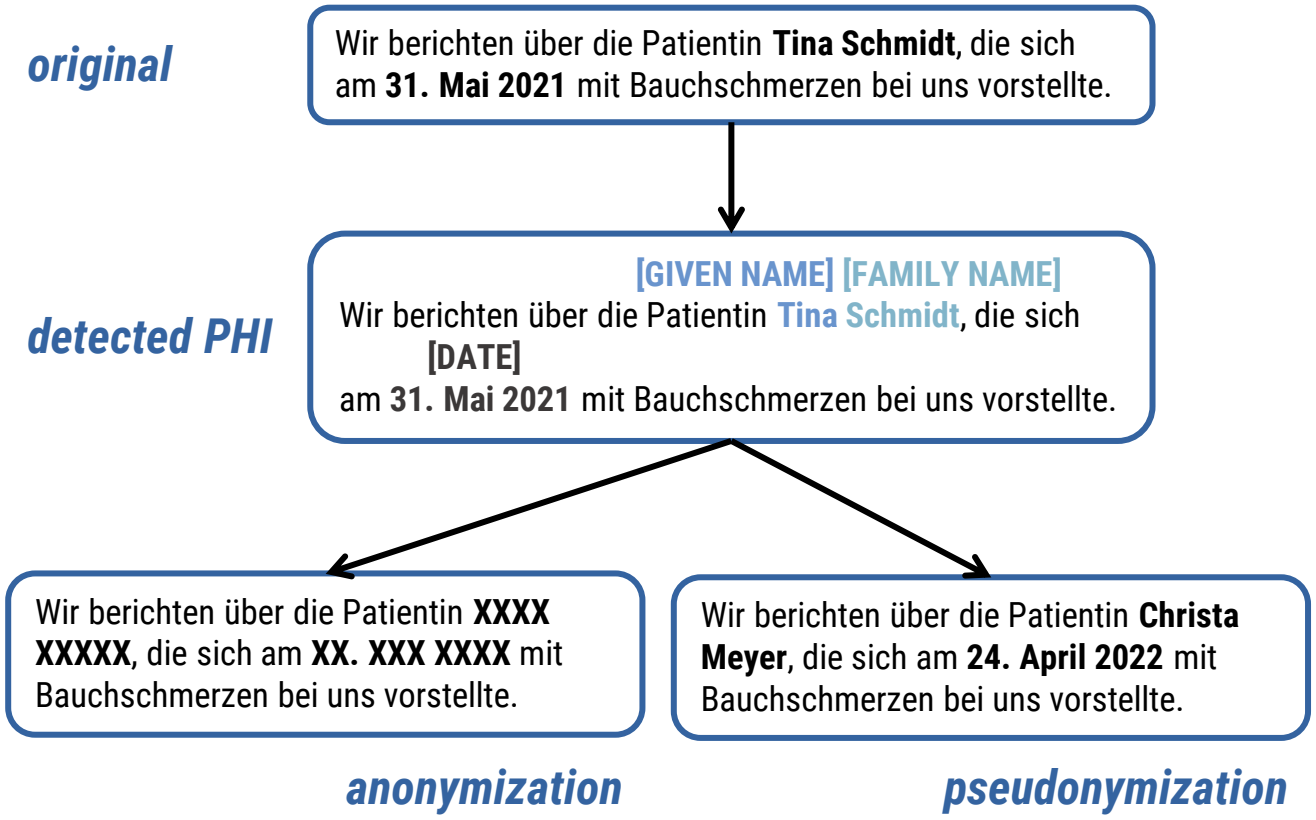
HOSPITAL DISCHARGE PROCEDURES

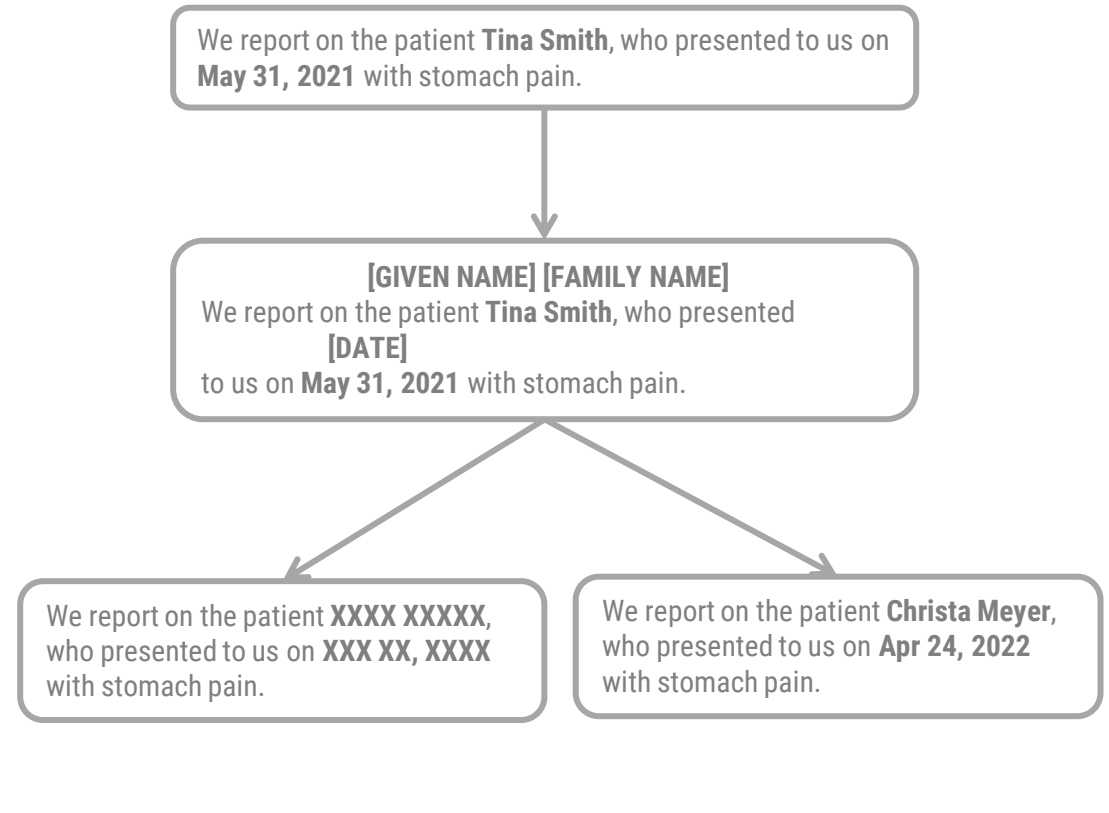
1. 32650 Thoracoscopy with chest tube placement and pleurodesis.

HISTORY OF PRESENT ILLNESS

The patient is a very pleasant, 70-year-old female with a history of breast cancer that was originally diagnosed in the early 70's. At that time she had a radical mastectomy with postoperative radiotherapy. In the mid 70's she developed a chest wall recurrence and was treated with further radiation therapy. She then went without evidence of disease for many years until the late 80's when she developed bone metastases with involvement of her sacroiliac joint, right trochanter, and left sacral area. She was started on Tamoxifen at that point in time and has done well until recently when she developed shortness of breath and was found to have a larger pleural effusion. This has been tapped on

Protected Health Information (PHI)





Protected Health Information

1	Names
2	Location
3	Dates
4	Phone numbers
5	Fax numbers
