e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-778

Can Physicians Recognize Their Own Patients in De-identified Notes?

Stéphane MEYSTRE[a,b,1] Shuying SHEN[a,b], Deborah HOFMANN[b], and Adi GUNDLAPALLI[a,b]
[a] *Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA*
[b] *VA Salt Lake City Health Care System, Salt Lake City, Utah, USA*

Abstract. The adoption of Electronic Health Records is growing at a fast pace, and this growth results in very large quantities of patient clinical information becoming available in electronic format, with tremendous potentials, but also equally growing concern for patient confidentiality breaches. De-identification of patient information has been proposed as a solution to both facilitate secondary uses of clinical information, and protect patient information confidentiality. Automated approaches based on Natural Language Processing have been implemented and evaluated, allowing for much faster text de-identification than manual approaches. A U.S. Veterans Affairs clinical text de-identification project focused on investigating the current state of the art of automatic clinical text de-identification, on developing a best-of-breed de-identification application for clinical documents, and on evaluating its impact on subsequent text uses and the risk for re-identification. To evaluate this risk, we de-identified discharge summaries from 86 patients using our 'best-of-breed' text de-identification application with resynthesis of the identifiers detected. We then asked physicians working in the ward the patients were hospitalized in if they could recognize these patients when reading the de-identified documents. Each document was examined by at least one resident and one attending physician, and with 4.65% of the documents, physicians thought they recognized the patient because of specific clinical information, but after verification, none was correctly re-identified.

Keywords. Natural Language Processing, Patient Data Confidentiality, Electronic Health Record De-identification, United States department of veterans affairs.

Introduction

The adoption of Electronic Health Record (EHR) systems is growing at a fast pace in the United States and in Europe. In the former, it reached more than 50% of physician practices and 80% of hospitals in April 2013, when only 17% of physician practices and 9% of hospitals were using an EHR in 2008. [1] This growth results in very large quantities of patient information becoming available in electronic format, with tremendous potentials, but also equally growing concern for patient confidentiality and privacy breaches.[2]

Confidentiality of the information entrusted by a patient to a healthcare provider has been a foundation of the confidence relationship established between them for centuries.[3] Breaching this confidentiality not only damages this relationship, but also

[1] Corresponding Author.

exposes the patient to financial, reputation, employment, and other identity theft disastrous consequences. In the U.S., the Health Insurance Portability and Accountability Act (HIPAA) protects the confidentiality of patient data [4] and requires informed consent of the patient and approval of the facility's Institutional Review or Ethics Board to use clinical data for research purposes, but this requirement can be waived if the data is de-identified.

The de-identification of patient information has been proposed as a solution to both facilitate secondary use of clinical information, and protect patient information confidentiality. Most clinical information found in the EHR is represented as narrative text,[5] and de-identification of text consists in the application of the HIPAA "Safe Harbor" rule, removing all protected health information (PHI). This is a tedious and costly manual endeavor,[6] and automated approaches based on Natural Language Processing (NLP) have been implemented and evaluated, allowing for much faster de-identification than manual approaches.[7] These approaches started with Sweeny's system [8] and focused on various selections of PHI, ranging from patient names only,[9] to all PHI categories defined in the Safe Harbor rule, or even everything that was not recognized as clinical information.[10]

The text de-identification process is composed of two main steps: PHI *detection*, and then PHI *removal or transformation*. The latter typically consists in replacing PHI with some tags or characters (e.g., 'Mr. Smith' becomes '<Patient_name>'), but another option is to replace PHI with synthetic but realistic substitutes (e.g., 'Mr. Smith' becomes 'Mr. Jones'). This second option is often called "PHI resynthesis" and has been experimented by our team [11] and Aberdeen et al.[12] It adds computational complexity but offers the advantage of allowing the rare instances of PHI that could have been missed to "hide in plain sight," notably improving the effectiveness of de-identification.[13]

The U.S. Veterans Healthcare Administration (VHA) Consortium for Healthcare Informatics Research (CHIR) is a multi-disciplinary group of collaborating investigators affiliated with VHA sites across the U.S. In the context of the CHIR, a de-identification project focused on investigating the current state of the art of automatic clinical text de-identification,[7,14] on developing a best-of-breed de-identification application for VHA clinical documents,[11] and on evaluating its impact on subsequent text analysis tasks [15] and the risk for re-identification of this text. When evaluated, most automatic clinical text de-identification systems allowed for a sensitivity of 88 to 99% when detecting PHI.[7,12] This means that some PHI is missed by these systems, and even if quite rare, does this matter? Would researchers using the de-identified clinical documents, or even healthcare providers who took care of the patients mentioned in the documents, be able to recognize these patients when examining the documents? All methods to assess the risk for re-identification were applied to a small number of structured and coded data only (e.g., demographics, location),[16,17] not to narrative text, and clinical documents are rich in clinical and social information that can be unique and could be used to re-identify a patient.

