Automatic Clinical Text De-Identification: Is It Worth It, and Could It Work for Me?

Stéphane M. Meystre, MD, PhD[1] , Hercules Dalianis, PhD[2], John Aberdeen, MS[3], Brad Malin, PhD[4]

[1] Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA
[2] Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden
[3] The MITRE Corporation, Bedford, Massachusetts, USA
[4] Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA

Abstract and Objective

The increased use and adoption of Electronic Health Records, and parallel growth in patient data available for secondary use by clinicians, researchers, and operational purposes, all cause patient confidentiality protection to become an increasingly more important requirement and expectation. The laws protecting patient confidentiality typically require the informed consent of the patient to use data for research purposes, a requirement that can be waived if the data are de-identified. Several methods to automatically remove identifying information from clinical text have been tested experimentally over the last 10 years, guided by the HIPAA "Safe Harbor" methodology. This panel will focus on the issues related with the automatic de-identification of clinical text. It will include an overview of the domain, a demonstration of good examples of such applications in English and in Swedish with their main authors sharing development and adaptation experiences, and a discussion of the HIPAA "Safe Harbor" de-identification quality and the risk for re-identification of de-identified data. The difficulties and issues related to this task will be debated, as well as the main methods used and the performance and adaptability of these methods.

Keywords:

Natural Language Processing, Patient Data Privacy, De-identification

Panel description

Introduction: With increased use and adoption of Electronic Health Record (EHR) systems, greater amounts of readily accessible patient data are available for use by clinicians, researchers, and operational purposes. As data become more accessible, protecting patient confidentiality is a requirement key to sharing and secondary use of patient data.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA; codified as 45 CFR §160 and 164) protects the confidentiality of patient data, and the Common Rule[1] protects the confidentiality of research subjects. These laws typically require the informed consent of the patient and approval of the Internal Review Board (IRB) to use data for research purposes, but these requirements are waived if data are de-identified. For clinical data to be considered de-identified, the HIPAA "Safe Harbor" technique requires 18 data elements (called PHI: Protected Health Information) to be removed.[2]

Anonymization and de-identification are often used interchangeably, but de-identification only means that explicit identifiers are hidden or removed, while anonymization implies that the data cannot be linked to identify the patient (i.e. de-identified is often far from anonymous). Scrubbing is also sometimes used as a synonym of de-identification. The de-identification of narrative text documents is often realized manually, and requires significant resources. Dorr et al.[3] have evaluated the time cost to manually de-identify narrative text notes (average of 87.2 ± 61 seconds per note), and concluded that it was time-consuming and difficult to exclude all PHI required by HIPAA. Already well aware of these issues, several authors have investigated automated de-identification of narrative text documents from the EHR. Some will be presented and discussed during this panel.

Several automatic de-identification applications have already been developed and tested with clinical text;[4] however, they are all specially adapted to the format and content of a few types of clinical documents in a specific institution. Similarly with other applications of NLP to clinical text, their generalizability is limited.[5] Certain methods are more generalizable than others, and certain methods perform better with different types of PHI than others. These adaptation and generalizability issues will be debated during this panel.

Panel overview: This panel will focus on the issues related with the automatic de-identification of clinical text. It will include an overview of the automatic de-identification of clinical text, a demonstration of good examples of such applications in English and Swedish with their main authors sharing adaptation experiences, and a discussion of the HIPAA "Safe Harbor" de-identification quality and the risk for re-identification of de-identified data. The difficulties and issues related to this task will be debated, as well as the main methods used and the performance and adaptability of these methods.

Learning objectives:

- Review automatic clinical text de-identification (overview, methods and resources used, performance of several existing systems) with attendees.
- Demonstrate two good examples of automatic clinical text de-identification applications, and share lessons learned about methods selection and adaptation.
- Teach attendees about the quality of de-identification, about the risk for re-identification.
- Discuss and share concrete examples of issues related to automatic de-identification methods and their performance.

Strategies to engage the audience: The panel organizer and presenters will ask the audience a few questions related with

their presentation and how it relates with the audience's experience.

The panel will also include a final discussion period, when panel presenters will invite the audience to share and discuss their own experience and how they relate to the presentations.

Panel organizer and participants

Clinical text de-identification at the VA [SMM]

Dr. Meystre, MD, PhD (University of Utah, Salt Lake City, Utah, USA) is a faculty member of the Department of Biomedical Informatics at the University of Utah (Salt Lake City, Utah, USA) with research activities focused on easing access to clinical data for research and clinical care purposes, using techniques such as Natural Language Processing for information extraction and automated de-identification, and automating ontologies development and management.

The U.S. Department of Veteran's Affairs (VA) is funding an informatics initiative called the Consortium for Healthcare Informatics Research (CHIR), focused on utilizing both structured and unstructured data previously unavailable for research and operational purposes. Evaluating existing de-identification methods[4] and potentially building and evaluating new methods and tools is one of the cornerstones of this initiative, and will be presented during this panel. Realized efforts shall fulfill the ethical and legal obligations of patient privacy and confidentiality.

Dr. Meystre is leading the CHIR de-identification project, to build a best-of-breed de-identification system for VA clinical documents, evaluate its interaction with subsequent text processing, and eventually how anonymous de-identified documents are. He will first introduce the audience to the principles of de-identification and to the issues related with text de-identification. He will then present the best-of-breed system developed for VA clinical text,[6] and share generalizability and adaptation to VA text difficulties.

