The SURROGATOR Framework for Context-Aware Surrogation of Privacy Sensitive Information in Medical text

Christina LOHR[a], Marvin SEIFERLING[b], Philipp WIESENBACH[b], Jakob FALLER[c] and Christoph DIETERICH[b,d]

[a]*Institute for Medical Informatics, Statistics and Epidemiology (IMISE),University Leipzig, Germany*
[b]*Klaus Tschira Institute for Integrative Computational Cardiology, University Hospital Heidelberg, Germany*
[c]*Medical Center for Information and Communication Technology, University Hospital Erlangen, Germany*
[d]*Partner Site Heidelberg/Mannheim, German Centre for Cardiovascular Research (DZHK), Heidelberg, Germany*

2026/05/26

MEDIZIN INFORMATIK INITIATIVE
GeMTeX

Federal Ministry of Research, Technology and Space

Privacy sensitive Data in clinical Documents

PHYSICIAN HOSPITAL DISCHARGE SUMMARY

Provider: Ken Cure, MD

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

Attachment Control Number: XA728302

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 two occasions and has rapidly reaccumulated so she was admitted at this time for thoracoscopy with pleurodesis. Of note, her CA15-3 was 44 in the mid 90's and recently was found to be 600.

HOSPITAL DISCHARGE PHYSICAL FINDINGS

Physical examination at the time of admission revealed a thin, pleasant female in mild respiratory distress. She had

GeMTeX's list of Personally Identifiable Information (PII):

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

(Lohr et al 2024, 2025)

PII Transformation Modes

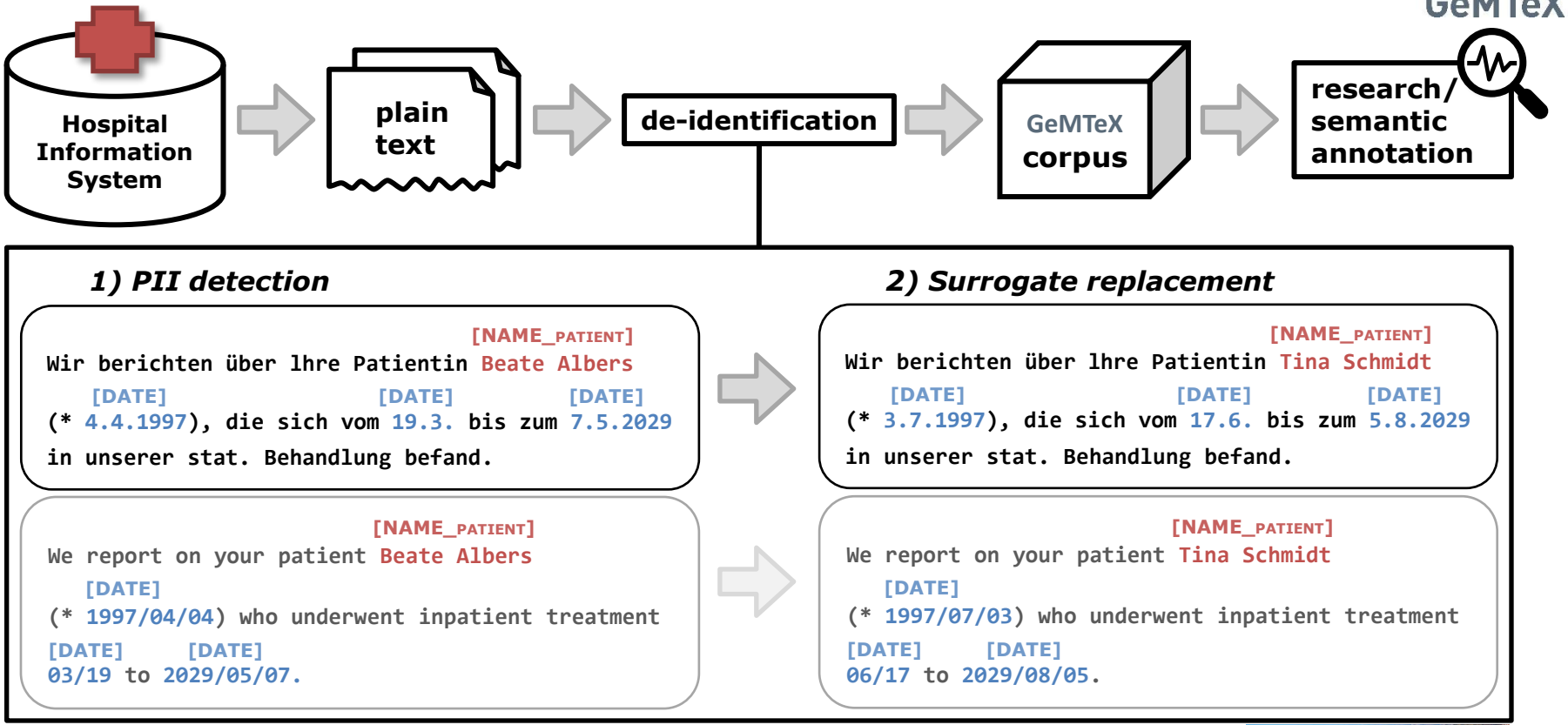
	Replacing text spans by
1) Redaction	character ‘X‘
2) Labeling	generic entity type; e.g., DATE
3) Traceable Surrogation	structured placeholder containing unique key for authorized re-linking; e.g., [** NAME PATIENT AB4D8F **]
4) Fictitious Surrogation	realistic, synthetically generated surrogates, consistent within a single document

