



Annotation Guidelines for Indirect Personal Identifiers (IPIs) in GraSCCo

Ibrahim Baroud, Vera Czehmann, Lisa Raithel, Sebastian Möller, Roland Roller
(Quality & Usability Lab, Technische Universität Berlin)

03 March 2026

1 Introduction

Medical notes include various direct and indirect identifiers that might reveal the identity of a patient. Personal identifiers within Personal Health Information (PHI), such as contacts, names, dates, IDs, locations and ages, can be reliably detected and then removed or obfuscated in a process known as de-identification (macro average F1: 0.936 on 2014 i2b2 Challenge) [1]. However, there is no clear, universal definition for indirect identifiers in texts, which makes processing them highly complex. Therefore, it is common practice to hire manual inspectors to highlight and process indirect identifiers in texts after de-identification. This, however, is very costly and time-consuming.

2 Motivation

The combination of indirect identifiers can lead to re-identification, and at the same time, manual inspection involves high time and labor costs. Therefore, we would like to analyze and label sensitive indirect identifiers in de-identified documents. Those annotations will be used as a gold standard dataset to measure the ability of (Large) Language Models (LMs) to detect sensitive indirect identifiers in de-identified documents.

3 Dataset

The texts we use for the annotation task comprise 63 clinical notes from the Graz Synthetic Clinical text Corpus (GraSCCo) dataset [2]. The German-language clinical text corpus contains more than 43,000 tokens covering a large variety of medical topics, such as neurology, oncology, and intensive care. These clinical texts originate from various sources such as hospitals, open access journals, and other texts published on the web. Each document in the corpus was manually anonymized by a medical doctor to diminish the possibility of re-identifying any persons mentioned in these texts. GraSCCo was annotated with the 18 PHI categories from HIPAA with the addition of the category *Profession*, and was released as an open-source dataset containing 1,438 PHI annotations.

4 Annotation task

Our annotation task employs all 63 documents from GraSCCo to label indirect identifiers on the token level. The sensitive information within the given documents can be descriptions of the patient's physical, mental, or circumstantial characteristics which might be detected by the patient's acquaintances or through other means such as a Google search. We exclude rare diseases, medications and medical devices or combinations of them from this analysis since they can be automatically detected using various tools such as [Metamap](#) [4]. Here are the categories for the sensitive indirect identifiers that must be considered during the annotation process:

1. **Body description (APPEARANCE):** The mention of a patient's (also infants) weight, height or a description of a patient's body or body modifications e.g. scar under the eye, very tall, very short, gained/lost weight over a specific period of time, tattoos, piercings,

or a split tongue. Only annotate descriptions of a patient's appearance that are permanently visible to the layman's eye. Do not annotate descriptions or mentions of medical procedures that imply scar/necrotic tissue, such as "patient has third-degree burns on scalp". Only annotate explicit mentions of appearance, such as "patient has multiple alopecic patches on scalp". Do not annotate diseases (diagnoses), even if they imply information about a patient's physique, e.g. "obesity". Do not annotate descriptions of a patient's physique, e.g. "normal/obese nutritional status".

2. **Circumstantial details (CURCUMSTANCES):** Any mention or description of an event (accident, storm, wildfire etc.) that caused, e.g., the patient's injury or happened in the clinical center, such as the patient being aggressive, rejecting help or medicine, explicitly agreeing to or requesting anything, information about level of participation, leaving AMA (also discussions in the regard with persons outside the family) or injuring hospital staff. Additionally, details about how the patient was brought into the hospital or mentions of statements, requests or complaints expressed e.g. by the patient. Do not annotate medical consequences of the event, e.g. "patient broke their arm".
3. **Socioeconomic, criminal history (SEC):** A mention of specific information about the patient's employment and occupation (e.g., "patient is a retired police officer", "goes to kindergarten") or criminal history, health insurance (e.g., "patient has no health insurance" or "patient has a legal guard") or social status such as homelessness or living in subsidized housing.
4. **Family details (FAMILY):** All mentions of family-related information about the patient, such as having a spouse, being adopted, having a twin sibling or having had a vitro fertilization pregnancy. Furthermore, specific descriptions of the family's medical history (e.g., parent died at age of X) or involvement (e.g., "patient's daughter serves as her health care proxy").
5. **Healthcare Facilities (HEALTH_FCLT):** All mentions of hospital names, hospital units, labs, departments, facilities, consulting services/teams, floor and rooms, medical branches, medical practices/offices, doctors and whether a patient receives inpatient or outpatient care.
6. **Relative Time (TIME):** All mentions of time-related information, e.g., yesterday, postoperative day number X, day of delivery number X, mentions of times when lab values were taken, or mentions of when medications should be taken (including medication schedules, e.g. "1-1-0", indicating medication to be taken mornings and noon, but not in the evening).
7. **Hobbies and Lifestyle (HOBIS_LFSTL):** All mentions about sports, playing an instrument and lifestyle in general, e.g., information about the patient's diet, tobacco, alcohol or other substance use (including frequency and quantities), unless a diagnosis is given such as "alcohol abuse", or private lifestyle.
8. **Details about a direct identifier (DETAILS_ID):** All mentions of PHI categories that e.g. were not detected and de-identified automatically or a description of a PHI. Any