6	Electronic mail addresses
7	Social Security numbers
8	Medical Record Numbers
9	Health plan beneficary numbers
10	Account numbers
11	Certificate/license numbers
12	Vehicle identifiers
13	Device identifiers
14	URLs
15	IP addresses
16	Biometric identifiers
17	Images
18	Other

International Research

- US Health Insurance Portability and Accountability Act (HIPAA) (1996):
  list of 18 criteria without any subcategories
- **i2b2 de-identification challenges** (2006 and 2014): PHI annotated text corpora (English only)

Situation in Germany

- **data protection defined by EU law**
- **no concrete legal specifications defined (hence, no equivalent to HIPAA PHI)**

Protected Health Information

1. NAMES
2. LOCATION
3. DATES
4. PHONE NUMBERS
5. FAX NUMBERS
6. ELECTRONIC MAIL ADDRESSES
7. SOCIAL SECURITY NUMBERS
8. MEDICAL RECORD NUMBERS
9. HEALTH PLAN BENEFICARY NUMBERS
10. ACCOUNT NUMBERS
11. CERTIFICATE/LICENSE NUMBERS
12. VEHICLE IDENTIFIERS
13. DEVICE IDENTIFIERS
14. URLS
15. IP ADDRESSES
16. BIOMETRIC IDENTIFIERS
17. IMAGES
18. OTHER

