STANFORD
SCHOOL OF MEDICINE

Stanford University Medical Center

Strategies for De-Identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies

Clete A. Kushida, M.D., Ph.D.
Professor, Stanford University Medical Center

Overview

- De-Identification and Anonymization Strategies
- Why Are Such Strategies Important?
- Database Sources, Key Words, Search Strategy
- Case Examples (Text, Images, Biological Samples)
- Discussion
  - Are Current De-ID Strategies Effective?
  - Which Strategies Are Best?
  - How Essential is Anonymization?
  - Do De-ID Strategies Alone Meet the Needs of Multicenter Research Studies?
  - What Approaches Can Be Used on a Multicenter Level to Ensure Privacy?
  - Further Work

Definitions

- De-identification and anonymization are strategies that are used to remove patient identifiers in electronic health record (EHR) data.
- *De-identification* of EHR data is the removal or replacement of personal identifiers so that it would be difficult to reestablish a link between the individual and his or her data.
- *Anonymization* refers to the irreversible removal of the link between the individual and his or her medical record data to the degree that it would be virtually impossible to reestablish the link.

Why Are Such Strategies Important?

- HIPAA Privacy Rule regulations (2000) permits covered entities to use/disclose data that have been removed of patient identifiers without obtaining an authorization and without further restrictions on use/disclosure. There are 18 “safe harbor” data identifiers under the Rule that constitute the minimal set of removed identifiers.
- Use of data removed of patient identifiers is one of three options available to investigators desiring to use medical data in research, besides obtaining informed consent from their patients or a waiver of informed consent from their IRB.

Why Are Such Strategies Important?

- As the use of EHRs has progressively increased, concerns have been raised about their utility to fundamentally improve the quality of patient care and the threat of unauthorized disclosure of PHI either unintentionally or by identity theft.
- Additionally, biomedical research is becoming increasingly dependent on the access, sharing, and management of EHR among clinical and research centers, especially those involved in observational and multicenter research studies.

Database Sources For Review

- BIOSIS Previews (via Thomson Reuters Institute for Scientific Information [ISI] Web of Knowledge, 1926-present)
- CINAHL (Cumulative Index to Nursing and Allied Health Literature, via EBSCOhost, 1937-present)
- Inspec (via Thomson Reuters ISI Web of Knowledge, 1898-present)
- MEDLINE (Medical Literature Analysis and Retrieval System Online, 1950-present)
- SciVerse Scopus (1823-present)
- Web of Science (via Thomson Reuters ISI Web of Knowledge, 1898-present)

