De-Identification of Clinical Text: Stakeholders' Perspectives and Acceptance of Automatic De-Identification

Stéphane M. Meystre, MD, PhD[a], Jonathan C. Silverstein, MD, MS[b], Guergana K. Savova, PhD[c], Valentina Petkov, MD, MPH[d], Bradley Malin, PhD[e]

[a] Medical University of South Carolina, Charleston, SC
[b] University of Pittsburgh School of Medicine, Pittsburgh, PA
[c] Boston Children's Hospital and Harvard Medical School, Boston, MA
[d] Surveillance Research Program, National Cancer Institute, Bethesda, MD
[e] Vanderbilt University Medical Center, Nashville, TN

Abstract:

Large quantities of patient clinical data are becoming available in an electronic format, generated by the fast-growing adoption of electronic health record (EHR) systems in the U.S. This growth creates tremendous potential but also a growing concern for patient confidentiality. Secondary use of this clinical data is essential to fulfill the potential for personalized healthcare, improved healthcare management, and effective clinical research. De-identification of patient data has been proposed as a solution to both facilitate secondary use of clinical data and protect patient data confidentiality. A substantial amount of clinical data in the EHR are represented as narrative text and de-identification of such data is a tedious and costly manual endeavor. Automated approaches based on natural language processing have been implemented and evaluated, allowing for much faster de-identification than manual approaches, with comparable or even improved protection. However, despite these advances, automatic de-identification of clinical text is not commonly used and accepted. This panel will focus on automatic de-identification of EHR text with perspectives from various stakeholders, reporting on a workshop organized by the National Cancer Institute on February 25-26, 2020 and supported by the Cancer Moonshot Initiative. Discussions will aim at broad sharing of opinions, ideas, advice, and practical experiences with clinical text de-identification.

Introduction:

Large quantities of patient data are becoming available in an electronic format. This data poses tremendous potential, but also concerns over patient confidentiality. The adoption of electronic health record (EHR) systems is growing at a fast pace in the U.S., encouraged by the Centers for Medicare and Medicaid Services incentive payments for meaningful use, and the prospect of improved healthcare quality. This growing adoption of EHR systems fuels the progression of clinical data available in electronic format. The secondary use of such data can provide a number of benefits, ranging from more effective scientific research to improved healthcare management and, ultimately, better quality in healthcare. Facilitating secondary use requires information sharing, but it is critical to ensure that patient privacy is maintained. In addition, since 2003, the National Institutes of Health (NIH) expects that all sponsored research projects (with at least $500,000 in direct costs in any year of funding) develop a data sharing plan for their research data in a way that upholds the confidentiality of the research subjects.[1] The Health Insurance Portability and Accountability Act of 1996 (HIPAA)[2] and the Common Rule[3] protect the confidentiality of patient and research subject data,[1] and require the informed consent of the patient or research subject and an approval of the Internal Review Board to use clinical data for research purposes. This requirement can be waived if the data is de-identified. However, obtaining consent can be difficult, if not impossible, when dealing with large populations and retrospective investigations. Requesting patient consent can also bias and adversely affect trial participation.[4] These reasons make clinical text de-identification desirable for clinical research, especially for large-scale research.

For flexibility, expressiveness, efficiency, and historical reasons, most detailed clinical information found in EHRs is still captured in free-text format, without structure or coding. De-identification of this clinical text may theoretically be accomplished by applying the Safe Harbor implementation of the HIPAA Privacy Rule, whereby explicit and quasi-identifiers about a patient are removed or hidden. This can be a tedious and costly manual endeavor, such that automated approaches based on natural language processing (NLP) have been developed and are comparable in performance to manual efforts .[5]

Although there has been significant progress made in the development of such systems the use of these systems remains limited. This is due, in part to various issues including, but not limited to 1) confusion due to terminology variation (e.g., anonymization vs. de-identification vs. scrubbing, vs. pseudonymization), 2) variation in the interpretation of the Safe Harbor method, 3) growing concerns for unauthorized access to clinical data (i.e., leaks), 4) considerations in regard to the risk for re-identification in de-identified clinical text, 5) limited options for accurate and sufficiently simple applications for automatic text de-identification, and 6) limited acceptance by providers and IRBs for the release of automatically de-identified text – particularly in the face of perceived large liabilities.

Panel overview:
This **interactive** panel will focus on automatic de-identification of EHR text with perspectives from various stakeholders. Discussions will aim at broad sharing of opinions, ideas, advice, and practical experiences with clinical text de-identification.

Topics discussed will include:

- Regulatory landscape for de-identification of clinical data for research use
- Current status, performance, and future development of clinical text de-identification systems
- Real world experience with de-identification of narrative clinical text for research (e.g., challenges, data sharing policies and practices)
- Important aspects to consider for de-identification of narrative text (e.g., re-identification risk)
- Stakeholders perspective and acceptance of automatic text de-identification

Learning objectives: During and after this session, participants should be better able to:

- Contrast characteristics and challenges of clinical text de-identification.
- Share experiences and ideas for improved acceptance and understanding of text de-identification.
- Evaluate practical options for text de-identification use.