HIPAA-PHI Original

Christina Lohr

↓

NAMES

Problem: internal structure of names

Christina Lohr

↙ ↘

GIVEN NAME FAMILY NAME

Protected Health Information

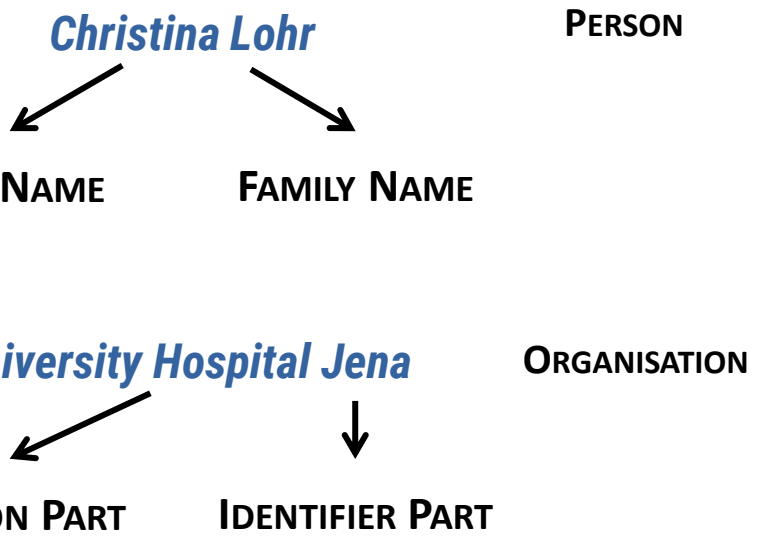
1. Names
2. Location
3. Dates
4. Phone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical Record Numbers
9. Health plan beneficary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers
13. Device identifiers
14. URLs
15. IP addresses
16. Biometric identifiers
17. Images
18. Other

HIPAA-PHI Original

Christina Lohr → **Names**

University Hospital Jena → **Names**

Problem: internal structure of names



Protected Health Information

1. NAMES
2. LOCATION
3. DATES
4. PHONE NUMBERS
5. FAX NUMBERS
6. ELECTRONIC MAIL ADDRESSES
7. SOCIAL SECURITY NUMBERS
8. MEDICAL RECORD NUMBERS
9. HEALTH PLAN BENEFICARY NUMBERS
10. ACCOUNT NUMBERS
11. CERTIFICATE/LICENSE NUMBERS
12. VEHICLE IDENTIFIERS
13. DEVICE IDENTIFIERS
14. URLS
15. IP ADDRESSES
16. BIOMETRIC IDENTIFIERS
17. IMAGES
18. OTHER

HIPAA-PHI Original

Christina Lohr → NAMES

University Hospital Jena → NAMES

Problem: internal structure of names

Christina Lohr → PERSON

Christina Lohr → GIVEN NAME, FAMILY NAME

University Hospital Jena → ORGANISATION

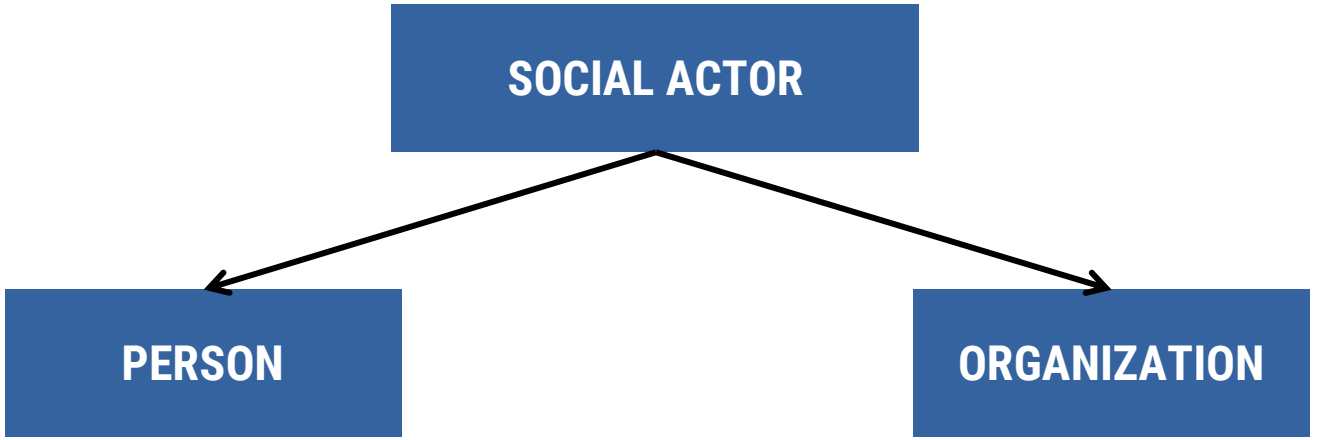
University Hospital Jena → COMMON PART, IDENTIFIER PART

Eder et al. (RANLP 2019):