The following sections describe our effort to evaluate this risk for re-identification of automatically de-identified clinical documents.

1. Methods

A corpus of 86 discharge summaries was automatically de-identified with 'BoB' (our *B*est-*o*f-*B*reed automatic clinical text de-identification application) and using PHI resynthesis. This application is described in more details in other publications.[11,18] The discharge summaries were extracted from the most recent documents in the EHR of 86 randomly selected patients hospitalized at the Salt Lake City VHA Medical Center, in the Acute Medicine Department, between 1 and 3 months before the beginning of our study.

A group of 5 attending physicians, 2 chief medical residents, 1 subspecialty fellow, 7 second or third year medical residents, and 4 interns (1st year medical residents) examined our corpus of de-identified discharge summaries. These medical providers were involved in the care of patients hospitalized in the Acute Medicine Department, and details of their recent rotations in the medicine wards are listed in Table 1.

Table 1. Physicians examining the de-identified discharge summaries

Identifier	Role	Service in Acute Medicine during the past 6 months (before the interview)
P1	Chief medical resident	2 weeks
P2	1st year resident (intern)	2 weeks
P3	3rd year resident	6 weeks
P4	3rd year resident	2 weeks
P5	3rd year resident	2 weeks
P6	1st year resident (intern)	5 weeks
P7	Subspecialty Fellow	2 months
P8	1st year resident (intern)	2 weeks
P9	3rd year resident	2 weeks
P10	Attending	Full-time (6 months)
P11	Attending	Full-time (6 months)
P12	Attending	Full-time (6 months)
P13	Attending	Full-time (6 months)
P14	Attending	Full-time (6 months)
P15	2nd year resident	6 weeks
P16	3rd year resident	7-8 weeks
P17	3rd year resident	6 weeks
P18	Chief medical resident	3 months
P19	1st year resident (intern)	3 weeks

Each de-identified discharge summary was examined by at least one resident or fellow and one attending physician (sometimes more), during 188 interviews when we presented the printed de-identified document to the physician, waited for them to read it, and then asked them if they recognized the patient presented in the document. We recorded their answer and if positive, verified the accuracy of the patient identity they proposed.

2. Results

The corpus of discharge summaries was automatically de-identified and the physician interviews were then administered during a period of about 2 months.

Three physicians (a chief medical resident and two residents) thought they had recognized a patient when reading their de-identified discharge summary. Their reasons to have this impression included specific procedures, diagnoses, signs, and imaging results, as listed in Table 2 below.

For these patients, the physicians suspected that they recognized the patient because of clinical details that reminded them of their patient. They often couldn't give any identifying information about the 'recognized' patient, but when they could, we verified this information, and eventually no patient was correctly recognized. (Table 3).

Table 2. Patients 'recognized' in de-identified discharge summaries

Patient identifier	Healthcare provider	Reason(s) to have 'recognized' the patient	Patient identity confirmed?
874	Resident (P4)	Procedures (no patient name remembered)	No
874	Resident (P2)	Diagnosis (wrong patient name)	No
975	CMR (P1)	Treatment, Diagnosis, Signs (wrong patient name)	No
994	CMR (P1)	Imaging (no patient name remembered)	No
996	CMR (P1)	Treatment (wrong patient name)	No

CMR = Chief medical resident

Table 3. Patient de-identified discharge summaries identification results

Identification type	Count	Proportion of patients
Patient identity unknown	82/86	95.35%
Patient identity incorrectly recognized (or without any patient identifier)	4/86	4.65%
Patient identity correctly recognized	0/88	0%

3. Discussion

As reported above, none of the 86 automatically de-identified discharge summaries could be re-identified. A few patients had characteristics that made the physician reviewing the discharge summary suspect that they had recognized the patient, but none was correctly identified.

Considering that physicians currently working in or who recently rotated through a hospital ward would be the most likely to recognize a patient described in a de-identified discharge summary, these results are encouraging in terms of risk for re-identification of de-identified clinical documents. Even if automatic de-identification applications are unable to reach perfect accuracy, the resynthesis of the detected PHI allows it to "hide in plain sight" and greatly reduces the risk of re-identification. More interesting is the recognized richness of unique or unusual clinical details that are noted in discharge summaries; these details were the reasons given for 'recognizing' a patient in our study. These details were eventually found not to be precise and specific enough to correctly re-identify patients in this study.