De-identification of Swedish Clinical Text [HD]

Hercules Dalianis, PhD, is a professor in Computer and Systems Science at Stockholm University, Sweden. Dalianis is leading the research group in Clinical Text Mining that has been working with the Stockholm EPR (Electronic Patient Record) Corpus for six years. Stockholm EPR Corpus consists of one million patient records in Swedish from over 800 clinics encompassing the years 2006-2010.[7]

The group's main focus has been factuality level detection in diagnosis[7] and negation detection in the free text.[9] The purpose has been to build information extraction tools for clinical text to assist physicians in their daily work but also tools for hospital intelligence for the hospital management as for example detecting the amount of hospital-acquired infections among inpatients.[10]

The group has carried initial work in annotation of PHI following the HIPAA standard.[11] These annotations have been used for training different de-identification tools based on machine learning algorithms (CRF, Conditional Random Fields).[12] Moreover, a rule-based system has been used to pseudonymize the de-identified text, also called resynthesis.[13] However, in general, the group has worked with the Stockholm EPR Corpus in a very protected, encrypted and off-line environment.[7]

The MITRE Identification Scrubber Toolkit [JA]

John Aberdeen is the project leader for the open-source MITRE Identification Scrubber Toolkit (MIST), which provides an environment to support rapid tailoring of automated de-identification to different record types, using automatically learned classifiers.[14] He will present results and lessons learned from a series of experiments training and testing MIST with different clinical note types.

Free text de-identification systems based on pattern matching and lists require considerable upkeep. By contrast, trainable statistical systems are updated by re-training with annotated examples of novel data or new record types. The latter provide a cleaner separation of the domain-specific clinical knowledge required to identify PHI from the software that applies such knowledge.

De-identification quality / re-identification risk [BM]

Over the past several years, Dr. Malin and his team at Vanderbilt University Medical Center have collaborated with researchers from the MITRE Corp. in the evaluation and extension of their de-identification system (i.e. MIST), and have performed systematic investigation into how the replacement of identifiers with false but realistic information influences the performance and reliability of natural language text de-identification tools.

Dr. Malin has developed and will present computational approaches to formally model and measure the re-identification risk of residual identifiers. His most recent work[15-17] provides a quantifiable approach to evaluating how residual personal identifiers, such as demographics, can be exploited for re-identification purposes. This research illustrates that the threats are often context dependent; for instance, it was recently shown that risk, from a technical and economic perspective, varies with respect to the geographic domain (e.g., U.S. state) from which the data was collected.

Dr. Malin is currently serving as an expert consultant to the Office for Civil Rights (OCR), at the Department of Health and Human Services, to draft guidance on the de-identification standards associated with the HIPAA Privacy Rule. Based on his experiences, he will present recent results from a workshop held by OCR and how medical privacy regulation will be revised.

Statement of the panel organizer

All participants listed in this proposal have agreed to take part in this panel.

References

1. GPO US. 45 C.F.R. § 46 Protection of Human Subjects. 2008. Available from: http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr46_08.html
2. GPO US. 45 C.F.R. § 164 Security and Privacy. 2008. Available from: http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html
3. Dorr DA, Phillips WF, Phansalkar S, Sims SA, Hurdle JF. Assessing the difficulty and time cost of de-identification in clinical narratives. Methods Inf Med. 2006;45(3):246-52.

4. Meystre, S. M., F. J. Friedlin, et al. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010;10: 70.
5. Hripcsak, G., G. J. Kuperman, et al. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med. 1998;37(1): 1-7.
6. Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. Journal of the American Medical Informatics Association. 2013 Jan 1;20(1):77–83.
7. Dalianis, H, M. Hassel, A. Henriksson and M. Skeppstedt. 2012. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. Proceedings of the Fourth Swedish Language Technology Conference, (SLTC-2012), Lund, Sweden, October 25-26, 2012, pp. 17-18.
8. Velupillai, S., H. Dalianis and M. Kvist. 2011. Factuality levels of diagnoses in Swedish medical text. MIE 2011, In User Centred Networked Health Care, A. Moen et al. (Eds.) IOS Press, 2011, European Federation for Medical Informatics, doi:10.3233/978-1-60750-806-9-559, pp 559-563
9. Skeppstedt M. 2011. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. Journal of Biomedical Semantics 2011, 2(Suppl 3):S3.
10. Ehrentraut, C, H. Tanushi, H. Dalianis and J. Tiedemann. 2012. Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records. A machine learning approach using Naïve Bayes, Support Vector Machines and C4.5. In the proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data, AND, December 9, 2012 held in conjunction with Coling 2012, Bombay.
11. Velupillai, S., H. Dalianis, M. Hassel and G. H. Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. International Journal of Medical Informatics (2009), doi:10.1016/j.ijmedinf.2009.04.005,
12. Dalianis, H. and S. Velupillai. 2010. De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields, Journal of Biomedical Semantics 2010, 1:6 (12 April 2010).
13. Alfalahi, A., S. Brissman and H. Dalianis. 2012. Pseudonymisation of person names and other PHIs in an annotated clinical Swedish corpus. In the Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, May 26, Istanbul, pp 49-54.
14. Yeniterzi R, Aberdeen J, Bayer S, Wellner B, Hirschman L, Malin B. Effects of personal identifier resynthesis on clinical text de-identification. J Am Med Inform Assoc. Mar 1;17(2):159-68.
15. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc. Mar 1;17(2):169-77.
16. Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J Investig Med. Jan;58(1):11-8.
17. Aberdeen J, Bayer S, et al. The MITRE Identification Scrubber Toolkit: Design, training and assessment. Intl J Med Inform. 2010;79:849-59.