- refining HIPAA-PHI into fine-grained categories

Protected Health Information – Replacement

1 NAMES
2 LOCATION
3 DATES
4 PHONE NUMBERS
5 FAX NUMBERS
6 ELECTRONIC MAIL ADDRESSES
7 SOCIAL SECURITY NUMBERS
8 MEDICAL RECORD NUMBERS
9 HEALTH PLAN BENEFICARY NUMBERS
10 ACCOUNT NUMBERS
11 CERTIFICATE/LICENSE NUMBERS
12 VEHICLE IDENTIFIERS
13 DEVICE IDENTIFIERS
14 URLS
15 IP ADDRESSES
16 BIOMETRIC IDENTIFIERS
17 IMAGES
18 OTHER

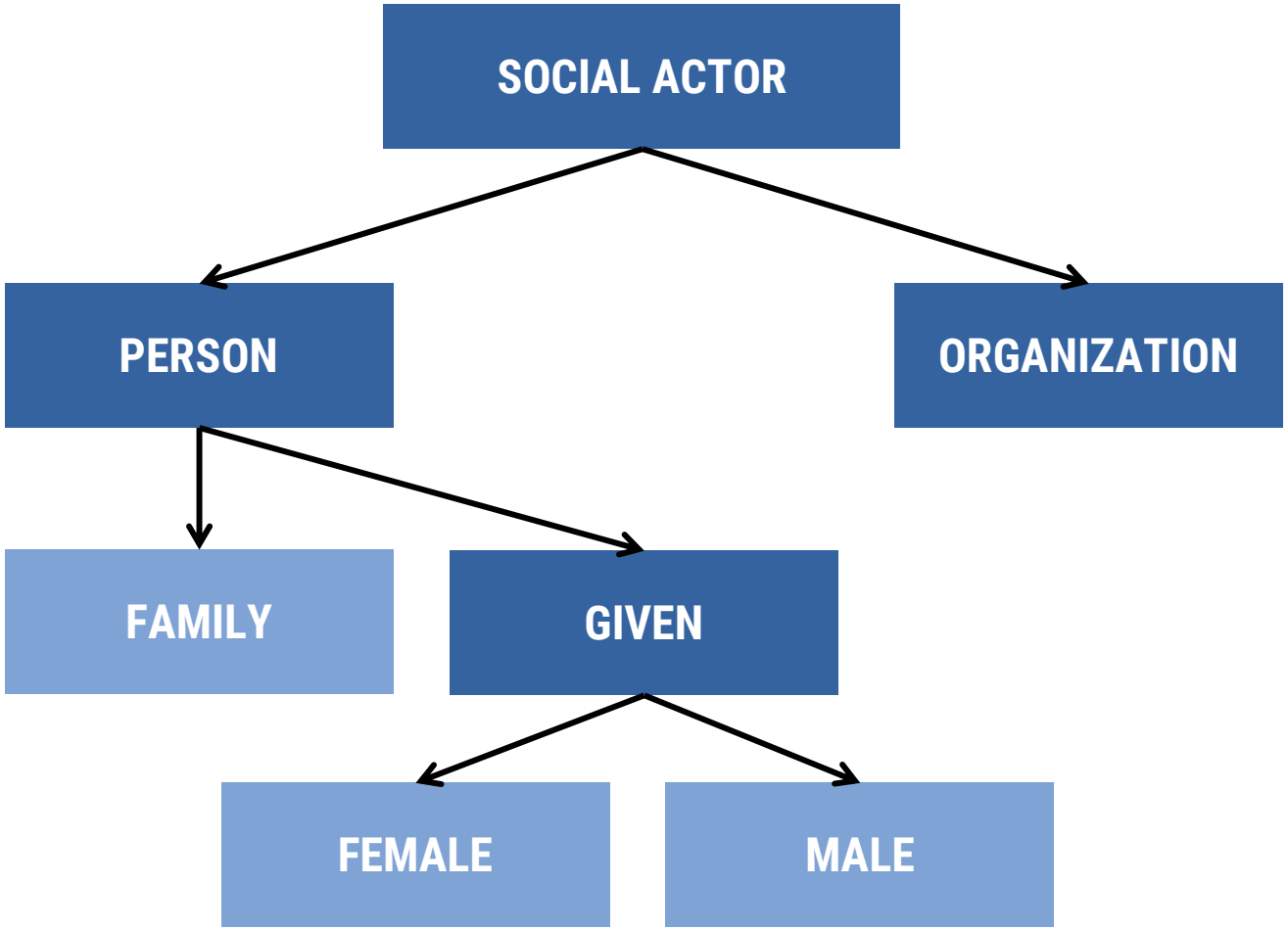


