MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190185

Generation of Surrogates for De-Identification of Electronic Health Records

Aipeng Chen[a], Jitendra Jonnagaddala[b,c], Chandini Nekkantti[d] , Siaw-Teng Liaw[b,c]

[a] *Prince of Wales Clinical School, UNSW Sydney, Australia*
[b] *School of Public Health and Community Medicine, UNSW Sydney, Australia.*
[c] *WHO Collaborating Centre for eHealth, UNSW Sydney, Australia*
[d] *CGD Health Pty Ltd, Australia*

Abstract

Unstructured electronic health records are valuable resources for research. Before they are shared with researchers, protected health information needs to be removed from these unstructured documents to protect patient privacy. The main steps involved in removing protected health information are accurately identifying sensitive information in the documents and removing the identified information. To keep the documents as realistic as possible, the step of omitting sensitive information is often followed by replacement of identified sensitive information with surrogates. In this study, we present an algorithm to generate surrogates for unstructured electronic health records. We used this algorithm to generate realistic surrogates on a Health Science Alliance corpus, which is constructed specifically for the use of development of automated de-identification systems.

Keywords:

Algorithms, data anonymization, electronic health records

Introduction

Unstructured electronic health records (EHR) such as discharge summaries, encounter notes, pathology reports and radiology are valuable source of information for basic, clinical and translational researchers [1-4]. These documents are often shared with researchers for clinical and biomedical research purposes. However, these documents may contain sensitive protected health information (PHI) such as patient name and unique patient identifiers. When private information is removed, patients are willing to share their medical records for research use [5]. To protect the privacy of patients, researchers need to share or access these documents in de-identified manner. Traditionally, the PHI was manually identified and removed before they are shared. However, with large number of documents, the manual process is tedious and not feasible to scale. In a study by Dorr et al [6], it was reported that average time required to manually de-identify one document (7.9 types of PHI per document) is 87.3 seconds. Various studies have shown that it is possible to automatically de-identify unstructured EHRs with acceptable accuracy [7-10].

Automated de-identification can be employed to replace traditional manual process. The process often consists of two main stages: identifying PHI and replacing identified PHI with surrogate information. Surrogates are realistic replacements of PHI that are close to real PHI but do not contain any private information. There are a few challenges in surrogate generation.

First, surrogates are supposed to be as realistic as possible. Since the documents are mostly intended to be shared for research, readability of the text need to be maintained. Researchers need to understand the text and ineffective surrogates can be distracting. Additionally, natural language processing systems also need the documents to be realistic so that the documents can be used to train the models to detect PHI in unseen documents. Second, the context of the document needs to be preserved. Some words, especially names and locations, can appear in different forms in one piece of text. For example, a name 'John Smith' can appear in the text as 'Smith, John', 'JS', 'Mr. Smith' or just 'Smith', they refer to one person and failing to maintain these complexities can lead to biased training during the development of de-identification systems.

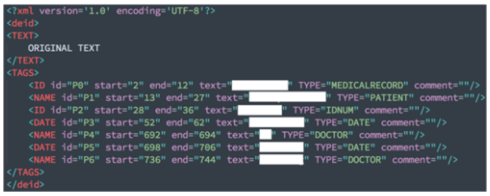
Third, the temporal information also needs to be preserved well so that the order of occurrence of the clinical events is maintained. Lastly, PHI can carry not only private information, but also information that could be useful in research. For example, a person's name could imply one's gender, which might be useful when no structured gender information is not available.

Some datasets use placeholders or other forms of obfuscation to remove PHI [11, 12]. Stubbs & Uzuner [13] discussed the challenges of generating realistic surrogates and described their algorithm to generate surrogates. Multiple strategies were applied to PHI of different categories in doing this algorithm. Alphabet and date shift were introduced to maintain the consistency of the context in the document. However, these methods were specific to the US and not directly translatable to Australian setting. As a result, the readability of the surrogated documents is low, and the performance of automated de-identification systems trained using this surrogated data might not perform well on Australian EHR. In this study, we improved previous strategies for surrogate generation to make it more relevant to Australian setting. Several matching strategies for different kinds of PHI were developed to match PHI with same meaning expressed in different formats. We have used the Health Science Alliance (HSA) biobank de-identification corpus. HSA biobank is an institutional biobank at UNSW Sydney for translational research.

