Clinical Text Automatic De-Identification to Support Large Scale Data Reuse and Sharing: Pilot Results

Stéphane M. Meystre, MD, PhD[1,2], Paul M. Heider, PhD[1], Youngjun Kim[1], PhD, Andrew Trice, BS[2], Gary Underwood, MS[2]

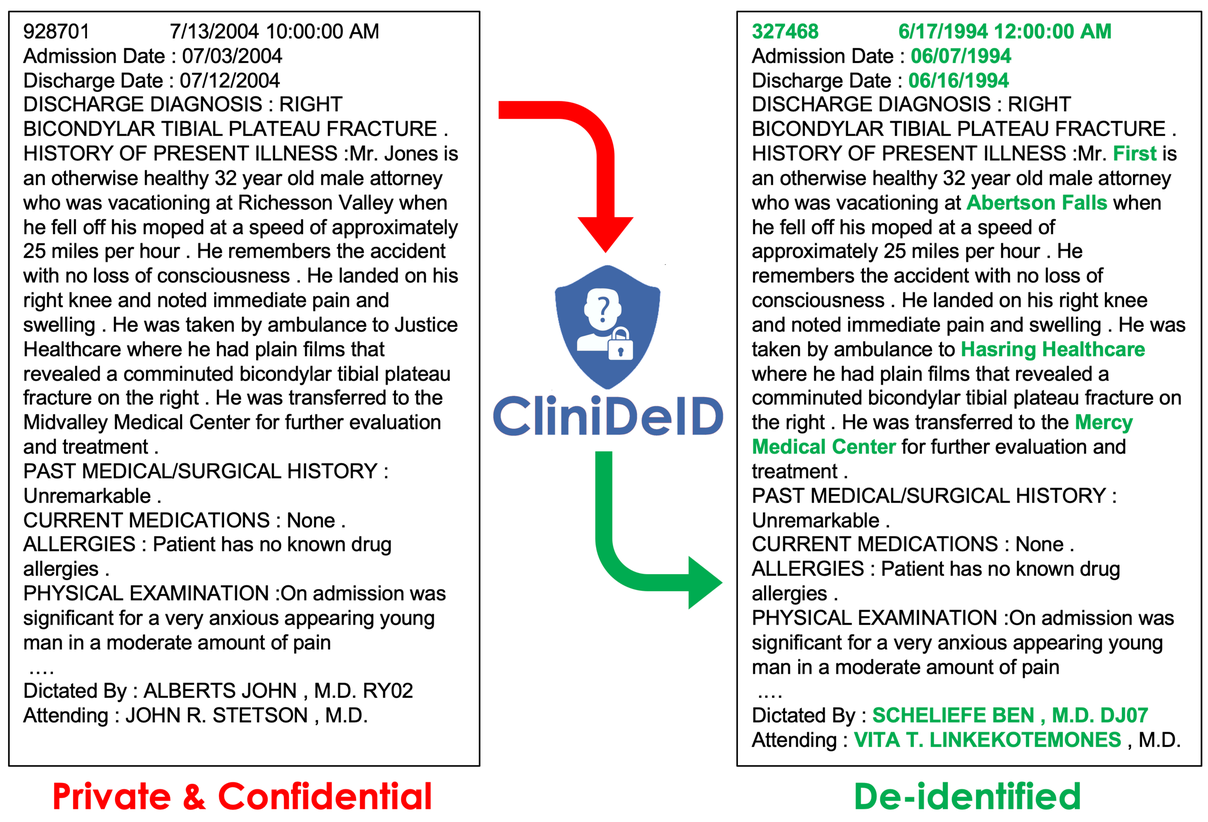
[1] Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC
[2] Clinacuity, Inc., Charleston, SC

Abstract

De-identification of patient data has been proposed as a solution to facilitate secondary uses of clinical data and protect patient data privacy.

Automated approaches based on Natural Language Processing have been evaluated, allowing for much faster de-identification than manual approaches.

This pilot study includes the evaluation of three versions of a new text de-identification application, with pattern matching, machine learning, and ensemble methods.



Results

- Reference standard annotated with average agreement between annotators ≥ 98% (Cohen's kappa).
- Evaluation of the NLP prototype accuracy done in two steps: 1) MUSC corpus of 250 annotated clinical notes and 2) 2014 i2b2 NLP challenge testing corpus of 514 annotated discharge summaries.[4]
- Highest recall with ensemble method and 2014 i2b2 challenge testing corpus, but lower when tested with a different corpus (MUSC).

Prototype evaluation results (PHI level)

	TP	FN	FP	Recall	Precision	F_1-measure
Rules-based and stepwise system (train and test 2014 i2b2)	10286	902	6999	0.919	0.595	0.723
Rules-based and stepwise system (train and test MUSC)	755	106	134	0.877	0.849	0.863
CRF-based system (train and test 2014 i2b2)	10476	794	1184	0.930	0.898	0.914
Ensemble method (train and test 2014 i2b2)	*	*	*	**0.983**	**0.978**	**0.980**
Ensemble method (train i2b2, test MUSC)	*	*	*	0.759	0.926	0.834

Introduction

The adoption of Electronic Health Record (EHR) systems is growing at a fast pace in the U.S. This growth results in very large quantities of patient clinical data becoming available in electronic format, with tremendous potentials, but also equally growing concern for patient confidentiality breaches.

De-identification of patient data has been proposed as a solution to both facilitate secondary uses of clinical data and protect patient data confidentiality.

The majority of clinical data found in the EHR is represented as text notes but de-identification of clinical text is a tedious and costly manual endeavor. Automated approaches based on Natural Language Processing (NLP) have been implemented and evaluated, allowing for much faster de-identification than manual approaches.[1] The HIPAA Safe Harbor method was used in most approaches.

Methods

This feasibility study included the following:

Creation of a reference standard for training and testing the text de-identification application; it included a random sample of 250 clinical narratives annotated by two domain experts (medical residents and fellows at MUSC) independently, and a third adjudicated if there was disagreement.

Three versions of the NLP application prototype were developed:

- **Rules-based** system: implemented the stepwise hybrid approach from BoB.[2] It included three main components: text pre-processing, high-sensitivity extraction, and a false positives filter. The former reuses equivalent components from BoB. High-sensitivity extraction reuses the pattern matching and dictionaries from BoB. The false positives filter implements machine learning classifiers (SVM) retrained with the 2014 i2b2 challenge corpus[3] or our small MUSC corpus.
- **CRF-based** system: consisted of pre-processing and feature extraction followed by a conditional random fields (CRF) classifier trained with the 2014 i2b2 challenge corpus.
- **Ensemble method:** implemented ensemble methods combining four different machine learning algorithms (CRF, SVM, MIRA, and a RNN (recurrent neural network)) all trained with the 2014 i2b2 challenge corpus and combined using a voting algorithm with a threshold of 1.

Testing of the prototype with the local reference standard of clinical narratives from the Medical University of South Carolina (MUSC). Validation and generalizability testing with a larger corpus of clinical narratives from another healthcare organization (Partners Healthcare, Boston, MA).

Acknowledgments: Research supported by the National Institute for General Medical Sciences (NIGMS) (R41GM116479).

References:
1. 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.
2. 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;20(1):77-83.
3. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform. September 2015:1-9.
4. Uzuner O, Stubbs A. Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks. J Biomed Inform. 2015;58:S1-S5.

Contact: meystre@musc.edu

www.clinacuity.com