Christina Lohr

University Hospital Jena

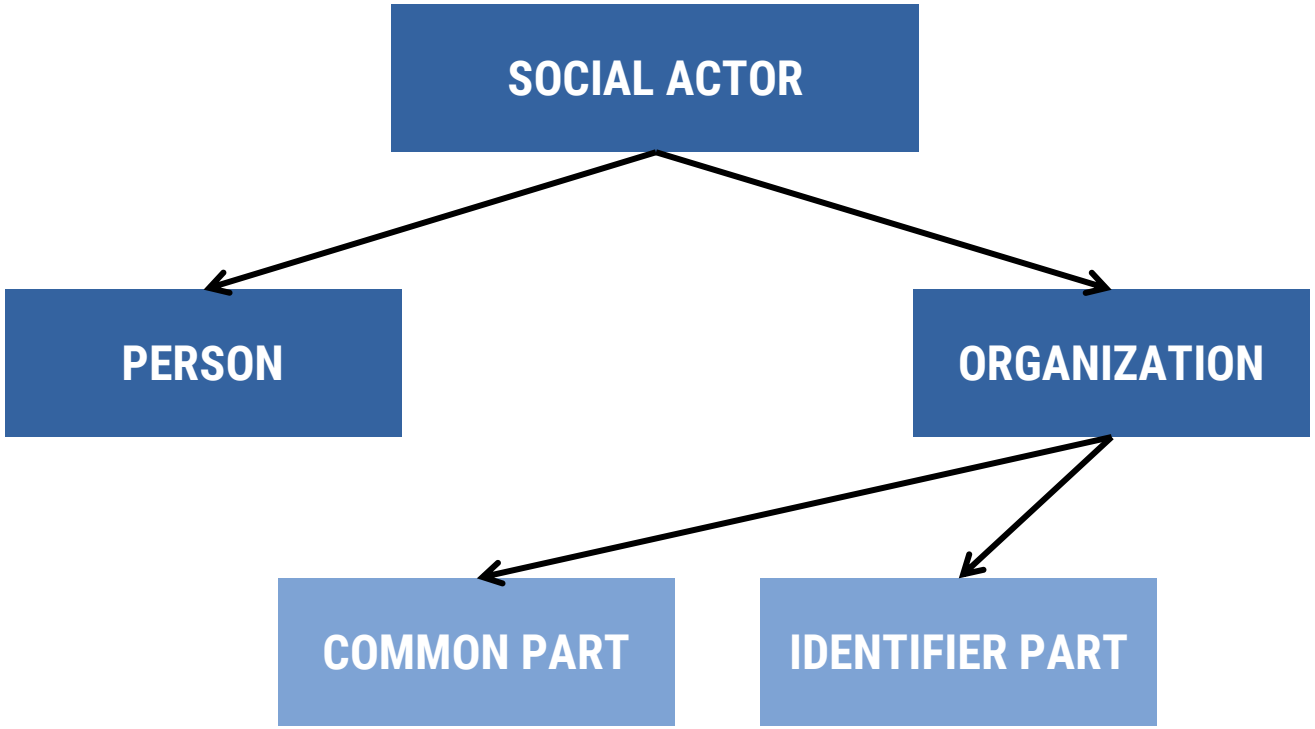
Protected Health Information – Replacement

1	NAMES
2	LOCATION
3	DATES
4	PHONE NUMBERS
5	FAX NUMBERS
6	ELECTRONIC MAIL ADDRESSES
7	SOCIAL SECURITY NUMBERS
8	MEDICAL RECORD NUMBERS
9	HEALTH PLAN BENEFICARY NUMBERS
10	ACCOUNT NUMBERS
11	CERTIFICATE/LICENSE NUMBERS
12	VEHICLE IDENTIFIERS
13	DEVICE IDENTIFIERS
14	URLS
15	IP ADDRESSES
16	BIOMETRIC IDENTIFIERS
17	IMAGES
18	OTHER



