Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research

Dilip Gupta, MD,[1] *Melissa Saul,*[2] *and John Gilbertson, MD*[1]

Key Words: Deidentification; Health Insurance Portability and Accountability Act; HIPAA; Safe-harbor elements; Confidentiality; Pathology reports

DOI: 10.1309/E6K33GBPE5C27FYU

Abstract

We evaluated a comprehensive deidentification engine at the University of Pittsburgh Medical Center (UPMC), Pittsburgh, PA, that uses a complex set of rules, dictionaries, pattern-matching algorithms, and the Unified Medical Language System to identify and replace identifying text in clinical reports while preserving medical information for sharing in research.

In our initial data set of 967 surgical pathology reports, the software did not suppress outside (103), UPMC (47), and non-UPMC (56) accession numbers; dates (7); names (9) or initials (25) of case pathologists; or hospital or laboratory names (46). In 150 reports, some clinical information was suppressed inadvertently (overmarking). The engine retained eponymic patient names, eg, Barrett and Gleason. In the second evaluation (1,000 reports), the software did not suppress outside (90) or UPMC (6) accession numbers or names (4) or initials (2) of case pathologists. In the third evaluation, the software removed names of patients, hospitals (297/300), pathologists (297/300), transcriptionists, residents and physicians, dates of procedures, and accession numbers (298/300).

By the end of the evaluation, the system was reliably and specifically removing safe-harbor identifiers and producing highly readable deidentified text without removing important clinical information. Collaboration between pathology domain experts and system developers and continuous quality assurance are needed to optimize ongoing deidentification processes.

In the sixth century BC, the Hippocratic oath made it very clear: "Whatever I shall see or hear in the course of my dealings with patients, it should not be published abroad, I will never divulge, holding such things to be holy secrets."[1] Twenty-six centuries and a scientific revolution later, in a world of distributed medical centers and extensive collaborative research, this Hippocratic principle has been formalized in 2 major bodies of legislation. The Health Insurance Portability and Accountability Act (HIPAA)[2] protects the confidentiality of patient information, while the "Common Rule"[3] protects the confidentiality and privacy of research subjects.

Access to clinical information, usually in the form of clinical documents, is fundamental to most areas of biomedical research. However, owing to HIPAA and Institutional Review Board (IRB) confidentiality concerns, research centers cannot share documents that identify patients. Under HIPAA guidelines, protected health information must not be disclosed for research purposes unless the patient grants authorization or the researcher obtains a waiver from the IRB, and each disclosure must be documented and available for review on patient request. These regulations regarding protected health information do not apply if data are deidentified. For clinical researchers to use free text information in a way that complies with standards of confidentiality, it is necessary to remove information that could identify individual patients. The HIPAA regulations require removal of the following patient identifiers: name, street address, city, county, zip code (with exceptions), dates (except year) directly related to an individual (eg, birth date, discharge date), telephone and fax numbers, e-mail addresses, Web uniform resource locators and Internet protocol addresses, social security numbers, account and medical record

numbers, health plan beneficiary numbers, certificate and license numbers, vehicle identifiers and serial numbers, device identifiers and serial numbers, biometric identifiers, full-face photographic images, and any other unique identifying number, characteristic, or code.

Deidentification of medical records involves 2 steps: (1) the identification of personally identifying references within medical text and (2) the masking, coding, and/or replacing of these references with values irreversible to unauthorized personnel.[4] Some computation methods have been described previously to achieve this goal in medical text documents.[5]

The Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, developed software to deidentify textual clinical reports, consistent with HIPAA, as part of the Integrated Advanced Information Management Systems (IAIMS) program. The IAIMS project is a National Institutes of Health–sponsored effort to help design and develop methods for integrating information across the health system. The resulting software, known as De-Id, works with clinical reports from the University of Pittsburgh Medical Center (UPMC; Pittsburgh, PA) report archive and MARS (MARS C Medical Archival System, Pittsburgh, PA) and deidentifies them, consistent with HIPAA.[6] Specifically, it replaces all information required to be removed by HIPAA, except that the system does not identify or remove full-face photographic images because it is designed to work on text-only databases. The program also replaces potential identifiers not included in the HIPAA safe-harbor elements, such as identifiers (eg, name, address, phone number) of health care providers (physician, health care worker, laboratories, and hospitals) and specimen identifiers (case numbers, outside accession numbers, and names of the referring hospitals). We sought to replace these identifiers, even though their removal is not specifically required under HIPAA, to decrease the possibility of potential patient identification and to protect providers should they or their practices be the subject of research projects.

