HHS Workshop on the HIPAA Privacy Rule's De-Identification Standard

March 8–9, 2010
Marriott at Metro Center, Washington, DC

A Commercial Approach to De-Identification

Dan Wasserstrom, Founder and Chairman
De-ID Data Corp, LLC

De-ID Data Corp, LLC
Founded to:
ENHANCE DATA ACCESS WHILE PROTECTING PATIENT PRIVACY

Founders Problem

Can't do NLP on Patient Records

If You Can't Get Access to Records!

History

- De-ID Data Corp founded in 2003 by a former public health physician and an NLP/Clinical Trials executive
- Application acquired under Technology Transfer Agreement from University of Pittsburgh
- Company dedicated to increasing data access for research while protecting patient privacy
- Focused the privacy component, leave the data mining to others…

Context

- Felt a commercial approach would provide a better opportunity for "knowledge transfer" from academia to healthcare services
  - Most hospitals and healthcare service providers do not have informatics departments
- Commercial entity offers product management discipline, ongoing investment in product improvement, maintenance, service and accountability

De-ID Use Case: NCI/caBIG™

- *GOAL:* Allow multiple hospitals to contribute data to a mutually accessible data warehouse for further text processing.
- *SOLUTION:* De-ID is used to process reports within an organization to enable de-identified reports to be sent to the shared data repository
- *OUTCOME*: Consistent de-identification across all institutions creates a successful HIPAA compliant collaborative research environment

Seven Years Later…

- De-ID has now been in the "De-IDentification business" for seven years
  - Watched the marketplace change
  - Years 1-3 people struggled with HIPAA but it didn't interfere with research and discovery
  - Years 4-7 people struggled with the concept of free text as data
- Now, the value of the unstructured (text-based) medical records is generally accepted
  - Even if the capabilities to extract meaning are not fully realized

Unique Focus: Unstructured Data (Text-based)

- De-IDentifying structured data is not a huge challenge
  - Process data to not include the fields containing PHI
- Using patient records and reports as 'data' requires removal of PHI from the narrative
  - Overmark (redaction ala magic marker) and you've diminished the data value of record
  - Undermark: considerable risk at allowing PHI to remain in the record

Managing Expectations

- Societal expectations are dissonant with the reality of healthcare
  - Textbooks say 10% of appendectomies should be expected to have no pathology, but De-IDentification should be 100% accurate
- Manual De-IDentification runs 70%-90% reliability[1], but high variability and high costs
- The standards focus on minimizing risk of disclosure not whether any individual element of PHI is removed

[1]Douglass MM, Clifford GD, Reisner A, Moody GB, Mark RG: **Computer- assisted deidentification of free text in the MIMIC II database.** ***Computers in Cardiology 2004, 31:341-344***

