Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs?

Cyril Grouin
LIMSI-CNRS, UPR 3251
Rue John von Neuman
91400 Orsay, France
grouin@limsi.fr

Nicolas Griffon
LIMICS, INSERM U 1142
CISMeF-TIBS-LITIS, EA 4108
CHU de Rouen, 76031 Rouen
nicolas.griffon
@chu-rouen.fr

Aurélie Névéol
LIMSI-CNRS, UPR 3251
Rue John von Neuman
91400 Orsay, France
neveol@limsi.fr

Abstract

De-identification aims at preserving patient confidentiality while enabling the use of clinical documents for furthering medical research. Herein, we aim to evaluate whether patient re-identification is possible on a corpus of de-identified clinical documents in French. Personal Health Identifiers are automatically marked by a de-identification system applied to the corpus, followed by reintroduction of plausible surrogates. The resulting documents are shown to individuals with varying knowledge of the documents and de-identification method. The individuals are asked to re-identify the patients. The amount of information recovered increases with familiarity with the documents and/or de-identification method. Surrogate re-introduction with localization from the same (vs. different) geographical area as the original documents is found more effective. The amount of information recovered was not sufficient to re-identify any of the patients, except when privileged access to the hospital health information system and several documents about the same patient were available.

1 Introduction

Research using clinical data requires the informed consent of patients involved. Privacy rules and regulation in France require that, in the absence of informed consent, clinical records used in research be anonymized or de-identified.

Anonymization consists in ensuring that health data used in the research can not be linked to individual patients. Alternatively, de-identification consists in removing or hiding personal health identifiers found in health documents (Meystre et al., 2010). In this study, we focus on the result of an automatic de-identification process. Both anonymization and de-identification aim at preserving patient confidentiality while enabling the use of clinical documents for furthering medical research. State-of-the-art automatic de-identification methods are often evaluated for their ability to redact a set of personal health identifiers (PHI) from clinical documents (Meystre et al., 2010). PHIs are defined according to the American Health Insurance Portability and Accountability Act (HIPAA) of 1996*.

In this study, we are investigating whether it is possible for individuals to recover patients' personal information based on the content of automatically de-identified documents. We caracterize the re-identification attempts using the skills, tools or information at the attacker's disposal. The targets of the re-identification attempts can be both surrogates wrongly used in replacement of original PHI and data not processed during the de-identification step (data missed during the de-identification process as well as data not being in the scope of this process).

Assessing whether patients can be re-identified after documents have been automatically de-identified is a difficult task, since the combination of seemingly innocuous pieces of information could endanger patient privacy (Benitez and Malin, 2010; Barbaro and Zeller Jr, 2006). The combination of a de-identification system that automatically tags PHIs in clinical text with the replacement of PHIs by plausible surrogates has been used to create realistic modified clinical records (Sweeney, 1996; Neamatullah et al., 2008) that are clinically and linguistically valid. This

*U.S. Department of Health Human Services, 1996 http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 31–39,
Lisbon, Portugal, 17 September 2015.

method is also referred to as "hiding in plain sight" obfuscation and was shown to contribute to increase the effective recall of automatic de-identification systems from about .94 to .99 (Carrell et al., 2013).

While the impact of de-identification on clinical information contained in the records has been studied (Deleger et al., 2013; Meystre et al., 2014b), fewer efforts have addressed the effective impact on patient privacy. Encouragingly, it was recently shown that doctors were not able to identify patients they had recently treated when relying on de-identified records (Meystre et al., 2014a). There is a need for other studies that evaluate whether re-identification is possible based on de-identified records.

In this study, we address re-identification attempts from the perspective of making a small de-identified clinical corpus available to the research community in circumstances such as a shared task or NLP challenge. Although a dataset released in the context of a shared task or challenge would require participants to sign a user agreement specifically binding recipients to *not* engage in re-identification attempts, this study considers an attack scenario where a negligent (or malignant) user would overlook this requirement.

In this context, we anticipate that the corpus would be accessed by individuals with a variety of backgrounds including researchers, developpers and clinicians. Furthermore, depending on the type of NLP task addressed by the challenge, there may be a need to include several documents pertaining to the same patient (e.g., to evaluate systems that create a cross-document patient timeline) or not (e.g., to evaluate systems that perform named-entity or concept recognition).