information that is not related to PHIs such as weight or medical units is not part of this category and should be annotated as described in the other categories.

- a. PHI descriptions regarding location (e.g. “patient lives in a halfway house” or “patient lives in prison”) should be annotated as `DETAILS_LOCATION`.
 - b. For consistency, we consider PHI categories according to [GeMTeX](#) ([HIPAA](#) categories adapted for German clinical reporting habits and legal requirements) [3]. The categories include `NAME`, `DATE`, `AGE`, `(LOCATION)`, `ID`, `CONTACT`, `PROFESSION`, `OTHER`, and respective subcategories.
9. **Details about location (`DETAILS_LOCATION`):** PHI descriptions regarding location, e.g. “patient lives in a halfway house” or “patient lives in prison”.
 10. **Ethnicity (`ETHNICITY`):** All mentions of ethnicity, e.g. “patient’s family has both Black and Hispanic heritage”.
 11. **Languages and speech (`LANGUAGES`):** All mentions of languages, e.g. “patient has been raised bilingual and uses both English and Spanish in their everyday life”, atypical speech patterns (unless diagnosed as speech disorder), and regional varieties.
 12. **Sexual orientation (`SEXUAL_ORIENTATION`):** All mentions of sexual orientation, e.g. “while the patient is in a heterosexual relationship, she identifies as bisexual”.
 13. **Others (`OTHER`):** Other kinds of non-medical information that might be sensitive to keep in the data.

5 Annotation Rules

1. Assume that the de-identified discharge summary written about a particular patient will be shared with researchers from another institute.
2. You want to reduce the risk of re-identification by annotating sensitive identifiers as described in Section 4.
3. Read the entire discharge summary to get a general sense of its content.
4. Read the discharge summary again and look for information similar to the examples mentioned in Section 4.
5. Highlight the sensitive information you find in the discharge summary and assign the respective category as described above.
6. Prioritize highlighting essential keywords and phrases rather than whole sentences.
7. Highlight the entity and the unit in the case of numerical information e.g. “postoperative day 17” or “weight is 2620g”, the whole span should be highlighted and not only the numerical values.
8. If the summary includes the sensitive information multiple times, highlight all of the occurrences.
9. Nested annotations are allowed, although separate annotation is preferred if possible. In case of `HOBIS_LFSTL`, nested annotation of frequency or other mentions of `RELATV_TIME` that refer to substance use or hobbies is not necessary.

6 Annotation Tool

We use INCEpTION (version 35.2) as the annotation tool due to our previous expertise with it and its easy usage for both annotation instructors and annotators [5].

7 References

- [1] Stubbs, A., Kotfila, C., & Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of biomedical informatics*, 58 Suppl(Suppl), S11–S19. <https://doi.org/10.1016/j.jbi.2015.06.007>.
- [2] Modersohn, L., Schulz, S., Lohr, C., & Hahn, U. (2022). GRASCCO - The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. *Studies in health technology and informatics*, 296, 66–72. <https://doi.org/10.3233/SHTI220805>.
- [3] Lohr C, Matthies F, Faller J, Modersohn L, Riedel A, Hahn U, Kiser R, Boeker M, Meineke F. De-Identifying GRASCCO - A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus. *Stud Health Technol Inform*. 2024 Aug 30;317:171-179.
- [4] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium* (p. 17).
- [5] Klie, J. C., Bugert, M., Boullosa, B., De Castilho, R. E., & Gurevych, I. (2018, August). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations* (pp. 5-9).