De-Id's main process is to locate identifiable text, as defined by the safe-harbor method or the limited data set in the document in question. To do this, De-Id implements a set of rules and dictionaries designed to identify the presence of any of the 17 (full-face photographic images not included) HIPAA-specified identifiers within electronically stored medical text. Examples of these rules include the examination of document headers for patient and provider names, use of the Unified Medical Language System (UMLS) Metathesaurus for identification of medical phrases (such as Gleason score) to retain in the document, pattern matching of numeric text to detect phone numbers and zip codes, the US Census dictionary to aid in the identification of names, and a variety of user-customizable dictionaries for identifiers and health care providers unique to an institution.[7] De-Id replaces identifiable text with deidentified but specific tags. Identifiers found multiple times in the report are replaced consistently with the same tag to improve readability of the report. Dates are replaced by tags, but date tags from 2 different dates retain the time interval between them. An example of a surgical pathology report deidentified by De-Id is given in **Appendix 1**.

There had been an initial, limited evaluation of the software's performance on a variety of clinical documents (history and physical examination reports, operative notes, discharge summaries, and progress notes). However, this initial evaluation did not include pathology reports.

Surgical pathology reports have a unique place in research among clinical documents, as information in pathology reports tends to retain its value much longer than that in clinical notes, radiology reports, and discharge summaries. Diagnoses confirmed by pathology form the basis for most clinical and research studies on disease, and the pathology laboratories retain tissue blocks indefinitely, so new research can be done on very old cases. Many retrospective studies involving pathology specimens and archival paraffin blocks require only deidentified patient information, and it is cumbersome to obtain specific consent from patients for new research on specimens obtained during surgery years ago. Furthermore, pathology reports are highly formatted, so it should be possible to identify specific sections in pathology reports that are associated with high research value but low risk of patient identification. Most pathology laboratories contain many years of electronically archived pathology reports containing diagnostic, demographic, and clinical information, but these reports tend to be free text and contain patient identifiers.

Before De-Id was used on pathology reports, the UPMC Department of Pathology requested an extensive evaluation of the system by pathologists. The study was to be limited to surgical pathology but involve reports from various periods. The evaluation would test the system's ability to completely deidentify reports and retain all important clinical information.

We describe the evaluation performed by the UPMC Centers for Pathology Informatics and Oncology Informatics and Clinical Research Informatics Service (CRIS). The evaluation required 3 cycles. At each cycle, from 300 to 1,000 reports were examined, and limitations in De-Id were identified and fixed. By the end of the study, the deidentification engine was successfully deidentifying pathology reports, removing HIPAA safe-harbor identifiers but overmarking no important clinical information. We provide information about a useful model for deidentification engines at other institutions and describe the unique deidentification challenges presented by pathology reports.

DOI: 10.1309/E6K33GBPE5C27FYU

Materials and Methods

Textual surgical pathology reports were selected randomly by the CRIS from the MARS archive (the central archive of textual clinical reports at UPMC) and processed by the De-Id engine. The software's mechanism is explained in the introduction. Briefly, the software locates identifiable text, implements a set of rules and dictionaries designed to identify the presence of any of the 17 HIPAA-specified identifiers, uses the UMLS Metathesaurus for identification of medical phrases, and replaces identifiable text with deidentified but specific tags. The study considered only surgical pathology reports and did not include autopsy, cytology, or special study reports unless they were part of a surgical pathology report. We did not consider reports with biometric identifiers, including fingerprints and voiceprints or full-face photographic images or any comparable images. The deidentified reports were sent to the Center for Pathology Informatics for manual evaluation. An encrypted linkage file with a counter for each case was created. It was stored on a secure server available only to CRIS. The linkage file contained the counter plus medical record number and unique document number as stored in MARS. CRIS held the encrypted linkage file until the end of the evaluation, at which time the file was destroyed.