This study was limited to one department in a VHA hospital in the U.S. and one clinical document type, and might not generalize easily to other similar settings. The number of de-identified documents examined was not large, but sufficiently powered to clearly give us an estimation of the risk for re-identification of automatically de-identified discharge summaries. Further studies are needed to validate and extend our results across hospital departments and document types. An important reason for re-identification is clinical information, especially when rare and unique, and larger studies with a higher likelihood of such rare and unique information would allow for a more accurate assessment of the risk for re-identification. Results of large studies would form the basis for informing institutional policy on large-scale de-identification of medical records and sharing of those records for secondary purposes.

Acknowledgements

We thank Abhisek Trivedi and Oscar Ferrandez Escamez for their help with the clinical documents de-identification. Research supported by VA HSR HIR 08-374. Views expressed here are those of the authors and do not necessarily represent those of the US Department of Veterans Affairs or the US Government. We wish to express our gratitude to the medical providers who participated in our study despite their busy schedules. The Chief Medical Residents deserve a special mention for their support.

References

[1] U.S. Department of Health and Human Services, Doctors and hospitals' use of health IT more than doubles since 2012, www.hhs.gov, (2013).

[2] J.K. Taitsman, C.M. Grimm, S. Agrawal, Protecting patient privacy and data security, *N. Engl. J. Med*, **368** (2013) 977–979.

[3] National Library of Medicine, The Hippocratic Oath, (2002).

[4] CFR Title 45 Subtitle A Part 164: Security and Privacy, GPO, 2008.

[5] A.W. Pratt, Medicine, Computers, and Linguistics, *Advanced Biomed Eng*, **3** (1973) 97–140.

[6] D.A. Dorr, W.F. Phillips, S. Phansalkar, S.A. Sims, J.F. Hurdle, Assessing the difficulty and time cost of de-identification in clinical narratives, *Methods of Information in Medicine*, **45** (2006) 246–252.

[7] S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, *BMC Med Res Methodol*, **10** (2010) 70.

[8] L. Sweeney, Replacing personally-identifying information in medical records, the Scrub system, *Proc AMIA Annu Fall Symp*, (1996) 333–337.

[9] R.K. Taira, A.A. Bui, H. Kangarloo, Identification of patient name references within medical documents using semantic selectional restrictions, *AMIA Annu Symp Proc*, (2002) 757–761.

[10] J.J. Berman, Concept-match medical data scrubbing How pathology text can be used in research, *Archives of Pathology & Laboratory Medicine*, **127** (2003) 680–686.

[11] O. Ferrandez, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore, S.M. Meystre, BoB, a best-of-breed automated text de-identification system for VHA clinical documents, *Journal of the American Medical Informatics Association*, **20** (2013) 77–83.

[12] J. Aberdeen, S. Bayer, R. Yeniterzi, B. Wellner, C. Clark, D. Hanauer, B. Malin, L. Hirschman, The MITRE Identification Scrubber Toolkit: design, training, and assessment, *Int J Med Inform*, **79** (2010) 849–859.

[13] D. Carrell, B. Malin, J. Aberdeen, S. Bayer, C. Clark, B. Wellner, L. Hirschman, Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text, *Journal of the American Medical Informatics Association*, **20** (2013) 342–348.

[14] O. Ferrandez, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore, S.M. Meystre, Evaluating current automatic de-identification methods with Veteran's health administration clinical documents, *BMC Med Res Methodol*, **12** (2012) 109.

[15] S.M. Meystre, O. Ferrandez, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Text De-Identification for Privacy Protection: a Study of its Impact on Clinical Text Information Content, *J Biomed Inform*, (2014, in press).

[16] K. Benitez, B. Malin, Evaluating re-identification risks with respect to the HIPAA privacy rule, *Journal of the American Medical Informatics Association*, **17** (2010) 169–177.

[17] K. El Emam, E. Jonker, L. Arbuckle, B. Malin, A systematic review of re-identification attacks on health data, *PLoS ONE*, **6** (2011) e28071–e28071.

[18] O. Ferrandez, B.R. South, S. Shen, S.M. Meystre, A Hybrid Stepwise Approach for De-identifying Person Names in Clinical Documents, *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012)*, Montreal, Canada, 2012: pp. 65–72.