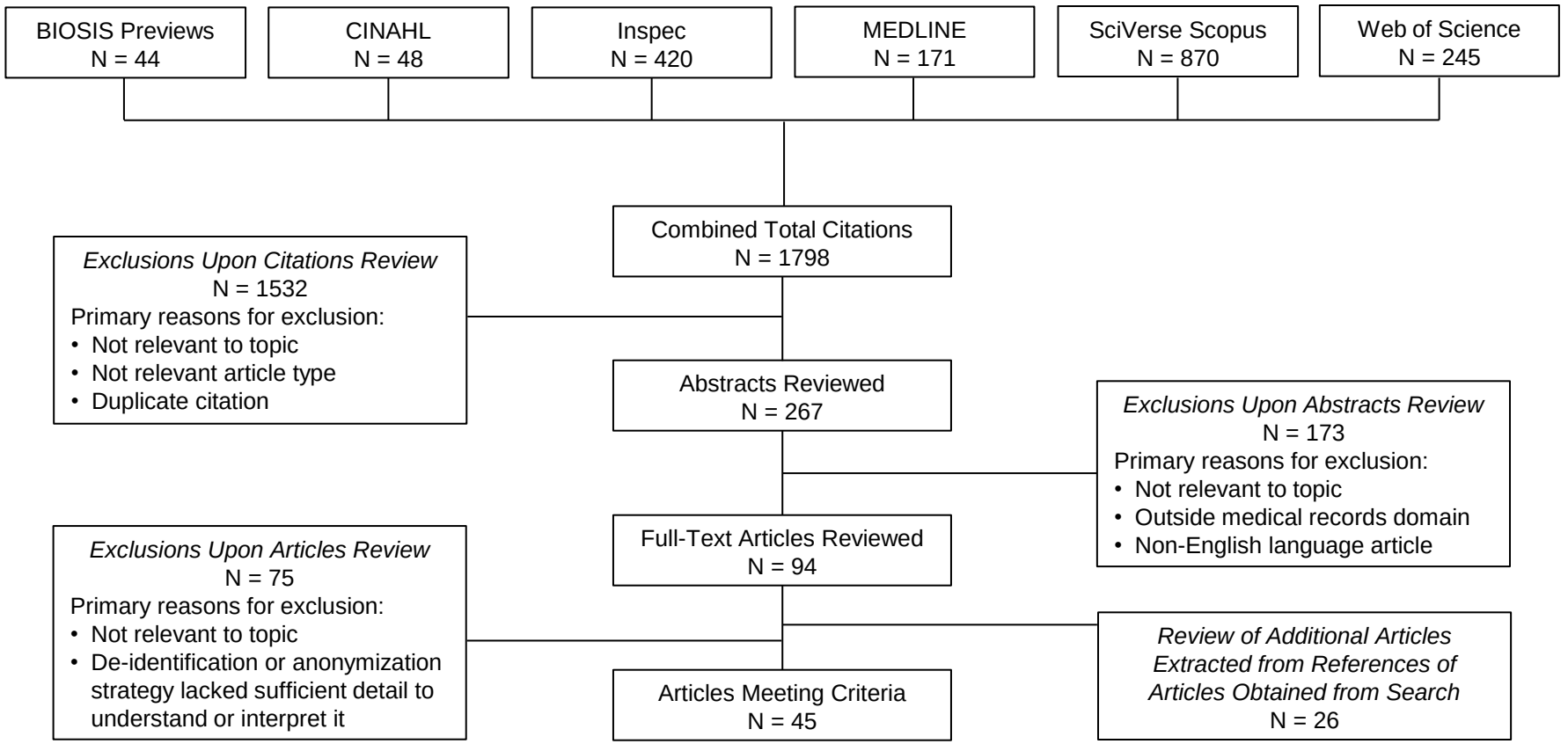
Key Words and Search Strategy

- Key Words: *deidentify, de-identify, deidentification, de-identification, anonymize, anonymization, data scrubbing, and text scrubbing*
- Articles were included if they were published up to June 30, 2011 and there was no restriction on earliest date of publication (i.e., earliest date obtained in search was 1996).
- Through the combined database search, 1798 prospective citations were identified

Search Strategy *(cont.)*

- The writing group chair conducted the review; however, five other members of the writing group independently reviewed the 120 full-text articles obtained after the abstracts review.
- Differences between the reviewers' judgments regarding inclusion or exclusion of articles were resolved by discussion; consensus was required from all six reviewers.
- The full text of 120 articles were reviewed and resulted in a final sample of 45 articles that met inclusion criteria for review.

Flow Diagram of Search Results



Case: De-Identification of Free Text

- Manual de-identification of PHI from free text in EHR can be tedious, costly, time-consuming, inaccurate, and unreliable.
- For example, resident clinicians can manually de-identify at a rate of about 18,000 words or 90 incidents of PHI per hour.

The automated software package, *deid*, scans the medical notes line-by-line, dividing them into individual words separated by whitespace

↓

deid identifies occurrences of PHI using dictionary-based look-ups and regular expressions

↓

deid replaces each PHI with a tag to indicate its corresponding category

Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.

Case: De-Identification of Free Text

DISCHARGE SUMMARY

Name: [**Known patient lastname**], [**Known patient firstname**]

[**Unit Number 626**]

Admission Date: [**2016-11-07**]

Discharge Date: [**2016-11-22**]

Date of Birth: [**1972-09-20**]

Sex: F

HISTORY OF PRESENT ILLNESS: Patient is a 44-year-old lady status post living related kidney transplant on [**2016-10-19**], who presented at [**Hospital 36**] for end-stage renal disease secondary to type 1 diabetes mellitus.

She presented to [**Hospital1 **] on [**2016-11-07**] with increased drainage from her surgical wound and JP, increased abdominal pain, and anuria x4 days. The patient reported constipation for a week. She denies flatus. She was complaining of nausea and vomiting. Her abdominal pain had become progressively worse left lower quadrant most notable. There is no radiation to the back or elsewhere. She denied any fevers, chills. She noted decreased p.o. intake recently. Her drainage from her wound incision and JP was notable for yellowish clear urine smelling fluid.

- On a test corpus of 1,836 notes with 296,400 words, there was 90 instances of false negatives (missed PHI), or 27 per 100,000 word count, with a recall (sensitivity) of 94.3%.
- Only one full date and one age over 89 were missed.
- No patient names were missed.