Methods

HSA Biobank Corpus

We constructed a large corpus of pathology reports that was annotated specifically for the use of development of automated de-identification systems for unstructured EHRs. The corpus consisted of 2100 pathology reports from 1833 patients. There were 38414 pieces of PHI identified in the corpus. Most of the PHI was tagged as names, locations, dates and IDs. Only few of the PHI were related to PHI categories: contact and age. There was no category of profession or other information observed in the HSA corpus. EHRs and their annotations were stored in the format of XML files as shown in the Figure 1.



```
<?xml version='1.0' encoding='UTF-8'?>
<deid>
<TEXT>
   ORIGINAL TEXT
</TEXT>
<TAGS>
   <ID id="P0" start="2" end="12" text="[redacted]" TYPE="MEDICALRECORD" comment=""/>
   <NAME id="P1" start="13" end="27" text="[redacted]" TYPE="PATIENT" comment=""/>
   <ID id="P2" start="28" end="36" text="[redacted]" TYPE="IDNUM" comment=""/>
   <DATE id="P3" start="52" end="62" text="[redacted]" TYPE="DATE" comment=""/>
   <NAME id="P4" start="692" end="694" text="[redacted]" TYPE="DOCTOR" comment=""/>
   <DATE id="P5" start="698" end="706" text="[redacted]" TYPE="DATE" comment=""/>
   <NAME id="P6" start="736" end="744" text="[redacted]" TYPE="DOCTOR" comment=""/>
</TAGS>
</deid>
```

Figure 1— Annotated Sample Document from the HAS Corpus in XML Format

PHI Categories

The definition of PHI categories was the same as it was in the i2b2 2014 de-identification shared task [14]. This in turn was developed based on the HIPAA (Health Insurance Portability and Accountability Act) established by the USA which defines 18 categories of PHI. The guideline expanded the original 18 categories to include more information. All patients' ages were included despite that only ages above 89 years were considered PHI in the HIPAA. These categories had been grouped into 8 main categories and 25 sub-categories. Detailed categories and sub-categories along with examples of each categories are presented in Table 1.

Table 1 – Detailed PHI Categories and Sub-Categories

PHI category	Sub-category	Example
Name	patient, doctor, username	John Doe, Dr. Max, Mr. Smith
Profession	none	lawyer, teacher
Location	room, department, hospital, organization, street, city, state, country, zip, other	peri-operative unit-pow, macquarie ward – rhw, 12 abc street
Age	none	23, 98
Date	none	24/12/1987, September 26th
Contact	phone, fax, email, url, ipaddress	+61-421123456 abc@gmail.com 194.223.1.1
IDs	social security number, medical record number, health plan number, account number, license number, vehicle id, device id, biometric id, id number	mrn: 9174338 id number: 12r1500257
Other	none	finger print, company logo

HSA Biobank Surrogate Algorithm

Names were collected from the Internet for the use of surrogate generation so that the surrogates can be closer to reality. Names are stored in separate files by categories. Names collected were names of individuals and locations. Names included first names and surnames. Location names included Australian states, cities, streets, organizations and hospital names. These files were loaded into the memory as lists and were later used to generate surrogates. The names and location information were specific to Australia. When a PHI entity was not found in the existing constructed surrogates map, PHI was considered as first-time occurrence, and a new surrogate was generated in various ways according to the PHI's category. In many cases, such as IDs, phone numbers, URLs and emails, it was easier to generate surrogates since they were merely combination of strings of digits and letters and sometimes some special characters such as commas, periods, parentheses. We simply replace them with randomly generated strings with the original format kept. There were some other PHI categories such as individual names, locations names lists were used so that the surrogates can appear more realistic. We formulated some rules to pick surrogates from these lists for some categories to improve the performance, which is discussed in later sections. The HSA biobank surrogate algorithm is available at https://github.com/TCRNBioinformatics/PHISurrogates/.

Name and Location

Surrogate generation for names turned out to be the most challenging part. Mapping names that could appear in different forms and preserving as much information as possible were the key challenges.

A name can appear in many different forms in one given document. For example, if a comma existed in the name, we considered the name included both first and last name. Accordingly, the algorithm removed the comma and replaced original information with surrogate names from the lists we constructed. After this replacement, names could appear as a full name, or combination of initials, or a combination of the initial of one name and another name as full. In the algorithm, we took the first two letters from both parts of the name as the key in the map. For example, "JOSM" in the case of "John Smith". If an abbreviation was observed in the text, a space character was used. For example, "J(space)SM" for "J Smith" and "JO (space)(space)" for "John". The reason for not using only the first initial letter was that different names might have identical initials and it could cause mistake in mapping. For example, "John Smith" and "Jack Scott" had the same initials "JS". On the other hand, "John" and "Johnny" could stand for the same person if "Johnny" was used as a nickname for "John". So, taking the whole name as a key could also cause mismatching in the map. With the two initials as key, a simple rule can be applied: if more than half of the letters in the key are same, they are considered as same name. For example, "JOSM", "J(space)S(space)", "JOS(space)", "J(SPACE)SM" and "J(space)S(space)" were all considered one name. As for the names as initials, we did not look up in the map but proceeded to the surrogate generation directly as the algorithm was able to map a full name's initials to the initials of its possible surrogate with the method describe above. For example, "JS" can be directly mapped to "LU" without looking up in the map or the name table so it is faster. The titles (i.e., "Dr.", "Mr.", "Miss.", "Mrs.") were extracted out initially. These titles were then added back to the surrogate names generated later.