preserve the document’s naturalness
what we prefer

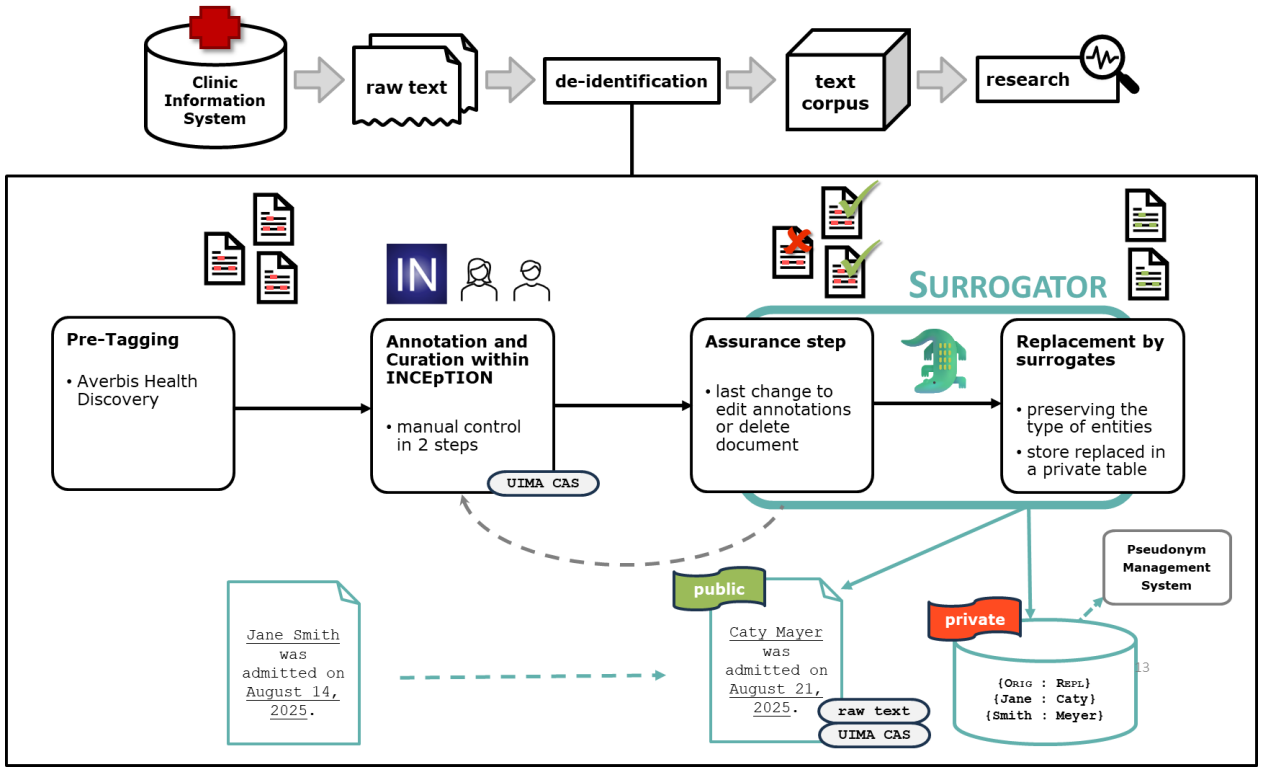
GeMTeX

- **Ger**man **M**edical **Tex**t Corpus
  - For Training, Evaluation and Finetuning of Large Language Models
  - 6 annotation sites
- De-Identification and our data protection concept

GeMTeX – De-Identification



GeMTeX – De-Identification & SURROGATOR



https://github.com/medizininformatik-initiative/GeMTeX/tree/main/surrogator

SURROGATOR

MEDICAL INFORMATICS INITIATIVE GERMANY

GeMTeX

Assurance step

- manual lookup by summary of PII annotations
- check non-PII sensitive data
- last change to edit annotations or delete document

Replacement by surrogates

- preserving the type of entities
- store replaced in a private table

Jane Smith was admitted on August 14, 2025.

public

Caty Mayer was admitted on August 21, 2025.

raw text

UIMA CAS

Pseudonym Management System

private

{ORIG : REPL}
{Jane : Caty}
{Smith : Meyer}

SURROGATOR's Strategies

PII	Key Technologies & Data Sources	Example		
		Original		*Surrogate*
NAME PATIENT	Gender detection, SPACY NER, curated name lists	***Dr. Inge Schmidt***	→	***Dr. Tina Meier***
DATE BIRTH, DATE DEATH	Rule-based date logic	***20.05.1950***	→	***01.04.1950***
DATE	Rule-based date logic	***Admitted 15.03.2024***	→	***Admitted 22.08.2023***
LOCATION *	SENTENCE TRANSFORMERS, OPENSTREETMAP	***Uniklinik Heidelberg***	→	***Klinikum Karlsruhe***
ID, CONTACT	Pattern-preserving regular expressions	***0176/1234567***	→	***0152/9876543***