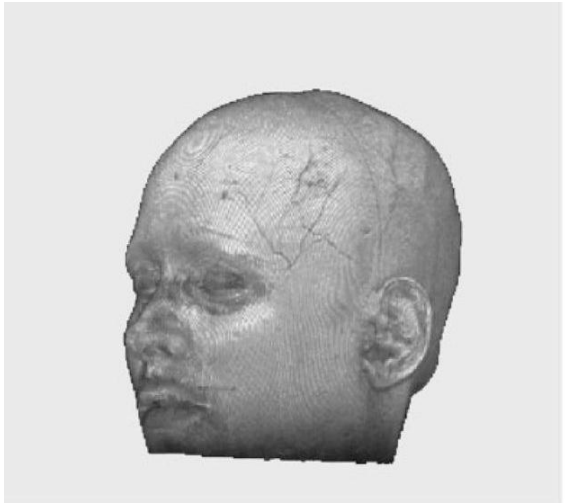
Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.

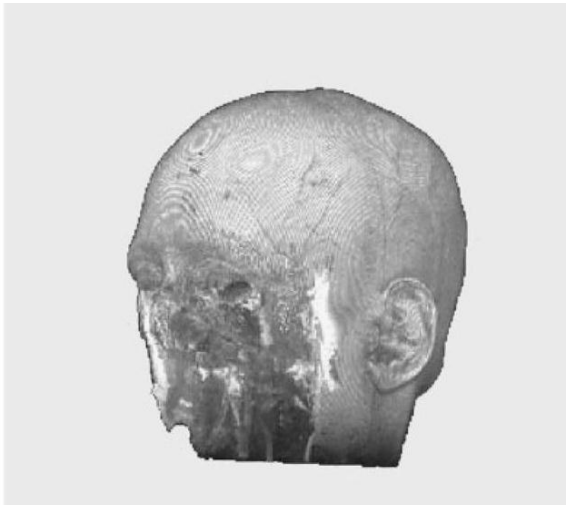
Case: De-Identification of MR Images

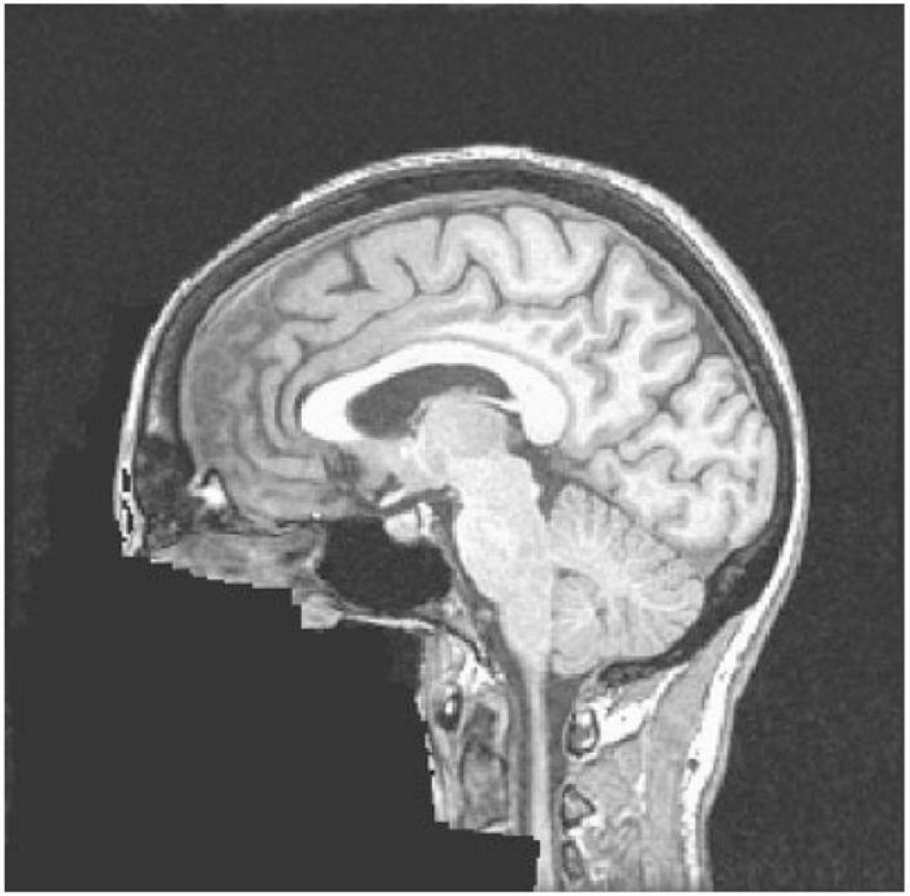
- An automated “defacing” algorithm used models of non-brain structures to remove identifiable facial features from MR volumes of 342 T1-weighted datasets:
  - Did an effective job of removing facial features without sacrificing brain tissue (none removed)
  - Could be performed relatively quickly (approximately 25 min on a dataset of 342)
  - Did not interfere with subsequent data processing, and in some cases, improved the quality of subsequent automated skull-stripping by removing more non-brain tissue.

