AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

Automatic Text De-Identification: How and When is it Acceptable?

Session S47; November 5, 2018

Stephane M. Meystre, MD, PhD, FACMI

Presenters

Stephane M. Meystre, MD, PhD (Medical University of South Carolina, Charleston, SC)

David S. Carrell, PhD (Kaiser Permanente Washington Health Research Institute, Seattle, WA)

John Aberdeen, MA (The MITRE Corporation, Bedford, MA)

Valentina Petkov, MD, MPH (Surveillance Research Program, National Cancer Institute, Bethesda, MD)

Jonathan C. Silverstein, MD, MS (University of Pittsburgh School of Medicine, Pittsburgh, PA)

Disclosures

I disclose the following relevant relationship with commercial interests:

- Shareholder of Clinacuity, Inc.

Learning Objectives

After participating in this session the learner should be better able to:

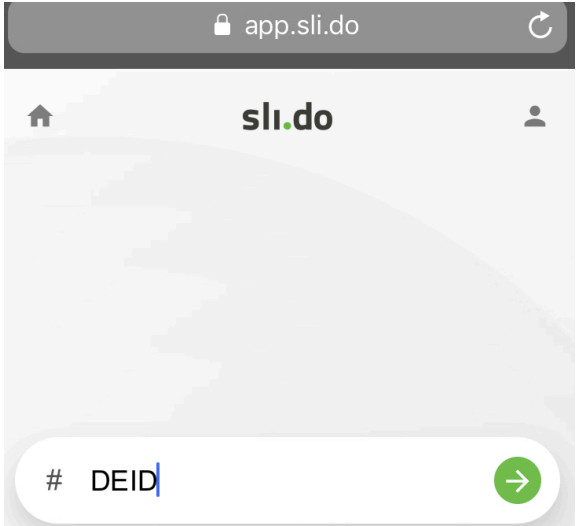
- Contrast characteristics and challenges of clinical text de-identification.
- Share experiences and ideas for improved quality, acceptance, and use of text de-identification.
- Evaluate practical options for text de-identification use.

Interactivity and Questions

Please participate and provide feed-back!

1. Tell us about your experience with text de-identification
2. Answer questions from presenters
3. Ask your questions anytime

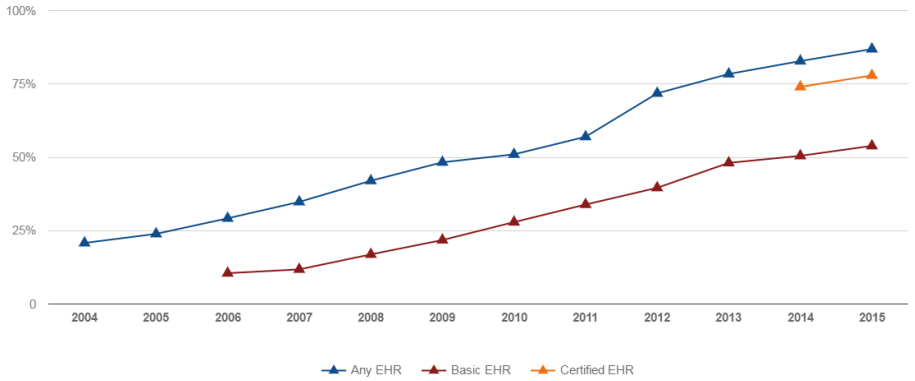
We use **Slido** as online survey system
Go to: **sli.do**
Event code: **#DEID**



Introduction

Fast growth in clinical data: The adoption of Electronic Health Record (EHR) systems is already well-advanced in several countries, or growing at a fast pace in other countries such as the U.S. where it is used in most hospitals.

Results in very large quantities of patient clinical data becoming available in electronic format, with tremendous potentials, but also growing concern for patient confidentiality and privacy breaches, legal and policy challenges, and data interoperability and integration difficulties.



Introduction

Privacy and confidentiality of clinical data:

In the U.S., the HIPAA (Health Insurance Portability and Accountability Act) protects the confidentiality of patient data. The Common Rule protects the confidentiality of research subjects. These laws typically require the informed consent of the patient and approval of the IRB to use data for research purposes, but these requirements are waived if data are de-identified.

De-identification means that explicit identifiers are hidden or removed. Often used interchangeably with *anonymization*, but the latter 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.