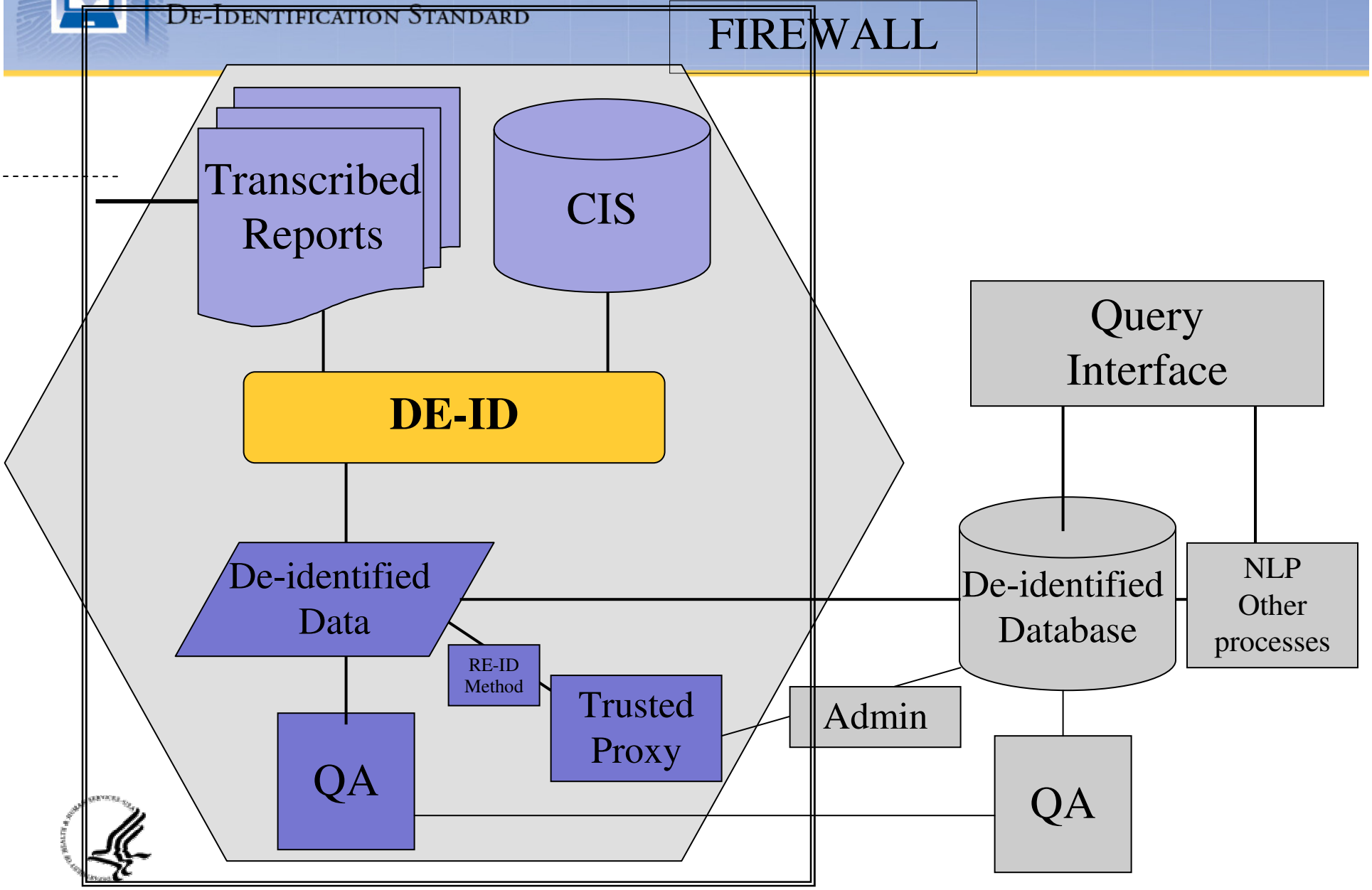
De-ID Package

- Software with reliable and valid baseline performance and consistent improvements in sensitivity and specificity
  - Sensitivity for names 100%; overall sensitivity 95%; specificity 89%
- Counsel on client-level quality improvement/oversight programs
- Focus on customer needs for data management workflow
  - Removes the burden of De-IDentification to allow focus on the research

What Does De-ID Do?

- De-ID creates a separate, De-IDentified version of a patient record or report
  - Outputs in text or XML
- De-ID does not "scrub" or remove text, but replaces text with specific tags in the form of proxies (e.g., for names) and offsets (e.g., for dates), preserving the data value of these elements and making them ready for data management and analytics.
- An optional linkage key can be generated for "re-identification" of the associated patient
- Capable of clustering De-IDentified records associated with the same patient

Generic Architecture



Example

Progress Note	DE-ID Progress Note
December 15, 2005. Rosalind Franklin is a 46 year old woman with a history of hypertension who was started on Norvasc on December 14 by Dr. Schwartz. She presented to the Hope Hospital ED today with a fine red pruritic maculopapular rash with no respiratory symptoms. Dr. Schwartz was called but unavailable, Dr. Lipton, covering advised D/C Norvasc and start Lisinopril 5mg	**DATE[Feb 14 2008]. **NAME[AAA BBB] is a **AGE[in 40s] year old woman with a history of hypertension who was started on Norvasc on **DATE[Feb 13] by Dr. **NAME[ZZZ]. She presented to the [**PLACE] today with a fine red pruritic maculopapular rash with no respiratory symptoms. Dr. **NAME[ZZZ] was called but unavailable, Dr. **NAME[YYY] covering advised D/C Norvasc and start Lisinopril 5mg