Bischoff-Grethe A, Ozyurt IB, Busa E, et al. A technique for the deidentification of structural brain MR images. Hum Brain Mapp 2007;28:892-903

Case: De-Identification of MR Images

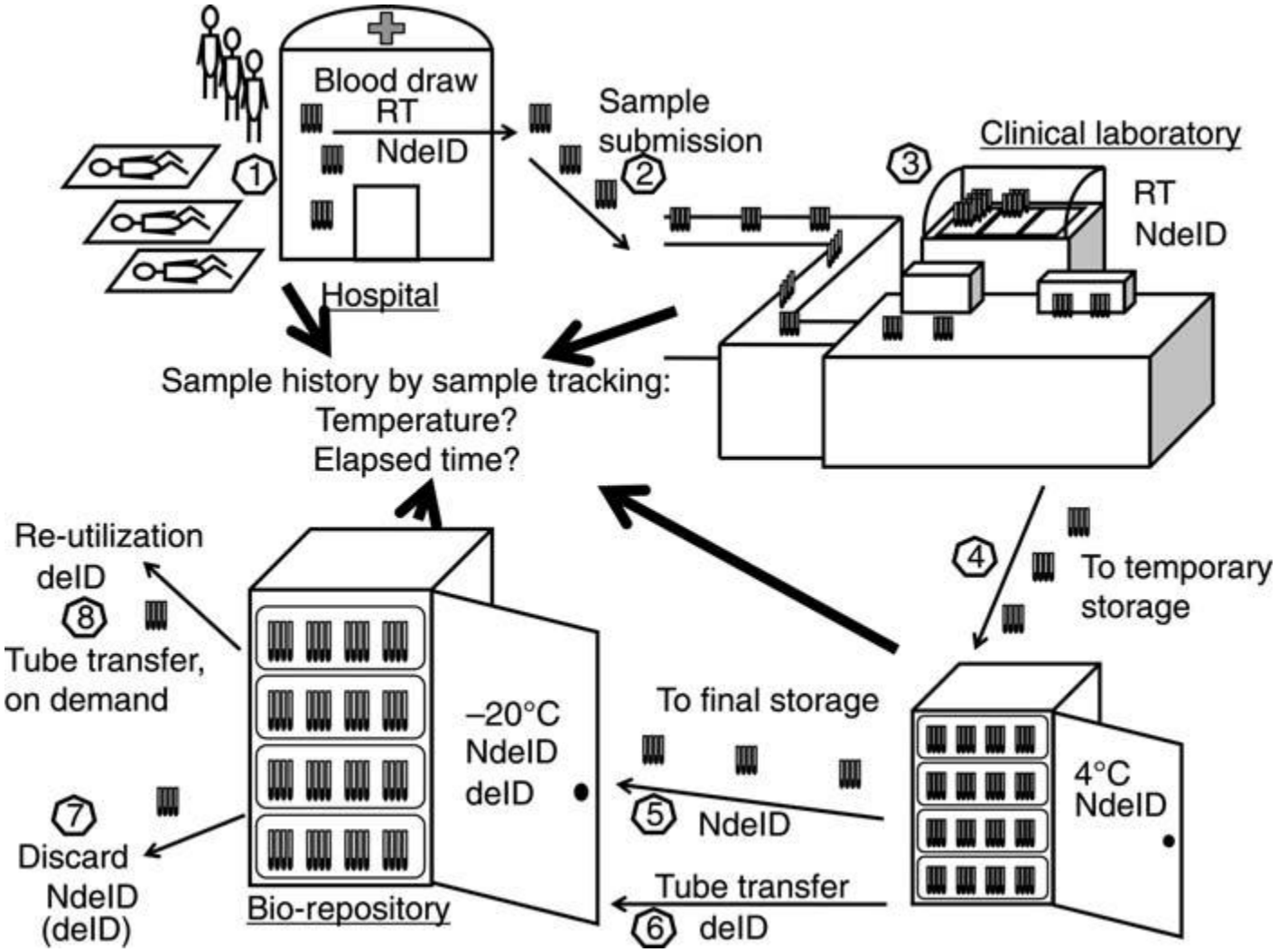






Bischoff-Grethe A, Ozyurt IB, Busa E, et al. A technique for the deidentification of structural brain MR images. Hum Brain Mapp 2007;28:892-903