Accordingly, we consider re-identification attempts by individuals with varied knowledge of clinical records and de-identification methods (medical doctors and computer scientists) on automatically de-identified records in French. In addition, we also assess the success of re-identification attempts on different types of datasets (documents pertaining to the same patient, vs. random patients) and surrogate re-introduction methods (using localization information similar to that of original documents, vs. different). The corpus used in our study has been automatically de-identified by a system, without any human intervention to check the outputs produced by the automatic process.

2 Background

The release of datasets containing personal information about the individuals who contributed to the creation of the data raises the concern of privacy protection. When such datasets are prepared for research purposes, the risks of privacy breach must be assessed and weighed against the potential benefits the research conducted using the data. In prior instances of data release, inadequate assessment of the possibility of privacy breach has led to public embarrassment and legal action (Barbaro and Zeller Jr, 2006). In light of this experience, extreme caution is needed prior to releasing sensitive data. The case of medical data such as those contained in Electronic Health Records requires specific attention, since the first rule of medical ethics as outlined in the Hippocratic Oath is to "first, do no harm". This makes it unethical to release medical data that could cause harm to a patient, e.g., through privacy breach.

The Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database (Saeed et al., 2002; Saeed et al., 2011; Lee et al., 2011) is an example of a success story in the clinical domain. In addition to applying a high-performing automatic de-identification method, the creators of MIMIC have drawn a data use agreement that requires the users to be informed about the sensitive nature of the data, and to contribute to privacy protection, should they identify any potential breach. To our knowledge, this is the only clinical database of this scale available for clinical and Natural Language Processing (NLP) research in English or in any other language. Smaller de-identified clinical datasets have also been released in conditions similar to MIMIC in the context of international NLP challenges, such as i2b2 with a variety of goals, including the evaluation of de-identification methods (Uzuner et al., 2007).

We believe that studies assessing the possibility of privacy breach on realistically de-identified data can lead to a better understanding of the risk benefit balance for dataset release. In addition, such studies can contribute to building confidence in de-identification systems and methods that are otherwise evaluated quantitatively.

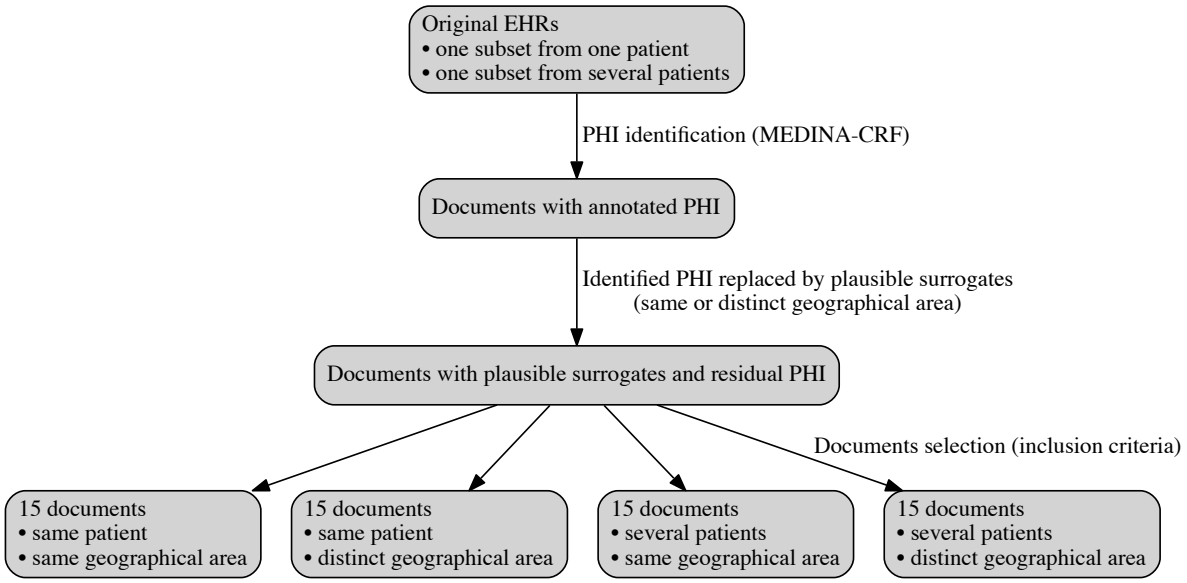

Figure 1: Production of corpora used in this study

3 Material and methods

3.1 Corpus preparation

The corpus used in this study was approved by the French administrative authority on data privacy[†] for research on Information Retrieval (IR) in large Electronic Health Records.