In order to preserve context, two rules were applied when generating surrogates. Firstly, to preserve the gender information, we constructed and split the names dictionary into three types; male first names, female first names, and surnames. A first name was first searched in both male and female first names list to determine its gender. For those that belonged to both, or not found in both, a random gender was given. Then a surrogate was picked from the dictionary according to the name's gender. Then alphabet shift of fixed length was applied. With this shift, the initial letter of a name was mapped to another letter in the alphabet, then a name starting with this letter was picked from the names dictionary lists as surrogate. We maintained various dictionaries for different sub-categories of locations, including countries, states, cities, streets, hospitals and organizations. Surrogates of locations were randomly picked from dictionaries accordingly.

Date and Age

A randomly generated date shift in the range of 1 to 730 days was applied. Dates and ages in records from same patient were shifted with the same length to maintain temporality. Similar to the names PHI category, dates could appear in various forms. For example, '18/12/2017' could appear as '18.12.2017' or '2017-12-18'. Similarly, a date '18/12' could be written as "December 18" or "Dec 18". Another challenge was the ambiguity of date strings caused by different date notation styles. For example, a date string "03/04/05" could possibly in the format of "DD/MM/YY" or "YY/MM/DD" or "MM/DD/YY". Since all the reports were retrieved from Australian hospital, we considered all the date strings are of either "DD/MM/YY" or "DD/MM/YYYY" format. We normalized all the non-standard date variations into ISO-8061 standard, "YYYY-MM-DD". The parsed dates were then shifted by adding a time shift in the format. As for the age information, applying a date shift was relatively easier. The date shift in days was converted into years and added.

ID and Contact

Surrogates for IDs were relatively easier to generate. For BIOID and IDNUM, surrogates started with two digits from 10-99, followed by an "N" and an "R" respectively, then seven digits from 1000000-9999999, as similar format was observed in our documents. Medical record numbers with seven digits varied between 1000000 and 9999999, followed by three alphabets. As for fax and phones, all surrogate numbers were generated in the Australian format: two digits from 01-08, followed by seven digits from 0000000 to 9999999. For emails, the format was a random string of length 32 plus "@gmail.com". Surrogates found from the list map or newly generated information was then used to replace the original PHI.

Results

We presented an algorithm for surrogate generation for de-identification of unstructured EHRs. This algorithm can be used either during the de-identification corpus construction or during the development of automated de-identification systems for unstructured EHRs. We tested our algorithm on 2100 annotated pathology reports from the HSA corpus. The algorithm took a total time of 2.94 seconds to process the 2100 documents and 0.0014 seconds per document on average. There was 38,414 pieces of annotated PHI in the corpus, 18.29 in each file on average. Most of the surrogates generated were names and locations (Table 2). Among all PHI entities in the HSA corpus, a total number of 1085 pieces of PHI were replaced with the previously generated surrogates found in the existing map before a new surrogate was generated (Table 2). Though it is not feasible to validate every surrogate generated, two authors have manually verified 5% of the documents from the HSA corpus to assess the readability and contextual information of the documents after surrogate generation. Our algorithm has generated surrogates as intended but few issues were observed, as we discuss in the next section.

Table 2— Count of Surrogates Generated by Categories and Found in the Lists Developed

Category	Count of surrogates	Count of surrogates from lists developed
Name	11789	284
Age	141	0
Contact	7	0
Location	9861	104
Date	7665	321
ID	8951	376
Professions	0	0
Other	0	0
Total No. of documents	2100	1085

Discussion

We discussed in detail strategies applied to tackle the key challenges in surrogate generation. Lists of names and locations were collected and used to make the surrogates as realistic as possible. Maps of existing surrogates were used to make sure that same surrogates were used to replace the same objects so that the context could be maintained. Formats of dates, IDs and contacts were saved according to the Australian standards since the EHRs were all retrieved from Australian hospitals. For names, a key of first two letters of both part of names was designed to deal with co-reference and ambiguity. Also, a date shift was applied for the generation of dates and ages so that the progress of date in the context can be preserved. The information of names was preserved by determining the possible gender of name by searching in the real name lists.