Introduction

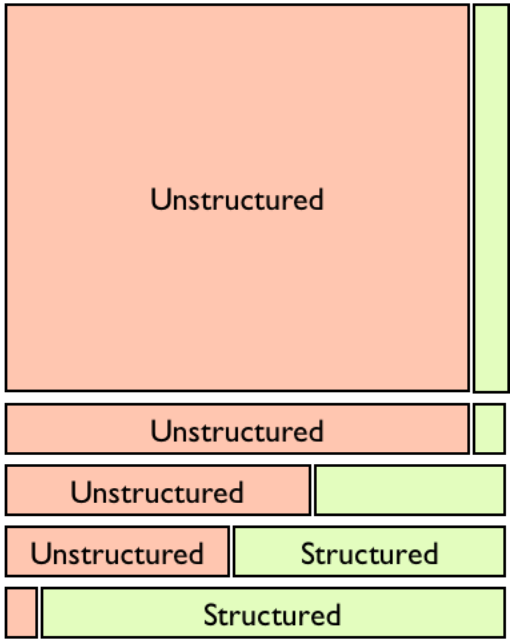
HIPAA Safe Harbor method: The following PHI about the patient, relatives, employers, or household members have to be removed:

1. Names
2. All geo-subdivisions smaller than a State
3. All elements of dates (except year)
4. Phone numbers
5. Fax numbers
6. Electronic mail addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators
15. Internet Protocol address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. **Any other unique identifying number, characteristic, or code**

Introduction

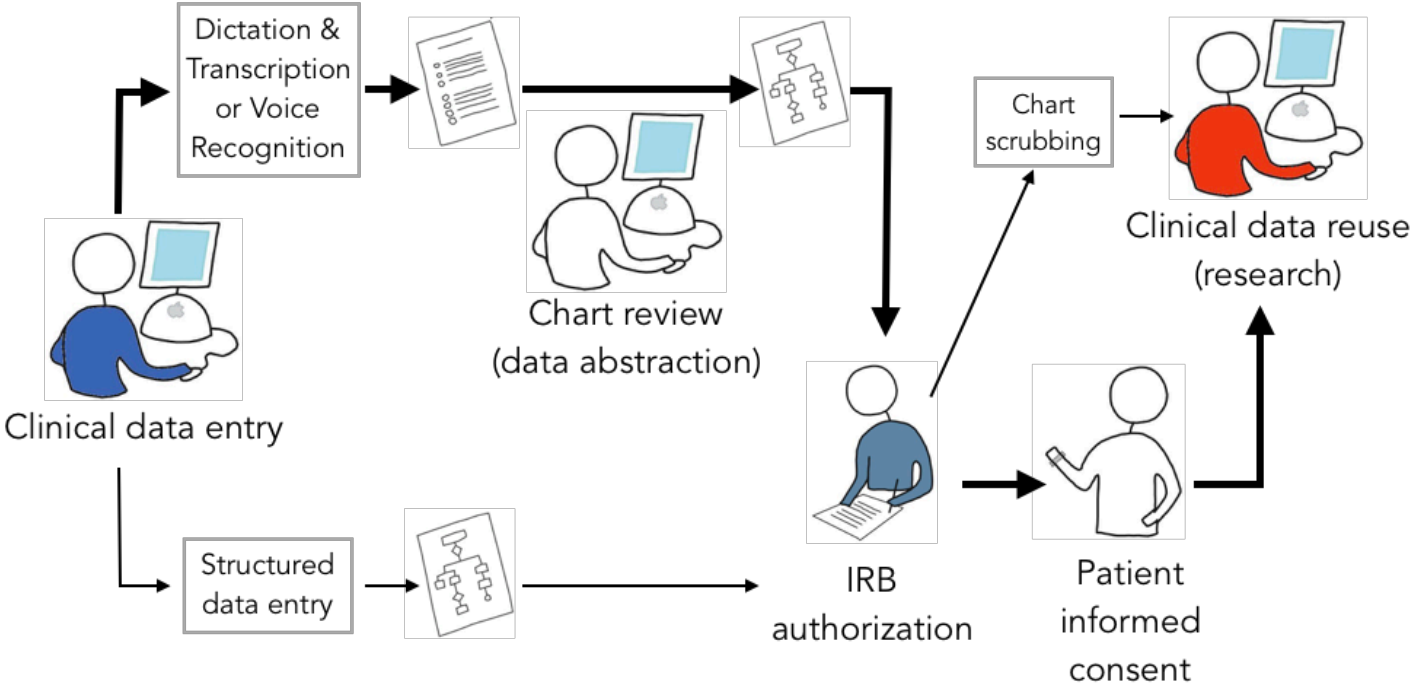
Most EHR data is unstructured text:

- Documents
  - History and Physicals
  - Clinical notes, Consult notes
  - Operative reports
  - Surgical pathology reports
  - Progress notes, Letters
  - Orders
  - Discharge summaries
- Imaging / Radiology
- Prescriptions (pharmacy; CPOE)
- Laboratory results
- Administrative information



Introduction

Current common example of data entry and reuse



Introduction

Why use NLP for text de-identification?

Manual text de-identification is a lengthy and costly process (about 90 s per document).

NLP can be used to automatically de-identify electronic clinical documents.