Twelve types of PHIs pertaining to patients, patient relatives and health professionals were targeted in our study: first names, last names, initials, addresses, cities, countries, zip codes, telephone and fax numbers, email addresses, hospital names, identifiers (such as social security numbers or medical device serial number) and dates (including patient date of birth).

In this study, we selected documents among the three most frequent types in the corpus: discharge summaries, correspondance and procedure or consult report.

We assess the chances of re-identification on a worst-case scenario, using a high-performing automatic de-identification method (Medina-CRF, see details below) on a corpus comprising 60 documents that will likely cause the system to fail identifying some PHI.

Rule-based criteria for finding files that we anticipate to be "hard to de-identify" for the automatic tool were compiled based on an error analysis.

[†]CNIL - Commission nationale de l'informatique et des libertés `http:www.cnil.fr`

They include:

Name criteria The tool often fails to identify complex names or part of complex names that include a hyphen or space (e.g., Dorothy Jane, Watterman-Smith).

Contact information criteria The tool often fails to identify contact information that appears in the content section of document (i.e., outside of header/footer sections), even when introduced by trigger words (e.g., "domicilié" *residing at*, "personne de confiance" *support person*)

Date criteria The tool often erroneously marks dates that are not linked to the patient record, e.g., dates of legal procedures quoted in the patient record. Marking these dates for replacement can compromise the confidentiality of the other dates in the file or record, because marked dates are shifted by a random number of days at the step of surrogate re-introduction.

The Medina-CRF de-identification tool for French clinical documents was designed by one of the authors (Developer 2). It is a statistical tool that was trained on a corpus of 100 gold-standard documents (Grouin and Névéol, 2014). The automatically tagged PHIs are replaced by plausible surrogates in order to create a de-identified corpus where PHIs may or may not pertain to the original documents.

The idea behind surrogate introduction is to apply the "hiding in plain sight" principle, with the

hypothesis that original PHIs will be less conspicuous in the corpus among surrogate PHIs. The original version of the surrogate replacement module was developed by one of the authors (Developer 2) and then extended by another author (Developer 1).

We assess the possibility of re-identification in different situations, relevant to clinical NLP research using de-identified records. Depending on the aim of a study, it may be necessary to use a corpus comprising documents pertaining to medical record of the same patient (e.g., patient timeline analysis) or documents pertaining to different patients (e.g., concept identification).

Our hypothesis is that re-identification might be more difficult for a corpus of documents pertaining to random patients (vs. same patients), as a corpus of documents from the same patient provides more information about a unique patient and also offers the possibility to cross-reference information between documents.

We also assess the possibility of re-identification with different settings of the surrogate re-introduction tool. Our surrogate re-introduction method relies on lists of surrogates for each type of PHI that can be marked by the de-identification tool. A setting of the tool allows the user to select a geographical area (at the level of French departments, equivalent to U.S. states) for the re-introduction of surrogates for cities, zip codes and hospitals.

We experimented with two settings of the tool, one where the geographical area of surrogates was the same as that of original PHIs, one where the geographical area was different.

The corpus was divided into four sections to study the variations in medical purpose and surrogate setting (see Figure 1):

1. 15 documents pertaining to the same patient with surrogate reintroduction from the same geographical area
2. 15 documents pertaining to random patients with surrogate reintroduction from the same geographical area
3. 15 documents pertaining to the same patient with surrogate reintroduction from a different geographical area
4. 15 documents pertaining to random patients with surrogate reintroduction from a different geographical area

For files collected from the same patient, we selected a random record from all records with at least one file meeting the "hard to de-identify" criteria. Then, we selected one random file meeting the "hard to de-identify" criteria for this record, and then selected other files randomly from the record.

For files collected from random patients we first selected three files meeting each type of criteria, and then selected other files randomly.

The file selection method was automatic, so that the authors who designed the method and participated in the annotation (Developers 1 and 2) knew of the selection criteria, but did not have prior knowledge of why a particular file in the corpus had been selected.

3.2 Gold standard set development

Two sets of gold standard annotations were created for the study corpus of 60 documents.

