*Public Health and Informatics*
*J. Mantas et al. (Eds.)*


*doi:10.3233/SHTI210163*

# Pseudonymization of PHI Items in German Clinical Reports

Christina LOHR[a,1], Elisabeth EDER[b] and Udo HAHN[a]
[a]*Jena University Language & Information Engineering (JULIE) Lab*
*Friedrich-Schiller-Universität Jena, Jena, Germany &*
*SMITH Consortium of the German Medical Informatics Initiative*
[b]*Institut für Germanistik, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria*

**Abstract.** We describe the adaptation of a non-clinical pseudonymization system, originally developed for a German email corpus, for clinical use. This tool replaces previously identified Protected Health Information (PHI) items as carriers of privacy-sensitive information (original names for people, organizations, places, etc.) with semantic type-conformant, yet, fictitious surrogates. We evaluate the generated substitutes for grammatical correctness, semantic and medical plausibility and find particularly low numbers of error instances (less than 1%) on all of these dimensions.

**Keywords.** pseudonymization of clinical reports, Protected Health Information (PHI), German-language clinical reports, surrogate generation

## 1. Introduction

One of the main reasons for the impressive advancements in nearly all branches of natural language processing (NLP) is the abundant mass of data accessible for training and operating NLP engines. While their petabyte dimension seems out of reach for clinical applications for the time being, massive regulatory obstacles to increasing volumes of clinical raw data and distributing them within the NLP community are in place. Such legal constraints securing individual data privacy are imposed on all sorts of personalized medical documents (almost) all over the world (cf. the General Data Protection Rule (2016/679) for the EU [1] or the Health Insurance Portability and Accountability Act (HIPAA) for the USA) [2]. HIPAA, for instance, enumerates 18 Protected Health Information (PHI) categories that need to be obfuscated when patient data leave the secured walls of any local clinical site. Human efforts to break this de-identification bottleneck are laborious, costly and error-prone. Current automatic approaches to de-identification achieve a detection rate for PHI items between 94% up to 99%, yet seem unable to close the remaining gap completely [3, 4].

In order to deal with such unrecognized PHI items in an ingenious way, the de-identification problem can be rephrased as a pseudonymization task. Pseudonymization (an approach originating from the seminal work of Sweeney [5]) replaces text stretches which contain confidential, i.e., privacy-sensitive, information by semantic type-

[1] Corresponding Author, JULIE Lab, FSU Jena, Fürstengraben 30, 07743 Jena, Germany. Contact: christina.lohr@uni-jena.de

conformant, natural, yet at the same time fictitious (and hence non-confidential) surrogates (e.g., the invented name "Jessica Johnson" consistently replaces the original patient name "Jennifer White" in the entire document). As an outstanding advantage, this kind of camouflage is particularly robust against unwarranted effects of de-identification errors, since potential leaks due to an unrecognized PHI item (say, "Jennifer White" remains unchanged as "Jennifer White") are not at all obvious to spot and thus hard to decipher in terms of re-identification of concrete individuals. As a consequence, privacy information might not be leaked despite incomplete de-identification, an assumption referred to as "Hiding in Plain Sight" (HiPS) [6].

The majority of work on medical pseudonymization, no surprise, has been conducted for the English language, with data coming from well-known repositories allowing access via Data Use Agreements (clinical notes from MIMIC II [7], Physionet and i2b2 [8, 9]), while several non-sharable datasets have been created as well [10, 11, 12]. As far as non-English clinical language is concerned, French [12], Danish [13], Swedish [14] and Dutch [15] EHRs have also been dealt with.

Deeper evaluation efforts related to the re-identification risk after pseudonymization began with two studies reporting encouraging evidence (on small-scale data sets though, incorporating less than 100 patient records) that experienced physicians were unable to re-identify patients they had been treating from pseudonymized record data [10, 12]. In another small-scale experiment testing the purported 'naturalness' of substitutions, evaluators, by and large, were unable to recognize pseudonymized documents (only 3.5% of these documents were correctly identified whereas in 1.5% of the cases they categorized non-pseudonymized documents as pseudonymized) [14]. Recent experiments on larger data scales study much more aggressive attack scenarios, both in machine-supported [3] and human expert-centered [4] re-identification settings. Both studies conclude that such massive attacks can attenuate, but not eliminate, the protective effect of pseudonymization. Furthermore, enormous man- and compute-power are needed on the attackers' side to organize and run such operations.