Alphabet shift was applied on names to maintain the context. There are advantages mapping the letters with a fixed alphabet shift. We could easily replace the initials with the shifted letters without concerning about the ambiguity of initials, such as "JS" for either "John Smith" or "Jack Scott" would both be replaced by "LU" with a shift of length 2. Additionally, mistakes in, mapping initials to names could be reduced too. A consistent shift of initials could make sure that the surrogates are distributed more evenly so that the surrogated documents can have a better performance when training automated de-identification systems. It is possible that the shift of initial letters in names can be inferred according to the frequency of letters in the context. It is possible that an initial can be used to identify a person, but since there are a limited number of names that appear in the documents, and the alphabet shift is generated randomly per document, it is almost impossible to deduct the original initials from the surrogate names. Even if the alphabet shift pattern is identified, only the initials of the names can be obtained, so the risk of re-identification is still negligible.

The findings in this study are subjected to several limitations. There was no profession observed in the corpus, the surrogate generation process on professions was not applied and tested. Patients' health status might be related to their profession and a surrogate need to preserve such information so that the potential relationship could be used for research. Also, as these documents are generated by clinicians, there is a possibility of spelling errors which could impact surrogate generation process. Our algorithm doesn't consider this and as a result it is possible that a misspelt PHI entity could be replaced with a non-contextual based surrogate. Date shift was applied to dates and ages in the algorithm. However, some date information remained in the documents such as holiday names, like Christmas and New

Year's Eve. This information can be used along together with the generated surrogates to infer the date shift and therefore get the original date and age. In future, we would like to address these limitations by improving our surrogate generation algorithm in turn reducing the risk of re-identification.

Conclusion

In summary, we presented an algorithm to generate realistic contextual surrogates for unstructured EHRs de-identification systems. The algorithm can also be used to construct a corpus for the development of de-identification systems. In the algorithm, replaced the PHI with realistic surrogates in order to maintain the quality and context of the documents. Australian names and date formats were used in this study. Our findings suggest that the algorithm presented in this study is capable of processing large number of documents within few seconds. However, the documents need to have PHI information already identified. Different strategies were applied to tackle different challenges. An existing surrogate map is maintained to make sure that same PHI, or PHI with the same meaning are replaced with the same surrogate so that the context is preserved. However, professions and PHI annotated as 'other', which can possibly be an important source for research use, and a critical risk of identification leak, are not observed in our HSA corpus. This study suffers from various limitations such as failure to handle misspellings and holiday information. Future work in this area is required, especially to reduce the risk of re-identification.

Acknowledgements

This study was conducted as part of the electronic Practice Based Research Network (ePBRN) and Translational Cancer research network (TCRN) research programs. ePBRN is funded in part by the School of Public Health & Community Medicine, Ingham Institute for Applied Medical Research, UNSW Medicine and South West Sydney Local Health District. TCRN is funded by Cancer Institute of New South Wales and Prince of Wales Clinical School, UNSW Medicine.

References

[1] Berman, J.J., Concept-match medical data scrubbing: how pathology text can be used in research. Archives of pathology & laboratory medicine, 2003. 127(6): p. 680-686.

[2] Pitson, G., et al., Developing a Manually Annotated Corpus of Clinical Letters for Breast Cancer Patients on Routine Follow-Up. Studies in health technology and informatics, 2017. 235: p. 196-200.

[3] Stubbs, A. and Ö. Uzuner, Annotating risk factors for heart disease in clinical narratives for diabetic patients. Journal of Biomedical Informatics, 2015. 58: p. S78-S91.

[4] Jha, A.K., The promise of electronic records: around the corner or down the road? Jama, 2011. 306(8): p. 880-1.

[5] King, T., L. Brankovic, and P. Gillard, Perspectives of Australian adults about protecting the privacy of their health information in statistical databases. International Journal of Medical Informatics, 2012. 81(4): p. 279-289.

[6] Dorr, D.A., et al., Assessing the difficulty and time cost of de-identification in clinical narratives. 2006. 45(03): p. 246-252.

[7] Golle, P. Revisiting the uniqueness of simple demographics in the US population. in Proceedings of the 5th ACM workshop on Privacy in electronic society. 2006. ACM.

[8] El Emam, K., et al., The re-identification risk of Canadians from longitudinal demographics. 2011. 11(1): p. 46.

[9] Institute, D.L.A.N.U.A.P.H.C.R., Contentious crop: harvesting information from electronic health records. 2005.

[10] Carrell, D., et al., Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. 2012. 20(2): p. 342-348.

[11] Kushida, C.A., et al., Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. 2012: p. S82-S101.

[12] Chakaravarthy, V.T., et al. Efficient techniques for document sanitization. in Proceedings of the 17th ACM conference on Information and knowledge management. 2008. ACM.

[13] Stubbs, A., et al., Challenges in Synthesizing Surrogate PHI in Narrative EMRs, in Medical Data Privacy Handbook. 2015, Springer. p. 717-735.

[14] Stubbs, A. and Ö. Uzuner, Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of Biomedical Informatics, 2015. 58, Supplement: p. S20-S29.

Address for correspondence

Dr. Jitendra Jonnagaddala; z3339253@unsw.edu.au