Intended audience: This panel is addressed to professionals with activities and interests in clinical data secondary use or reuse, as enabled by data de-identification. It will mostly interest professionals planning to use text de-identification or currently experiencing difficulties with text de-identification.

Expected discussion and strategies to engage the audience: Participation of the audience in discussions will be key, for broader experience, idea, and advice sharing. The panel moderator and presenters will start with brief presentations of key challenges and experiences with text de-identification, and ask the audience questions related to their presentation and how it relates with the audience's experience. Panel moderator and presenters will invite the audience to share and discuss their own experiences and how they relate to the presentations.

Panel organizer and participants:

Bradley Malin, PhD, FACMI, NAM will moderate this panel and present aspects to consider for de-identification of narrative information, including how to assess the risk for re-identification of de-identified narratives. Dr. Malin will provide illustrations of how text de-identification has been applied in various settings, including Vanderbilt University Medical Center and clinical study reports from clinical trials submitted to the European Medicines Agency (where he serves as a member of an advisory board on data privacy). He will further provide insight into how such practices are planned for application in the context of the NIH-sponsored All of Us Research Privacy. Dr. Malin is the Vice Chair for Research Affairs in the Department of Biomedical Informatics at Vanderbilt University Medical Center. He is also a Professor of Biomedical Informatics, Biostatistics, and Computer Science at Vanderbilt University Medical Center, where he runs the Health Data Science Center. Since 2018, he has served as the chair of the Committee on Access, Privacy, and Security for the All of Us program.

Stephane Meystre, MD, PhD, FACMI will present an overview of the current status, performance, and future development of clinical text de-identification systems. He will then present an example of such systems in more details (CliniDeID). Dr. Meystre, is Associate Professor and SmartState Chair in Translational Biomedical Informatics at the Medical University of South Carolina (Charleston, SC) 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, patient trial eligibility discovery and automated text de-identification. His extensive experience on the latter topic spans a large-scale text de-identification project at the VA resulting in the development of a best-of-breed system (nicknamed ‘BoB’), subsequent developments to further improve the accuracy and generalizability of this system, and continuing development and strengthening resulting in CliniDeID®, a commercial product recently launched by Clinacuity, Inc.

Jonathan Silverstein, MD, MS, will present real world experiences with de-identification of narrative clinical documents for research, insisting on its importance, challenges, data sharing policies and practices. Specifically, he will focus on the decades long experience at the University of Pittsburgh/UPMC in the use of EHR text in research: from large early collection of this clinical text for research (e.g. the MARS system); to leading informatics research in natural language processing (e.g. NegEx) and text de-identification (e.g. De-ID); to practical use at scale across many research and operational projects (e.g. R3). Dr. Silverstein currently serves as Professor and Chief Research Informatics Officer at the University of Pittsburgh (Pitt), where his responsibilities include data provisioning and honest brokering for research at UPMC, an integrated global nonprofit health enterprise including 40 hospitals and 700 clinical locations and at UPMC’s academic partner the University of Pittsburgh (5th in NIH total funding). The data provisioning service, Health Record Research Request (R3) operates as a core facility at Pitt under BAA with UPMC with support from the CTSA program. Dr. Silverstein will detail essential features of Pitt/UPMC policy structure and the R3 service that enable de-identified clinical information construction, use at scale, and sharing for numerous research projects.

Guergana Savova, PhD, will present a stakeholder’s perspectives and acceptance of automatic text de-identification issues – that of an NLP or artificial intelligence practitioner. What are the main considerations to keep in mind for downstream use of de-identified clinical text? How does the de-identification style affect the algorithms and their usability and portability? Dr. Savova is Associate Professor at Computational Health Informatics Program (CHIP) and Harvard Medical School. Her research interests are in a Natural Language Processing, a sub-field of Artificial Intelligence that has witnessed astounding developments in the last several years. The mission of Dr. Savova’s lab is processing all health-related language with the goal to advance biomedicine. Dr. Savova and her lab have been part of many high-impact federally funded projects. They have contributed to the open-source community Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES at ctakes.apache.org) and Deep Phenotyping for Cancer (DeepPhe at https://github.com/DeepPhe/DeepPhe-Release) which have been widely used. The Apache Software Foundation recognized cTAKES as one of its top 20 most influential projects. Dr. Savova is a co-founder of cTAKES and DeepPhe.

Statement of the panel organizer: All participants listed in this proposal have agreed to take part in this panel.

References

1. NIH Statement on Sharing Research Data. 2003. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

2. CFR Title 45 Subtitle A Part 164: Security and Privacy. Available from: https://www.govinfo.gov/app/details/CFR-2019-title45-vol2/CFR-2019-title45-vol2-part164

3. CFR Title 45 Part 46: Protection of Human Subjects. 1991. Available from: https://www.govinfo.gov/app/details/CFR-2016-title45-vol1/CFR-2016-title45-vol1-part46

4. Dunlop AL, Graham T, Leroy Z, Glanz K, Dunlop B. The impact of HIPAA authorization on willingness to participate in clinical research. Ann Epidemiol. 2007;17(11):899-905.

5. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010;10:70.