Case: De-Identification of Biosamples



Furuta K, Yokozawa K, Takada T, et al. De-identification procedure and sample quality of the post-clinical test samples at the bio-repository of the National Cancer Center Hospital (NCCH) in Tokyo. Jpn J Clin Oncol 2011;41:295-298

Case: De-Identification of Biosamples

- This repository contains approximately 250,000 samples with an average influx of 90,000 samples per year of which approximately 80% need to be de-identified.
- This process differs from data scrubbing patient identifiers on the physical sample since a tube transfer procedure is used for sample de-identification.
- This is a manual de-identification procedure that is subject to human error.

Furuta K, Yokozawa K, Takada T, et al. De-identification procedure and sample quality of the post-clinical test samples at the bio-repository of the National Cancer Center Hospital (NCCH) in Tokyo. Jpn J Clin Oncol 2011;41:295-298

Are Current De-ID Strategies Effective?

- Current de-identification strategies have impressive recall and precision rates.
- No existing system is perfect, and there is the possibility that certain PHI will not be de-identified.
- Limitations of many current systems include:
  - Inability to detect misspellings, typographical errors, and proper names that share characteristics with non-PHI
  - Restrictions in managing only certain types of data; algorithms that are not designed to handle diverse PHI (e.g., hard-coded PHI in output files)
  - Difficulty in compensating for variation in nomenclature

Which Strategies Are Best?

- For heuristic, lexical, and pattern-based systems, studies evaluating these systems have reported good performance (especially precision) but experienced domain experts must spend significant time and effort.
- For statistical learning-based systems, they are able to be used “out of the box” with minimal redevelopment time and learn how to identify PHI from the data itself rather than relying on precompiled, manually-constructed sets of data.

Which Strategies Are Best? *(cont.)*

- For both images and biological samples, there are too few studies with a paucity of quantitative data to judge the best approach
- Biological samples have the added Common Rule anonymization requirements needed for IRB exemption that do not appear to be satisfactorily addressed by the current approaches.

How Essential is Anonymization?

- In theory, anonymization is important since it places the patient's or research participant's right to privacy as the top priority in any anticipated or unanticipated scenario, and dramatically minimizes the release of sensitive information that may discriminate or stigmatize the individual from a social or economic perspective.
- In practice, it still may be possible to identify an individual from supposedly anonymized data sets, especially with respect to rare diseases within a specific geographical area.

Do De-ID Strategies Alone Meet the Needs of Multicenter Research Studies?

- *Besides the de-identification of individual documents, what can be done to ensure the privacy of data sets?*
- *What approaches can be used on a multicenter level to ensure patient or participant privacy?*

What Approaches Can Be Used on a Multicenter Level to Ensure Privacy?

- De-identification and anonymization strategies are important, but are one component of an integrated data collection and management system.
- Some institutions use honest brokers, which collect and provide data to research investigators in a manner whereby it would not be reasonably possible for investigators to identify the participants directly or indirectly.

Further Work

- Management of identifiers for the protection of genetic information, particularly with respect to protecting the privacy of identities to which DNA sequences were derived.
- This area of genomic privacy is particularly challenging for the biomedical community, given the immense quantity of data that needs to be processed, stored, and shared, as well as the consequences that identifying genomic data may have on an individual's health, employment, and insurance status.

Acknowledgements

- Deborah A. Nichols, M.S., *Stanford University*
- Rik Jadrnicek, *Microflow DBMS Inc.*
- Ric Miller, *Microflow DBMS Inc.*
- James K. Walsh, Ph.D., *Sleep Medicine and Research Center, Chesterfield, MO*
- Kara Griffin, *Sleep Medicine and Research Center, Chesterfield, MO*
- Pamela R. Hyde, M.A., *Stanford University*
- AcademyHealth
- Gurvaneet Randhawa, M.D., MPH and AHRQ

Funding was provided by a contract from AcademyHealth. Additional funding was provided by AHRQ 1R01HS1973 (Comparative Outcomes Management with Electronic Data Technology (COMET) Study).