One gold standard set comprises annotations of all PHIs on the original corpus (see gold standard 1 in Figure 2). It was obtained by revising the original corpus with automatically marked PHIs. This gold standard is used to determine the performance of the de-identification tool on the study corpus.

Another gold standard set comprises annotations of the original PHIs that were not replaced by surrogates (see gold standard 2 in Figure 2). This was obtained by comparing the study corpus (after surrogates were re-introduced) to the original corpus (with automatically marked PHIs). One annotator prepared this gold standard corpus after they had produced their annotations on the original corpus, in an experimental setting similar to that of other annotators. This gold standard set is used to determine the number of "unmarked" PHIs in the study corpus, and to compute the performance of annotators to identify original PHIs.

3.3 Re-identification experiment

The corpus was shown to individuals with varying knowledge of the documents and corpus processing method: one clinician practicing in the hospital that supplied the corpus for this study, two informatics researchers who designed and developed Medina (the de-identification tool and surrogate re-introduction tool), and three other researchers without specific knowledge of the corpus or de-identification methods. Each individual was asked to mark PHIs that they believed to be

original, i.e., that may reveal information about the patients. The annotations were made using the BRAT rapid annotation tool (Stenetorp et al., 2012).

The annotators that were not familiar with either the corpus or the de-identification method (Researchers 1-3) were told briefly that clinical documents were processed automatically to replace twelve types of PHIs. They were given the specific list of PHIs, which was encoded in BRAT as categories available for creating annotations. They were told that the automatic system was not perfect, and that some of the PHIs present in the documents might be original PHIs, that they had to try and identify.

The annotators were told that the four sections of the corpus corresponded to a document selection from either the same or random patients. Researchers 1-3 were told that the geographical setting of surrogate re-introduction varied between corpus sections. However, they were not told that one setting was the original geographical location, while the other was not.

After the annotators had worked on the documents, they were asked to provide any specific information on any patient that they believed to have re-identified in the course of the study. They could use any tool at their disposal to attempt re-identifying the patients. In practice, the tools used included: a generic search engine, an online reverse look-up directory service and a hospital health information system.

For each individual, we computed the performance of identifying original PHIs as well as inter-annotator agreement (IAA) with other individuals in terms of F-measure. Performance of PHI identification and IAA were assessed both overall for the entire corpus as well as for each of the four sub-sections.

4 Results

Table 1 shows the distribution of PHIs in the corpus. About 10.0% were original PHIs, while 90.0% were re-introduced surrogates.

Table 2 shows the detailed performance of the automatic de-identification tool, for exact matches. The overall performance on the corpus was 0.93 F-measure, with 0.96 precision and 0.90 recall, which can be considered state-of-the-art.

Figure 2 shows an excerpt of a sample corpus document. This document was selected as "hard

PHI type	Total	Unmarked
Last name	541	18 (3.3%)
First name	487	17 (3.5%)
Initials	39	35 (89.7%)
Address	60	21 (51.7%)
City	153	39 (25.5%)
Zip Code	67	12 (17.9%)
Phone	282	0 (0.0%)
Email	42	0 (0.0%)
Identifier	20	16 (80.0%)
Date	233	17 (7.3%)
Hospital	166	24 (14.5%)

Table 1: Distribution of total and unmarked PHIs in the final corpus

Category	Precision	Recall	F-measure
Last name	0.97	0.95	0.96
First name	0.98	0.96	0.97
Initials	0.67	0.05	0.09
Identifier	1.00	0.25	0.40
Hospital	0.74	0.53	0.62
Address	0.98	0.82	0.89
Zip code	1.00	0.79	0.88
City/Country	0.99	0.95	0.97
Date	0.94	0.97	0.96
E-mail	1.00	1.00	1.00
Telephone	0.99	1.00	0.99
Overall	0.96	0.90	0.93