At the Center for Pathology Informatics, the reports were distributed for evaluation on paper to 4 pathologists (including D.G. and J.G.) with training in pathology and informatics. The reports were at least 2 years old. Evaluation was guided by a formal training session and written "Instructions to Reviewers" **Appendix 2**. Briefly, the reviewers were asked to examine each report twice. First, they were to find the text that should have been deidentified but was not (eg, underdeidentified or undermarked text). They were to circle each error and classify the type of identifier that had been missed by the software. The classes included P (patient identifiers), D (dates), M (provider or institution), and S (specimen identifiers). Patient identifiers could include items such as name, address, city, state, zip code, phone number, medical record number, social security number, employer, job title, e-mail address, account numbers, and names of relatives.

After reading the report, the reviewer provided an assessment of the extent of deidentification. Specifically, the reviewer was asked to estimate the following probability: "What is the probability that the identity of the patient could be determined by the researcher (who has no previous information on this case) from the remaining information in this report? If 100 clinical researchers at UPMC read this record, assess the expected number (n) of them who would correctly identify the patient from just the information in the record." The standard we used was that the researcher would have access to clinical archive (MARS) and laboratory information services (CoPath Plus, Cerner, Kansas City, MO) and could do simple searches (search by number or name) but could not perform complex or advanced searches (search or match text strings).

After evaluating the undermarking errors, reviewers reread the report, searching for instances in which the program inadvertently removed clinical text (overmarking) as if it were identifying text. Overmarked text was found by the presence of tag and by context. Reviewers wrote their opinions about what had been overmarked. Finally, they rated the importance of the suppressed text for understanding the report and added any comments they thought appropriate.

After the initial review, the marked-up reports were sent to the principal investigators (D.G. and J.G.) who read each report, examined the markups, and summarized the results in an Excel spreadsheet (Microsoft, Redmond, WA). Because the principal investigators rereviewed each report, the summation acted as a quality control. After summation of results, a document was prepared that outlined the successes and failures of the system, including the general classes of problems identified by the research. The results were discussed with the De-Id development team, which used the results to improve the system's performance. In general, the results identified situations that had not been anticipated by the developers; minor software engineering was required to rectify the errors.

The deidentification engine was evaluated 3 times. Problems identified at each pass were discussed with the CRIS team between evaluations to improve deidentification of pathology reports. A new set of 1,000 reports (250 reports each from the years 1985, 1990, 1995, and 2000) was reviewed in a similar manner in the second evaluation. A mechanism to "scrub" the non-UPMC (outside) accession number was developed. An additional 300 reports were evaluated in the third evaluation.

Results

The main results of each evaluation are summarized in **Table 1**.

First Evaluation

Our initial data set included 967 surgical pathology reports (CoPath Plus) for the year 2000, chosen at random from hospitals of the UPMC network. A sample deidentified surgical pathology report is shown in Appendix 1. These data represent approximately 1% of the surgical pathology reports during the year 2000. The reports ranged from "gross only" to complex resections that included multiple special procedures.

DOI: 10.1309/E6K33GBPE5C27FYU

Table 1
Important Findings in Each Evaluation of a Deidentification Software Engine*

	First Evaluation (n = 967; year 2000)	Second Evaluation (n = 1,000; 250 each for years 1985, 1990, 1995, and 2000)	Third Evaluation (n = 300; year 2000)
Undermarking errors			
Accession numbers	103 (10.7)	96 (9.60)	2 (0.7) abbreviated numbers
UPMC	47 (4.9)	6 (0.60) in the comment section	—
Non-UPMC		90 (9.00)	—
With outside hospital information	27 (2.8)	—	—
With no outside hospital information	29 (3.0)	—	2 (0.7)
Dates	7 (0.7)	0 (0.00)	0 (0.0)
Physician information			
Pathologist name	32 (3.3)	4 (0.40)	3 (1.0)
Clinician name	2 (0.2)	0 (0.00)	0 (0.0)
Pathologist initials	25 (2.6)	2 (0.2)	0 (0.0)
Hospital name	46 (4.8)	0 (0.00)	3 (1.0)
Overmarking errors	150 (15.5)	0 (0.00)	0 (0.0)
Block label interpreted as address	50 (5.2)	0 (0.00)	0 (0.0)

UPMC, University of Pittsburgh Medical Center, Pittsburgh, PA.
* Data are given as number (percentage) of total reports evaluated.