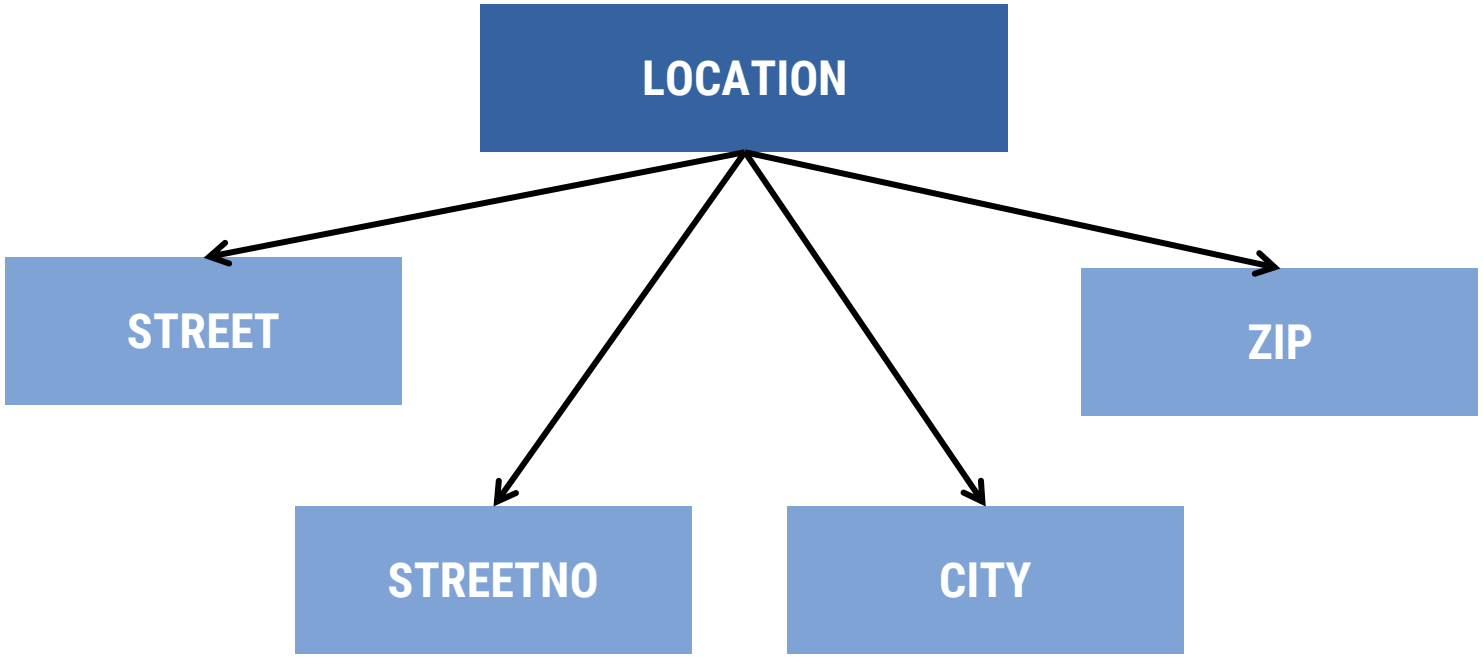
Protected Health Information – Replacement

1 NAMES
2 LOCATION
3 DATES
4 PHONE NUMBERS
5 FAX NUMBERS
6 ELECTRONIC MAIL ADDRESSES
7 SOCIAL SECURITY NUMBERS
8 MEDICAL RECORD NUMBERS
9 HEALTH PLAN BENEFICARY NUMBERS
10 ACCOUNT NUMBERS
11 CERTIFICATE/LICENSE NUMBERS
12 VEHICLE IDENTIFIERS
13 DEVICE IDENTIFIERS
14 URLS
15 IP ADDRESSES
16 BIOMETRIC IDENTIFIERS
17 IMAGES
18 OTHER



Protected Health Information – Replacement

1 Names
2 Location
3 Dates
4 Phone numbers
5 Fax numbers
6 Electronic mail addresses
7 Social Security numbers
8 Medical Record Numbers
9 Health plan beneficary numbers
10 Account numbers
11 Certificate/license numbers
12 Vehicle identifiers
13 Device identifiers
14 URLs
15 IP addresses
16 Biometric identifiers
17 Images
18 Other



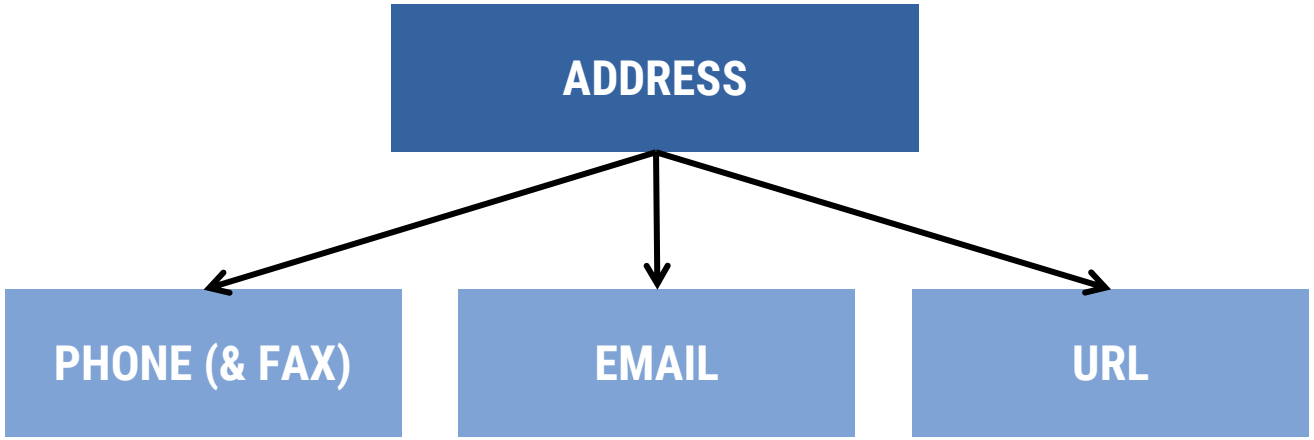
Protected Health Information – Replacement