Table 2: Performance of Medina-CRF on the study corpus

to identify" per our contact information criteria as it contains the trigger word "personne de confiance" *support person*, along with a contact phone number for the patient's spouse. While this particular PHI was correctly identified and substituted by the automatic system, additional information about the patient's family was not. Documents are shown to annotators without any markings (processed text). On the gold standard 2 section of the figure, surrogate PHIs are shown in italic font, and original PHIs (that were not substituted by the automatic processing) are underlined. In this example, the original PHIs were the residence location of the patient's children - the passage reports "Marital Status: married. 3 Children (2 in Marseille, 1 in Corse)." For this particular document, two annotators (Developer 1 and 2) correctly identified that the two original PHIs were, indeed, original. One annotator (Researcher 1) identified that

original text	**gold standard 1**
Mary Smith née le 05/08/1928 Mariée, 3 enfants (2 à Marseille et 1 en Corse) Profession: sans profession ... Personne de confiance: époux Tél: 06 41 69 31 72 ... Pathologie pancréatique en 1993 ... Dr. Daniel Lucas, Médecin attaché.	Mary Smith née le 05/08/1928 Mariée, 3 enfants (2 à Marseille et 1 en Corse) Profession: sans profession ... Personne de confiance: époux Tél: 06 41 69 31 72 ... Pathologie pancréatique en 1993 ... Dr. Daniel Lucas, Médecin attaché.
processed text, shown to annotators	**gold standard 2**
Jane Doe née le 04/07/1927 Mariée, 3 enfants (2 à Marseille et 1 en Corse) Profession: sans profession ... Personne de confiance: époux Tél: 06 02 41 57 15 ... Pathologie pancréatique en 1992 ... Dr. Gregory House, Médecin attaché.	*Jane Doe* née le *04/07/1927* Mariée, 3 enfants (2 à Marseille et 1 en Corse) Profession: sans profession ... Personne de confiance: époux Tél: *06 02 41 57 15* ... Pathologie pancréatique en *1992* ... Dr. *Gregory House*, Médecin attaché.

Figure 2: Sample corpus document. Original PHIs (annotated in the gold standard corpora) are underlined. For illustration purposes on this figure, surrogate PHIs are shown in *italic font*.

the country PHI "Corse" was original. Relying on their knowledge of the "hard to identify" criteria, Developer 1 also marked the phone number "06 02 41 57 15" as original PHI, when it was in fact a surrogate.

On average, the annotators each spent 2 hours working on the corpus to produce the annotations.

Table 3 presents the performance of PHI identification by annotator, ordered by prior knowledge of data and method; we can classify them into three groups, represented by double bars: advanced knowledge of both documents and method, advanced knowledge of either documents or method, little knowledge of either documents or method. The table presents results for each of the four sections of the corpus (lines 2 to 5) as well as overall (line 6).

Table 4 presents the inter-annotator agreement for PHI identification.

Patient re-identification using the generic search engine and online reverse look-up directory service was unsuccessful. However, two patients could be re-identified using the hospital health information system.

5 Discussion

Performance of original PHI identification Table 3 shows that overall, PHI recognition is low. It suggests that the ability to identify original PHIs is associated with prior knowledge of the documents and/or corpus de-identification method. The highest PHI recognition is 0.50, which is not very high performance. Researchers 1-3 had no prior knowledge of either the method or the documents. After the experiment, Researcher 1 correctly identified the hospital that supplied the documents. No individual was able to supply more specific information about any of the patients based on the corpus alone.

Table 4 shows that the higher inter-annotator agreement was observed between the annotators with the highest performance for PHI recognition, Developer 1 and Clinician. Nonetheless, agreement was only 0.33, which is considered very low (Artstein and Poesio, 2008). This indicates that the "hiding in plain sight" strategy is working well, and that the original PHIs are not obvious to the annotators.

Corpus	1	2	3	4	Overall	
Dev1	34	35	31	42	142	n
	0.71	0.57	0.61	0.67	0.62	P
	0.33	0.54	0.40	0.50	0.50	R
	0.45	0.56	0.49	0.57	0.57	F
Clin	13	11	19	25	68	n
	0.62	0.64	0.47	0.76	0.61	P
	0.11	0.18	0.19	0.34	0.20	R
	0.19	0.29	0.27	0.47	0.30	F
Dev2	285	59	28	41	413	n
	0.16	0.19	0.71	0.51	0.23	P
	0.64	0.30	0.43	0.38	0.46	R
	0.26	0.23	0.53	0.43	0.30	F
Res1	30	8	6	15	59	n
	0.47	0.50	0.33	0.80	0.54	P
	0.19	0.11	0.04	0.21	0.15	R
	0.27	0.18	0.08	0.34	0.23	F
Res2	0	66	0	43	109	n
	0.00	0.02	0.00	0.07	0.04	P
	0.00	0.03	0.00	0.05	0.02	R
	0.00	0.02	0.00	0.06	0.03	F
Res3	26	24	26	10	86	n
	0.00	0.00	0.00	0.00	0.00	P
	0.00	0.00	0.00	0.00	0.00	R
	0.00	0.00	0.00	0.00	0.00	F