Undermarking Errors

Undermarking errors were identified in 4 main areas: accession numbers, dates, provider names, and hospital names.

1. Accession numbers: This area was the dominant undermarking error in the first evaluation. Pathology laboratories use the accession number, generated by the laboratory information system, as the main identifier for specimens and reports. The program removed all accession numbers in the report headers, but it is not uncommon to have accession numbers embedded in the text in a variety of places. In the initial data set of 967 reports, 103 (10.7%) contained embedded accession numbers. De-Id had not attempted to suppress the embedded numbers because the software developers did not understand the nature and importance of accession numbers before the evaluation.

As specimen identifiers, the accession numbers, although not mentioned specifically in the HIPAA safe-harbor list of identifiers, are associated with a strong potential for patient identification. As a practical matter, the effectiveness of an accession number in identifying a patient depends on the origin of the accession number and the other information available in the report. Our study identified 3 specific situations: (1) UPMC accession numbers: Of the 967 reports examined, 47 reports (4.8%) contained unmasked UPMC accession numbers. This was considered a serious breach of confidentiality because virtually anyone at UPMC with access to the clinical or laboratory information systems could identify a patient by using this number. (2) Non-UPMC accession numbers with information on the outside hospital: Slides and specimens often are sent to outside hospitals and consultants for second opinions. The consulting pathologist often records the accession number and name of the outside hospital as part of the second-opinion report. Of the 967 reports, 27 (2.8%) included the "outside accession number" and information about the hospital that provided it. This was considered a substantial risk for identification, depending on the amount of information available about the outside hospital. (3) Non-UPMC accession numbers with no information on the outside hospital: 29 reports (3.0%) included an "orphan" outside accession number (accession number without the name of the hospital or laboratory that generated it). Our reviewers considered this a less substantial risk for identification.

2. Dates: The system was very effective in suppressing dates in reports. Of 967 reports, only 7 disclosed dates. All of the errors occurred in the "Clinical History" part of the report, and the majority were the date of the last menstrual period. None of the dates posed any threat of patient identification. The system blocked all encounter and procedure dates and all patient ages except one (see "Other Undermarking Errors").

3. Provider names: Although removal of the names is not required by HIPAA, De-Id replaces them. The evaluation turned up 5 distinct types of provider undermarkings: (1) name of case pathologist; (2) name of consulting, advising, or frozen section pathologist (or resident); (3) name of clinician; (4) pathologist's initials; and (5) initials of the transcriptionist. Of these, the initials of the transcriptionist were by far the most common but the least important. At UPMC, one can learn virtually nothing about patients from the initials of the transcriptionists because UPMC has a central transcription system. Pathologists' initials also were considered of little risk. We identified 25 reports (2.6%) in which pathologists' initials were not suppressed.

DOI: 10.1309/E6K33GBPE5C27FYU

Only 2 reports disclosed the identity of clinicians. The engine failed when the names of the clinicians were hidden in free text of the clinical history section. In one of the reports, 2 clinicians were identified ("Drs X and Y performed the delivery"). The De-Id system was not searching for the construct "Drs X and Y."

In 9 reports, the case pathologist's name was not suppressed. However, these 9 reports involved only 2 pathologists. The names appeared (unsuppressed) at the signature section of the report at the end of final diagnosis. Both were dental pathologists (with the DDS credential rather than MD) and were members of the dental school (not the medical school) faculty. The system had not included dental faculty in the dictionaries, nor had it been programmed to check for a DDS credential. Both limitations were rectified.

In an additional 23 reports (2.4%), pathologists were named. In 2 reports, this occurred in the "Comment" section in the form of "Drs XX and YY concur." In remaining reports, this occurred in the "Intraoperative Consultation" section. The frozen section diagnosis line mentioned "Drs J.Smith/A.Jones" (with no spaces between the first initial and last name or between the slash and the names). This does not seem to be a serious confidentiality problem because the pathologists were not, in general, the formal sign-out pathologists on the case and tended not to be associated with the case in the laboratory information system. We do not believe that a pathologist's name poses a substantial risk to patient confidentiality in these contexts.

4. Hospital names: Of the 967 reports, 46 (4.8%) mentioned hospital or laboratory names. Of these reports, 22 involved UPMC hospitals. The risk posed by these identifications varied depending on the presence or absence of an associated accession number.