## 2. Methods

**Data.** We used two German-language clinical corpora, namely the Jena part of the national reference corpus 3000PA [16] which contains 1,106 real discharge summaries sampled from the Jena University Hospital's information system (approved by the local ethics committee (4639-12/15)), and JSYNCC [17], a complementary corpus made of German medical textbook documents mimicking clinical reports for educational purposes (chosen to boost the volume of experimental data and genre diversity). 3000PA was annotated for HIPAA PHI categories with an instance-based inter-annotator agreement (IAA) of F1=0.96 [16]. JSYNCC was also annotated for these HIPAA categories, with an IAA of F1=0.79. For this study, we had to adjust the HIPAA category system, since, e.g., in the original schema, a generic Name category was introduced whereas in our refined type system (see description below), we further divide this category into *Female Name* and *Male Name*, as well as *Given* and *Family Name*. The same applies to *location identifications*, e.g., *streets* and *cities*, which are generally summarized as Locations by HIPAA. Basic corpus statistics of the original data, the original annotated HIPAA PHI entities ($PHI_{HIPAA}$), and the evaluation subset are summarized in Table 1.

**Table 1.** Amount of Doc(uments), Sent(ences), Tokens, and PHI entities, distinguishing the original HIPAA categories from the refined PHI annotation type scheme and surrogates for identified PHI items, both for the reference corpora and their associated evaluation subsets

| Corpus | | Doc. | Sent. | Tokens | $PHI_{HIPAA}$ | $PHI_{Refinded}$ | $PHI_{Surrogates}$ |
|---|---|---|---|---|---|---|---|
| **3000PA** | Original | 1106 | 196k | 1.709k | 44,165 | 57,303 | 51,120 |
| | Eval. subset | 100 | 20k | 181k | 5,163 | 6,115 | 4,643 |
| **JSynCC** | Original | 903 | 33k | 406k | 3,960 | 4,406 | 4,374 |
| | Eval. subset | 200 | 9k | 102k | 1,185 | 1,250 | 1,154 |

**Refined PHI Annotation Type Scheme for Pseudonymization.** Originally, the PHI type scheme we employ here was developed for the pseudonymization of German email documents [18]. It contains five top-level categories. The first one, *SocialActor*, is split into three subtypes—*Organization*, *Person*, and *User*. *Organization* includes all types of *legal actors*, e.g., *companies* and *institutions*. *Natural actors* fall under human *Person* (*including patients*, *their relatives*, *clinical staff*) and are assigned the subtype *Names*, with further subtypes *Family Name*, and *Given Name*, the latter subdivided into *Female Name* and *Male Name*. Finally, *User Name* covers all kinds of artificial names for users of IT systems. *Date* is the second top-level type covering e.g., *date of birth*, starting and ending dates of hospital stays. The third top-level type, *Formal Identifier*, subsumes *Password* as user-defined access authorization code for technical appliances, and *Unique Formal Identifier* to capture id codes for persons (patient ids, typist ids, etc.). The fourth top-level type, *Location*, subsumes *Street Name*, *Street Number*, *Zip Code*, and *City Name*. Finally, the fifth top-level type, *Address*, comprises *Email Address*, *Phone Number*, and *URL*, including other forms of domain names.

For the clinical application, we extended this type hierarchy by two subtypes. The top-level *Organization* type was assigned the subtype *Medical Unit* which is technically divided into a *common* and an *identifier part*: identifying location and person names of institutions and station or room numbers are subject to obfuscation whereas its common part is kept as-is in the surrogation step (the specific disease characteristics of a patient treated at, say, a "Department of Dermatology" cannot be naturally preserved when department names are arbitrarily exchanged). The second extension relates to the *Person type*, with an additional subtype, *Physical Attributes*, which subsumes a person's Age in years, *Height* in (centi)meters, and *Weight* in (kilo)grams.