1 Names
2 Location
3 Dates
4 Phone numbers
5 Fax numbers
6 Electronic mail addresses
7 Social Security numbers
8 Medical Record Numbers
9 Health plan beneficary numbers
10 Account numbers
11 Certificate/license numbers
12 Vehicle identifiers
13 Device identifiers
14 URLs
15 IP addresses
16 Biometric identifiers
17 Images
18 Other

DATES

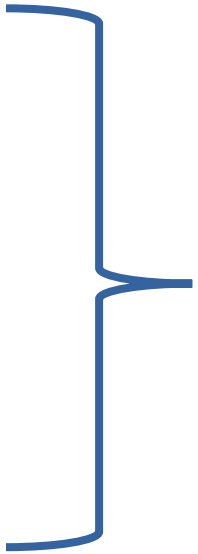
Protected Health Information – Replacement

1 Names
2 Location
3 Dates
4 Phone numbers
5 Fax numbers
6 Electronic mail addresses
7 Social Security numbers
8 Medical Record Numbers
9 Health plan beneficary numbers
10 Account numbers
11 Certificate/license numbers
12 Vehicle identifiers
13 Device identifiers
14 URLs
15 IP addresses
16 Biometric identifiers
17 Images
18 Other



Protected Health Information – Replacement

1 Names
2 Location
3 Dates
4 Phone numbers
5 Fax numbers
6 Electronic mail addresses
7 Social Security numbers
8 Medical Record Numbers
9 Health plan beneficary numbers
10 Account numbers
11 Certificate/license numbers
12 Vehicle identifiers
13 Device identifiers
14 URLs
15 IP addresses
16 Biometric identifiers
17 Images
18 Other