Other Undermarking Errors

There were 5 reports that we believed posed a risk of identification owing to uncommon identifiers or unique clinical histories. As shown in **Table 2**, The reports included 1 medical device with an identification number, 1 research protocol number, information about a Bosnian war veteran, and information about a person with a gunshot wound that might have been discussed on the local news. Currently, no reliable mechanism is available to identify and remove such unique combinations. The most interesting problem involved a 91-year-old patient. The De-Id system would have picked this up and reported **AGE<in 90s>-year-old, but the word "year" was misspelled as "yea." To prevent this occurrence in the future, the word "yea" was included in the system's age-finding rule because it was thought that the pattern "## yea" in a medical document most likely would represent a misspelling of the word *year.*

Overmarking Errors

Overmarking errors occurred when the system eliminated clinical information, interpreting it as an identifier. The number of overmarking errors was small and seemed to have minimal impact on the interpretation of the reports. The output from the system was very readable. Of 967 reports, 150 included overmarking. In 50 reports, a block label was interpreted by the program as a postal address. Another frequent problem was that the term H&E often was replaced by H\T\E. This affected the readability of the reports. A more serious problem was the infrequent removal of special stain or immunohistologic stain information.

The evaluators were asked to score each overmark from 1 (not important) to 10 (report unreadable). Of the information from the reports, approximately 99% was intact. **Table 3** indicates the overmarks that received the worst scores (a score of 5 or higher). The engine retained eponymic patient names, such as Barrett (esophagitis), Raynaud (hypothyroidism), Castleman (lymphadenopathy), de Quervain (tendinitis), and Gleason (prostate cancer grade). Interestingly, the program also retained the names of authors in references quoted in the surgical pathology report.

Second Evaluation

We were able to suggest specific changes that made the deidentification engine more competent in the specific domain of surgical pathology reports. We identified the areas in which the program failed. A new set of 1,000 reports (250 reports each from the years 1985, 1990, 1995, and 2000) was

Table 2
Reports Posing a Risk of Patient Identification Owing to Uncommon Identifiers or Unique Clinical History

Section of Pathology Report	Text With Identification Risk
Gross	Quantum-II Model 254-xx SSI-SST-500 (device name and number)
Multiple	OSS SYS99-3573; UPCI protocol 98-056 (research protocol number)
History	The patient is a 91- yea-old male who presents with no given clinical history.
History	Age<in 30s> year old male with a history of having been hit by shrapnel in the Bosnian war, underwent craniotomy for removal of methylmethacrylate skull plate
Multiple	**AGE<>-year-old black male status post gunshot wound to left hip

DOI: 10.1309/E6K33GBPE5C27FYU

Table 3
Overmarking Errors*

Score	Section of Pathology Report	Extract	Probable Text Suppressed
5	Comment	Positive for **INITIALS and cytokeratin 7....	EMA? CK?
5	Final diagnosis	**name<> of Gram and Grocott stains to follow	Results?
5	Gross	Labeled **NAME<> portion of left tube	Distal?
5	History	**NAME<> grade 3 + 3 = 6	Gleason?
7	Comment	The tumor is positive for ... **INITIALS<>, receptors	ER? PR?
7	Multiple	Neoplastic DNA **ADDRESS HI, *Hin*dIII and BglI; A **NAME<> Stain...	?
8	Final diagnosis	Within soft tissues **ADDRESS IN AREA (S)...	?
5	Intraoperative	Lymphadenitis, **name<> carcinoma	No evidence of carcinoma

CK, cytokeratin; EMA, epithelial membrane antigen; ER, estrogen receptor; PR, progesterone receptor.
* *Overmarking* is the inadvertent removal of clinical text as if it were identifying text. Overmarks were scored on a scale of 1 (not important) to 10 (report unreadable).

reviewed in a similar manner in the second evaluation. Our findings were as follows:

1. There was no difference in the ability of the De-Id program to deidentify the reports from different time frames (ie, 1985, 1990, 1995, 2000).

2. The software did not hide any important clinical information. It retained terms such as Barrett esophagus or Hashimoto thyroiditis. The term "Hodgkin's" was overmarked in a diagnosis of lymphoma in 1 case.