**Pseudonymization System.** After transforming PHI-sensitive text mentions as defined by the above entity type system, pseudonymization requires generating a surrogate for each PHI instance by transforming the original text mention into a type-conformant artificial mention substitute. For this task, we extended a rule-based surrogate generation system designed for non-medical purposes (see [18] for details). Following clinical conventions for obfuscations, we implemented cut-offs at Age > 89 and shifts of Height and Weight by some constant increment while preserving the Body Mass Index (BMI) as a non-identifying attribute.

## 3. Results

The PHI columns in Table 1 display the number of text mentions of standard HIPAA categories ($PHI_{HIPAA}$), those resulting from our refined type system ($PHI_{Refined}$), and the surrogates produced on the basis of the latter ($PHI_{Surrogates}$). From the 3000PA and

JSYNCC corpus, we selected 100 and 200 documents, respectively, for evaluation. Our evaluators were all medical students and native speakers of German. Overall, the results are more than promising. We encountered no morpho-syntactic error at all (which is not a trivial result because of the rich morphological constraints German language has to obey). Only the 3000PA corpus produced five low-level language-bound semantic errors due to spelling alternatives for German umlauts (e.g., ö→oe). We found 23 domain knowledge-bound medical errors in 3000PA and three in JSYNCC. Some of them were due to implausible date conversions, only two errors were due to false annotations (e.g., a typist's name was annotated as medical staff name). Some odd results, not necessarily false, also popped up. For instance, typical Arabic or Asian names were replaced with typical German substitutes, yet a migration background of that patient was linked with the anamnesis.

## 4. Discussion

Porting the pseudonymization system originally developed for email documents to the clinical domain turned out to be a comparatively straightforward task, once clinic-specific adaptations of the type system had been made. Furthermore, standard HIPAA categories had to be refined at a finer level of granularity to allow for natural and informative surrogates (e.g., preserving (fe)male names). With an overall error rate of way below 1%, the clinical pseudonymization system we describe here is robust enough to be deployed on a larger, routine scale. While naturalness of substitutions seems preserved, we have not been dealing with the re-identification risk of our approach up until now (see the brief discussion in the final paragraph of Section 1).

## 5. Conclusion

We here presented the first pseudonymization system for German clinical documents. With only a few adaptation steps (updates of the type hierarchy by clinically relevant PHI categories and changes at the code level to cope with these extensions) the original pseudonymization system primarily built for an email corpus [19] can be easily reused in the clinical domain. The evaluation results show that error rates for machine-generated surrogates are negligible. Due to its modular design and transparent implementation (cf. [18]), the pseudonymization procedure can easily be ported to other (European) languages as well (but note that even within the English-speaking language community such a transfer is not so straightforward as it seems because, e.g., different date schemes or zip code patterns have to be dealt with [11]).

Integrating the perspectives from two entirely different application domains (clinical reports vs. emails) revealed interesting insights. In particular, it turned out that finer levels of granularity than those defined by HIPAA-style categories were needed and seemingly unrelated categories (e.g., physical attributes of a patient) could be subsumed under unifying types. Hence, this work also contributes to a better understanding of what constitutes person-identifying information under digital privacy considerations. The PHI annotations for JSYNCC, corresponding type-conformant pseudonyms and programming code for the pseudonymization system are available under
https://doi.org/10.5281/zenodo.4584505

**Acknowledgements**
This work was supported by BMBF within the SMITH project under grant 01ZZ1803G. We thank the evaluation team, as well as André Scherag, Danny Ammon, and all members of the Data Integration Center of the Jena University Hospital for their support.