Table 3: Performance of PHI identification in terms of number of PHIs annotated (n), Precision (P), Recall (R) and F-measure (F). Clin=Clinician, Dev=Developer, Res=Researcher. The corpus subsets are listed as per the description in section 3.1: 1=same patient, same location; 2=random patients, same location; 3=same patient, different location; 4=random patients, different location

	Dev1	**Clin**	**Dev2**	**Res1**	**Res2**
Clin	0.32	–			
Dev2	0.21	0.10	–		
Res1	0.21	0.11	0.18	–	
Res2	0.00	0.00	0.00	0.01	–
Res3	0.01	0.01	0.03	0.01	0.00

Table 4: Inter-Annotator Agreement in terms of F-measure (Clin=Clinician, Dev=Developer, Res=Researcher)

Methods for re-identification attempts The tools available to the annotators to attempt re-identifying the patients mainly consisted of information publicly available over the internet.

One annotator (Researcher 1) systematically checked hospital names, person names and locations using a generic search engine, and was able to identify the hospital that the patients were treated in. Two annotators (Developer 1 and Clinician) used an online reverse look-up directory service for all phone numbers and addresses that they believed might be original PHIs; however, the queries did not return any results.

One annotator (Clinician) had access to the hospital Health Information System (HIS). They reported that the information they were able to recover about any single patient in the random sets was likely not sufficient to submit a valid query into the system[‡]. However, for patients with several documents available, recouping an approximate stay date with diagnostic codes such as ICD or CPT either found directly in a document (these codes are not considered PHIs) or inferred using medical coding knowledge, they were able to pull the patient record from the HIS and therefore identify the patients. For the two patients in our study corpus (corpus subsets 1 and 3) the re-identification required several attempts at querying the HIS and took 20 minutes for one patient, 30 minutes for the other.

Other annotators did not report using any re-identification strategies that relied on data sources outside the documents themselves.

The most powerful tool used is definitely the hospital HIS, by an individual with access credentials. The system search services are set-up in a way that requires the users to provide sufficient information about the patients before a record can be retrieved. In our study, it did not help with patient re-identification when there was only one document available about the patient.

However, when there were several documents available about the same patient, the patient could be identified in 30 minutes or less. In this case, patient identification required (i) access to the HIS; (ii) knowledge of how records are coded and retrieved in the HIS; and (iii) medical knowledge to identify or infer diagnostic codes from the patient documents.

[‡]In practice, re-identification was attempted for a few documents that the Clinician thought had re-identification potential. Attempts to re-identify a patient were stopped after 30 minutes when they were not successful. After a few failures, the clinician focused his efforts on the patients for which more than one document were available.

Perfomance variation according to medical purpose and surrogate setting We can see from table 3 that overall, the performance of PHI identification seems to be higher when the geographical area of re-introduced surrogate is different from that of the original PHIs (row 4 vs. row 2, row 5 vs. row 3). This indicates that the "hiding in plain sight effect" is stronger when location surrogates come from the same geographical area as the original PHIs.

The initial argument against using same-area location surrogates was that, as surrogates are re-introduced randomly, a surrogate could happen to be the same as the original PHI, therefore voiding the replacement operation. We did identify a few occurrences of this phenomenon in our data set, when the surrogate and original PHI were in fact different strings (e.g., "Bois-Guillaume" vs. "BOISGUILLAUME").

The PHI identification results from table 3 do not clearly indicate that PHI identification was made easier in the corpus subsets of documents from the same patients vs. random patients (line 2 vs. line 3 and line 4 vs. line 5). However, when the hospital HIS was available, patients with multiple documents available could be identified while patients with only document available could not.

Implications for the design of a clinical corpus to be used in an NLP challenge or shared-task The results of our study suggest that re-identification attempts from researchers without privileged access to the hospital health information system (which is expected to be the case of most individuals accessing a corpus through a challenge) will not be successful.