De-ID Metrics

- Published metrics of early version
  - Gupta, et al Am J Clin Pathol 2004;121:176-186
  - Results from three separate evaluations demonstrated marked ability to improve software performance on overmarking and undermarking
- Additional evaluations
  - As part of the product improvement process, two independent evaluations demonstrated
    - 100% specificity for names
    - Increases in phrase-level sensitivity (undermarking) from 95% to 98.5%
    - Increased in phrase-level specificity at 99.8%
    - Increases in report-level specificity (overmarking) from 85%-89%

De-ID Budget Considerations

- Manual De-IDentification;
  - reviewer salary, time, management , errors, variable costs
- Automated open source De-IDentification;
  - expensive to install, hardware, staff time, variable costs
- Automated commercial De-IDentification;
  - installation and ongoing technical support included in initial license fees, fixed costs

De-ID Use Cases

De-ID Use Case: DNA BioBank

- *GOAL:* How to combine genotype and phenotyping information for broad research purposes
- *SOLUTION:* Enable the phenotyping information available to the entire research project
- *OUTCOME*:

Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine

DM Roden[1–3], JM Pulley[4], MA Basford[1,4], GR Bernard[2,4], EW Clayton[5,6], JR Balser[3,4] and DR Masys[7]

Our objective was to develop a DNA biobank linked to phenotypic data derived from an electronic medical record (EMR) system. An "opt-out" model was implemented after significant review and revision. The plan included (i) development and maintenance of a de-identified mirror image of the EMR, namely, the "synthetic derivative" (SD) and (ii) DNA extracted from discarded blood samples and linked to the SD. Surveys of patients indicated general acceptance of the concept, with only a minority (~5%) opposing it. As a result, mechanisms to facilitate opt-out included publicity and revision of a standard "consent to treatment" form. Algorithms for sample handling and procedures for de-identification were developed and validated in order to ensure acceptable error rates (<0.3 and <0.1%, respectively). The rate of sample accrual is 700–900 samples/week. The advantages of this approach are the rate of sample acquisition and the diversity of phenotypes based on EMRs.

J Clin Pharm Therapeutics 84(3) 2008; 362-369

De-ID Use Case: Data Warehouse

- *GOAL:* Develop a leading-edge data warehouse to support patient operations and clinical research
- *SOLUTION:* De-ID installed in-line to process reports and feed a Web-based portal for all clinicians, administrators, and researchers
- *OUTCOME*: Real-time access to the hospitals historical records and reports by the entire professional staff, from bedside nurses to clinical researchers.

De-ID Use Case: State Database

- *GOAL:* Enhance access to state-level cancer database
- *De-ID SOLUTION:* De-ID offered a web-services schema directly linking the client's Java/SQL program to the De-ID web service.
- *OUTCOME*: Clinicians across the state can make a query via website and De-ID processes the data stream and creates a De-IDentified data table delivered via the users internet browser

De-ID Use Case: Clinical Trial Recruitment

- *GOAL:* Finding patients closely matched to clinical trial eligibility criteria
- *SOLUTION:* Create De-IDentified, HIPAA compliant versions of reports that could be searched and sorted for eligibility criteria. Once such records were found, De-ID's Linkage Key function could be used via trusted proxy to contact the physician responsible for that patient.
- *OUTCOME*: More accurate and rapid identification of eligible candidates for trials

Role of Commercial Providers

- Reliability and accountability for software performance
- Continuous product improvements
- Support for data management workflow
- Removes a barrier to data access and analysis without requiring investment of time and personnel against the HIPAA issue
- Costs include technical support

Value

- De-ID software has itself created 'de facto' standards in the marketplace
  - Consistency across institutions and data sets
  - Transparency of reliability and validity data
  - Simplicity of connectivity to data management systems
  - Basic performance to national standards while supporting customer capability to "tune" the software through dictionary management

Thank You

Dan Wasserstrom

Founder and Chairman

DE-ID Data Corp, LLC

dan@de-idata.com