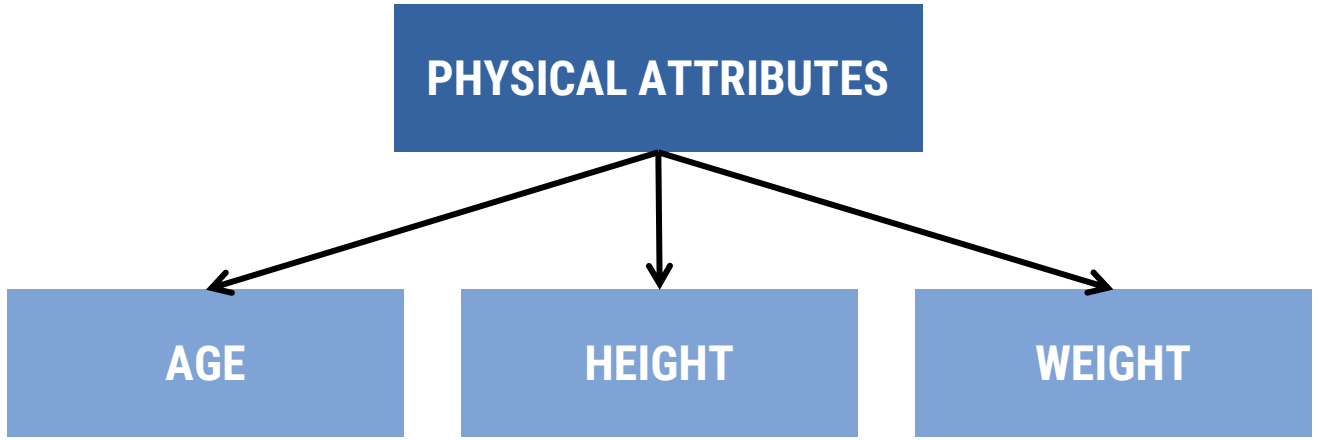


FORMAL IDENTIFIER

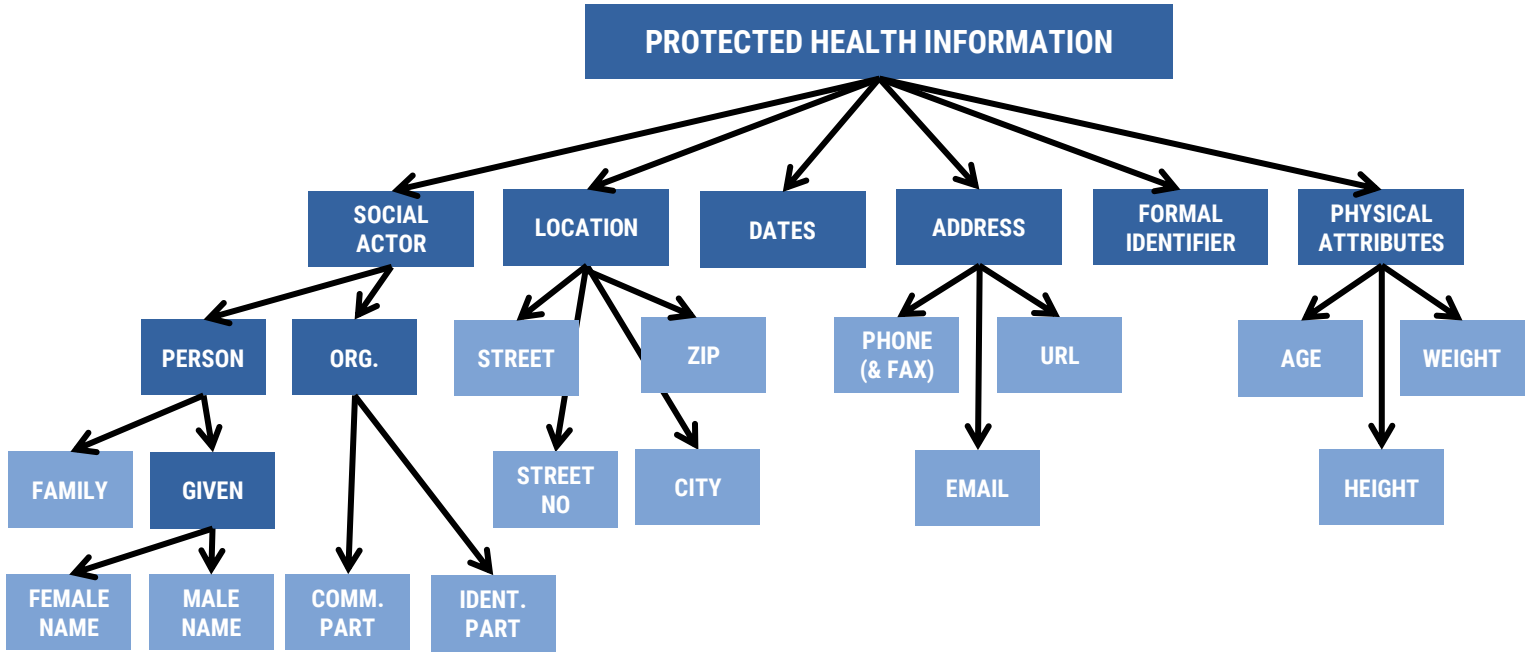
Protected Health Information – Replacement

1. Names
2. Location
3. Dates
4. Phone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical Record Numbers
9. Health plan beneficary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers
13. Device identifiers
14. URLs
15. IP addresses
16. Biometric identifiers
17. Images
18. Other

Extension:



Protected Health Information – Summary



Pseudonymization System

- extension of a rule-based surrogate generation system designed for emails (Eder et al., RANLP 2019)
- input: German text with PHI annotations
- output: German text with pseudonymized PHI items
- claim: easily adaptable to other Western European languages

Evaluation

data

- 300 German discharge summaries and case studies
- document sizes: ~1000 words and ~ 20 PHI items

evaluation criteria

- syntax, e.g., „Herr Meyer**s** Aufnahme“ → „Herr Schmidt**s** Aufnahme“
- semantic and clinical context, e.g., gender, age, treatment story in a plausible context

results

- error rate < 1 %
  - spelling alternatives of German umlauts (ä→ae)
  - false annotations
  - wrong date conversions

Conclusion

- first pseudonymization system for German clinical text documents
- re-engineering of HIPAA categories:
  - taxonomic structure
  - fine-grained extension of relevant categories
- open-source code and example annotations available:
  ➔ https://doi.org/10.5281/zenodo.4584505

Thank you!

Acknowledgements

This work was supported by BMBF within the SMITH project under grant 01ZZ1803G in cooperation with Jena University Hospital.

Kindly contact me at:

christina.lohr@uni-jena.de

www.julielab.de