Evaluation by GRASCCO

Graz Synthetic Clinical Text Corpus

63	synthetic discharge summaries
5,430	sentences
43,667	tokens

licence	CC 0 1.0 Universal („No copyright“)
download	https://doi.org/10.5281/zenodo.6539131
PII ann.	https://doi.org/10.5281/zenodo.15747389
more details	Lohr C, et al. **GeMTeX's De-Identification in Action: Lessons Learned & Devil's Details**. Stud Health Technol Inform. 2025 Sep 3;331:274-282. (https://pubmed.ncbi.nlm.nih.gov/40899551/)

Evaluation by GRASCCO & LLM's

1. Data Utility via NER
   - openai/gpt-oss-20b on the original corpus versus our surrogate version.
   - **F1-Score** (original) **0.73** → **F1-Score** (surrogate) **0.70**

2. Privacy via 2AFC Linkage Test
   - attacker (openai/gpt-oss-20b) chose between the true PII and its surrogate for a masked entity
   - accuracy **50.77%**, statistically equivalent to random chance (**50%**)

https://github.com/dieterich-lab/SurrogatorEval

The SURROGATOR Framework for Context-Aware Surrogation of Privacy Sensitive Information in Medical text

Christina LOHR, Marvin SEIFERLING, Philipp WIESENBACH, Jakob FALLER and Christoph DIETERICH

https://github.com/medizininformatik-initiative/GeMTeX/tree/main/surrogator

https://pubmed.ncbi.nlm.nih.gov/42175112

Christina Lohr
christina.lohr@imise.uni-leipzig.de
https://chlor.github.io

Christoph Dieterich
christoph.dieterich@med.uni-heidelberg.de
https://www.dieterichlab.org

2026/05/26

Federal Ministry of Research, Technology and Space