The text de-identification process is composed of two main steps:

- **PHI detection**, and then
- **PHI removal or transformation**: replacing PHI with some tags or characters (e.g., ‘Mr. Smith’ becomes ‘<Patient_name>’), or replace PHI with synthetic but realistic substitutes (e.g., ‘Mr. Smith’ becomes ‘Mr. Jones’) = PHI "resynthesis”

Introduction

Text de-identification, from original text

928701 7/13/2004 10:00:00 AM
Admission Date : 07/03/2004 Discharge Date : 07/12/2004
DISCHARGE DIAGNOSIS : RIGHT BICONDYLAR TIBIAL PLATEAU FRACTURE .
HISTORY OF PRESENT ILLNESS :
Mr. Jones is an otherwise healthy 32 year old male attorney who was vacationing at Richesson Valley when he fell off his moped at a speed of approximately 25 miles per hour . He remembers the accident with no loss of consciousness . He landed on his right knee and noted immediate pain and swelling . He was taken by ambulance to Justice Healthcare where he had plain films that revealed a comminuted bicondylar tibial plateau fracture on the right . He was transferred to the Midvalley Medical Center for further evaluation and treatment .
PAST MEDICAL/SURGICAL HISTORY : Unremarkable .
CURRENT MEDICATIONS : None .
ALLERGIES : Patient has no known drug allergies .
PHYSICAL EXAMINATION :
On admission was significant for a very anxious appearing young man in a moderate amount of pain .
…
Dictated By : ALBERTS JOHN , M.D. RY02
Attending : JOHN R. STETSON , M.D.

Introduction

Text de-identification, from original text to identified PHI

928701 7/13/2004 10:00:00 AM
Admission Date : 07/03/2004 Discharge Date : 07/12/2004
DISCHARGE DIAGNOSIS : RIGHT BICONDYLAR TIBIAL PLATEAU FRACTURE .
HISTORY OF PRESENT ILLNESS :
Mr. Jones is an otherwise healthy 32 year old male attorney who was vacationing at Richesson Valley when he fell off his moped at a speed of approximately 25 miles per hour . He remembers the accident with no loss of consciousness . He landed on his right knee and noted immediate pain and swelling . He was taken by ambulance to Justice Healthcare where he had plain films that revealed a comminuted bicondylar tibial plateau fracture on the right . He was transferred to the Midvalley Medical Center for further evaluation and treatment .
PAST MEDICAL/SURGICAL HISTORY : Unremarkable .
CURRENT MEDICATIONS : None .
ALLERGIES : Patient has no known drug allergies .
PHYSICAL EXAMINATION :
On admission was significant for a very anxious appearing young man in a moderate amount of pain .
…
Dictated By : ALBERTS JOHN , M.D. RY02
Attending : JOHN R. STETSON , M.D.

Introduction

Text de-identification, from original text to identified and resynthesized PHI

327468 6/17/1994 12:00:00 AM
Admission Date : 06/07/1994 Discharge Date : 06/16/1994
DISCHARGE DIAGNOSIS : RIGHT BICONDYLAR TIBIAL PLATEAU FRACTURE .
HISTORY OF PRESENT ILLNESS :
Mr. First is an otherwise healthy 32 year old male attorney who was vacationing at Abertson Falls when he fell off his moped at a speed of approximately 25 miles per hour . He remembers the accident with no loss of consciousness . He landed on his right knee and noted immediate pain and swelling . He was taken by ambulance to Hasring Healthcare where he had plain films that revealed a comminuted bicondylar tibial plateau fracture on the right . He was transferred to the Mercy Medical Center for further evaluation and treatment .
PAST MEDICAL/SURGICAL HISTORY : Unremarkable .
CURRENT MEDICATIONS : None .
ALLERGIES : Patient has no known drug allergies .
PHYSICAL EXAMINATION :
On admission was significant for a very anxious appearing young man in a moderate amount of pain .
…
Dictated By : SCHELIEFE BEN , M.D. DJ07
Attending : VITA T. LINKEKOTEMONES , M.D.

Introduction

Significant progress with automated text de-identification has been made.

Many **problems remain**, including:

- Terminology confusion (e.g., anonymization vs. de-identification vs. scrubbing, vs. pseudonymization)
- Variation in the interpretation of what information must be protected under the HIPAA “Safe Harbor” method.
- Growing concerns for unauthorized access to clinical data (‘leaks’) deepened by recent incidents with health insurers like Anthem (80 million records) and Premera (affecting 11 million individuals).

Introduction

Other **problems** include:

- Concerns as to whether de-identification as defined in HIPAA is sufficiently protective of an individual's identity to justify ‘‘hands off’’ approaches.
- Limited understanding of the risk for re-identification of de-identified clinical text against both human and automated attack methods.
- Limited acceptance by providers and IRBs for release of automatically de-identified text, given perceived potentially large liabilities, coupled with unclear rewards for sharing.
- Limited options for accurate and sufficiently simple applications for automatic text de-identification.


Thank you!
Remember polls & questions:
Go to: sli.do
Event code: #DEID
Email me at:
meystre@musc.edu
MUSC
Medical University
of South Carolina
Changing What's Possible | MUSC.edu
MUSC Health
Medical University of South Carolina
JAMES W. COLBERT EDUCATION CENTER AND LIBRARY