## References

[1] EUR-Lex, Available at https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04 .
[2] HHS.gov, Available at https://www.hhs.gov/hipaa/index.html.
[3] Carrell D, Cronkite D, Li M, Nyemba S, Malin B, Aberdeen J, and Hirschman L. The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with hiding in plain sight. JAMIA 2019; 26(12):1536–44.
[4] Carrell D, Malin B, Cronkite D, Aberdeen J, Clark C, Li M, Bastakoty D, Nyemba S, and Hirschman L. Resilience of clinical text de-identified with "hiding in plain sight" to hostile reidentification attacks by human readers. JAMIA 2020; 27(9):1374–82.
[5] Sweeney L. Replacing personally-identifying information in medical records, the SCRUB system. In: AMIA 2004 – Proceedings of the 1996 AMIA Annual Fall Symposium. Washington, D.C., USA, October 26-30, 1996. p. 333–7.
[6] Carrell D, Malin B, Aberdeen J, Bayer S, Clark C, Wellner B, Hirschman L. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. JAMIA 2013; 20(2):342–8.
[7] Douglass M, Clifford G, Reisner A, Moody G, and Mark R. Computer-assisted de-identification of free text in the MIMIC II database. In: CinC 2004 – Proceedings of the 31st Annual Scientific Conference on Computers in Cardiology 2004. Chicago, Illinois, USA, September 19-22, 2004. p. 341–4.
[8] Deléger L, Lingren T, Ni Y, Kaiser M, Stoutenborough L, Marsolo K, Kouril M, Molnar K, Solti I. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. JBI 2014; 50:173–83.
[9] Hartman T, Howell MD, Dean J, Hoory S, Slyper R, Laish I, Gilon O, Vainstein D, Corrado G, Chou K, Po MJ, Williams J, Ellis S, Bee G, Hassidim A, Amira R, Beryozkin G, Szpektor I, Matias Y. Customization scenarios for de-identification of clinical notes. BMC Med Inform Decis Mak. 2020 Jan 30;20(1):14.
[10] Meystre S, Shen S, Hofmann D, Gundlapalli A. Can physicians recognize their own patients in de-identified notes? Stud Health Technol Inform. 2014;205:778-82.
[11] Chen A, Jonnagaddala J, Nekkantti C, Liaw ST. Generation of Surrogates for De-Identification of Electronic Health Records. Stud Health Technol Inform. 2019 Aug 21;264:70-73.
[12] Grouin C, Griffon N, and Névéol A. Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? In: LOUHI 2015 – Proceedings of the Workshop on Health Text Mining and Information Analysis @ EMNLP 2015. Lisbon, Portugal, 17 Sept. 2015, p.31-9
[13] Pantazos K, Lauesen S, and Lippert S. Preserving medical correctness, readability and consistency in de-identified health records. Health Inform J 2017; 23(4):291–303.
[14] Dalianis H. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In: Proceedings of the Workshop on NLP and Pseudonymisation @ NoDaLiDa 2019. Turku, Finland, September 30, 2019. p. 16–23.
[15] Trienes J, Trieschnigg D, Seifert C, and Hiemstra D. Comparing rule-based, feature-based and deep neural methods for de-identification of Dutch medical records. In: HSDM 2020 – Proceedings of the Health Search and Data Mining Workshop @ WSDM 2020. Houston, TX, USA, Feb. 3, 2020. p. 3–11.
[16] Kolditz T, Lohr C, Hellrich J, Modersohn L, Betz B, Kiehntopf M, Hahn U. Annotating German Clinical Documents for De-Identification. Stud Health Technol Inform. 2019 Aug 21;264:203-207.
[17] Lohr C, Buechel S, and Hahn U. Sharing copies of synthetic clinical corpora without physical distribution: a case study to get around IPRs and privacy constraints featuring the German JSYNCC corpus. In: LREC 2018 – Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018. p. 1259–66.
[18] Eder E, Krieg-Holz U, and Hahn U. De-identification of emails: pseudonymizing privacy-sensitive data in a German email corpus. In: RANLP 2019 – Proceedings of the 12th International Conference on "Recent Advances in Natural Language Processing:" Varna, Bulgaria, September 2-4, 2019. p. 259–69.
[19] Eder E, Krieg-Holz U, and Hahn U. CODE ALLTAG 2.0: a pseudonymized German-language email corpus. In: LREC 2020 – Proceedings of the 12th International Conference on Language Resources and Evaluation. Marseille, France, May 11-16, 2020, pp. 4466-77.