It is also important to point out that the identification of the patient identities in this study were only possible because the de-identification was performed automatically and some original PHIs (dates) could be found in the documents.

In the context of data release for a challenge or shared-task, the de-identification process should include multiple rounds of manual review of PHI to ensure that no original PHIs were left.

In summary, this study suggests that patient privacy can be reasonably preserved in a corpus comprising documents pertaining to random patients, with same-area geographical surrogate re-introduction and manually reviewed de-identification.

Limitations The main limitation in this study is the size of the corpus, which comprises 60 documents. This size was chosen to keep the annotation time manageable. It is comparable to the size of the corpus (85 documents) used previously by Meystre et al. (2014a). The study of variations leads us to partition the corpus into four subsets of 15 documents, which can only provide indicative results. The study will need to be reproduced on a larger scale.

Also, one important category of individual likely to identify patients from the content of de-identified files includes patients themselves, or patients' relatives and acquaintances. For instance, an individual who personally knows the patient that our sample file pertains to (see Figure 2) might read this document and realize that the information (stay at home mother of 3 children who experienced a pancreas disorder in the past) matches the circumstances of their acquaintance. However, we have not been able to devise an adequate experimental setting to evaluate this chance. Arguably, the chance might be similar to that of patient re-identification by a doctor who had personally attended to the patient within the past three months. It was found that doctors were not able to re-identify their own patients from de-identified documents (Meystre et al., 2014a).

6 Conclusion

In spite of shortcomings of the de-identification system identified by the developpers in a thorough error-analysis, patient privacy was not compromised by individuals without privileged access to the relevant hospital health information system.

When access to the hospital health information system is available, patients can be re-identified by recouping information found in more than one document, and medical knowledge of medical coding. However, patient privacy is preserved when only one document per patient is available.

Furthermore, less information can be recovered when location surrogates for the same geographical area as the original files are used.

Acknowledgement

This work was supported by the French National Agency for Research under grant CABeRneT[§] ANR-13-JS02-0009-01. The authors thank the

[§]CABeRneT: Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle

Biomedical Informatics Department at the Rouen University Hospital for providing access to the LERUDI corpus for this work, and the annotators who kindly participated in this study.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–96.

Michael Barbaro and Tom Zeller Jr. 2006. A face is exposed for aol searcher no. 4417749. *The New York Times*, August 9.

Kathleen Benitez and Bradley Malin. 2010. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc*, 17(2):169–77.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2013. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *J Am Med Inform Assoc*, 20(2):342–8, Mar-Apr.

Louise Deleger, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li, Keith Marsolo, Anil Jegga, Megan Kaiser, Laura Stoutenborough, and Imre Solti. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *J Am Med Inform Assoc*, 20(1):84–94.

Cyril Grouin and Aurélie Névéol. 2014. De-identification of clinical notes in french: towards a protocol for reference corpus developpement. *J Biomed Inform*, 46(3):506–515, Aug.

Joon Lee, Daniel J Scott, Mauricio Villarroel, Gari D Clifford, Mohammed Saeed, and Roger G Mark. 2011. Open-access MIMIC-II database for intensive care research. In *Proc IEEE Eng Med Biol Soc*, pages 8315–8.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol*, 10(70).

Stephane Meystre, Shuying Shen, Deborah Hofmann, and Adi Gundlapalli. 2014a. Can physicians recognize their own patients in de-identified notes? In *Stud Health Technol Inform*, volume 205, pages 778–82.

Stephane M Meystre, Oscar Ferrández, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2014b. Text de-identification for privacy protection: a study of its impact on clinical text information content. *J Biomed Inform*, 50:142–50, Aug.

Ishna Neamatullah, Margaret M Douglass, Li-Wei H Lehman, Andrew Reisner, Mauricio Villaroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak*, 8(32).

Mohammed Saeed, Christine Lieu, Greg Raber, and Roger G Mark. 2002. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput Cardiol*, 29:641–4.

Mohammed Saeed, Mauricio Villaroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. 2011. Multiparameter intelligent monitoring in intensive care ii (MIMIC-II): A public-access intensive care unit database. *Crit Care Med*, 39(5):952–60.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Juni'chi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proc of EACL Demonstrations*, pages 102–7, Avignon, France. ACL.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. In *AMIA Annu Fall Symp Proc*, pages 333–7, Washington, DC.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, 14(5):550–63.