3. All patient identifiers, such as name, address, phone number, unique identifier numbers (eg, social security numbers), birth dates, and e-mail addresses, were masked, except in 1 case in which a device number was visible but the device type had been suppressed.

4. In 6 reports, the software did not hide previous UPMC accession numbers in the "Comment" section (scattered within descriptive text) of the report.

5. The program continued to struggle with outside accession numbers. At the time of testing, the developers were still evaluating an algorithm that would reliably remove accession numbers without overmarking laboratory values or numeric data. Of 1,000 reports, 90 (9.00%) displayed outside accession numbers. Although outside accession numbers were still not replaced, the program removed the names of hospitals or contributing private groups, addresses, and zip codes. In 1 report, the name of the outside hospital was mentioned at another place not in relationship with slide accession numbers, and the program failed to suppress that.

6. The initials of pathologists (2 reports) or their names (4 reports) were found in the frozen section diagnoses. These pathologists were not the sign-out pathologists, so their names were not in the headers of the reports.

7. In one case, the name of a specific research protocol and the school (Graduate School of Public Health) was mentioned, but there was no patient identifier.

The system had improved substantially between the first and second evaluations. The program did not remove important clinical information. Except for accession numbers, the system performed at a high level of specificity and reliability. However, a third evaluation was required to document progress on the accession numbers.

Third Evaluation

The third evaluation involved 300 reports from the year 2000. The system performed extremely well. The program replaced the names and other identifiers of the patients and the dates of procedures. The program also replaced names and initials of transcriptionists, residents, and physicians. In 3 reports, the names of pathologists associated with frozen sections were displayed (see "Second Evaluation"), and in 3 reports, the name of an outside hospital or laboratory was retained in the clinical history or description of the consult material. In 2 reports, the program failed to replace the name of a clinical trial.

With its new algorithm, the system handled accession numbers extremely well. Wherever full accession numbers were seen in reports, they were suppressed by the system. In 2 reports, abbreviated accession numbers (eg, "9829" as opposed to "S99-9829") were present and were not suppressed.

The De-Id engine reliably and effectively deidentified patient, provider, and specimen information in surgical pathology reports. Since the end of the evaluation, De-Id has been used to deidentify more than 35,000 pathology reports, and minor problems continue to be identified and fixed.

Discussion

The confidentiality of protected health information is an important component of medicine. It is becoming increasingly important and difficult to satisfy the concerns of patient confidentiality and biomedical research. Access to pathology documents lies at the center of these legitimate but contending concerns. Generated in the care of individual patients, pathology reports contain highly identifiable text

DOI: 10.1309/E6K33GBPE5C27FYU

and some of the most basic data needed for clinical research. In an attempt to provide researchers the information they need without violating confidentiality, there has been extensive research toward the development of automated systems that can reliably deidentify textual clinical reports.

Thomas et al[8] described a method for removing names in pathology reports by using an augmented search and replace method, taking advantage of the fact that the vast majority of proper names in pathology reports occur in pairs. Their team created a clinical and common usage word list and used substitution methods. Their method found 98.7% of 231 proper names in the narrative sections of pathology reports. Three single proper names were missed in 1,001 pathology reports (0.30%). Unfortunately, the system was limited to the removal of patient names. Other identifiers, such as those described by HIPAA, remained in the medical documents. Mechanisms relying on scrubbing have been described by Sweeney[9] and Berman.[5]

In contrast with many pilot and developmental deidentification mechanisms and algorithms described by others, the De-Id engine attempts to comprehensively remove all HIPAA safe-harbor identifiers except photographic images. Furthermore, it keeps the medical documents readable by leaving specific tags, so readers know the type of information that was replaced. This mechanism also makes it possible to request the information that was replaced by the De-Id engine. When a person, location, date, or specimen is encountered multiple times, the same replacement tag is used, increasing continuity and readability of the medical text, and if the report contains multiple dates, the date tags retain the intervals between the dates.

Another important aspect of De-Id is its clinical scope. The engine was designed to work with archives of all types of clinical documents. At UPMC, De-Id is implemented to work with electronic archives of millions of clinical documents collected during 28 years from more than 200 clinical applications and 18 hospitals. De-Id is, therefore, capable of handling the complexity of a multihospital environment and, more importantly, designed to deidentify a wide variety of textual medical documents from surgical reports to SOAP (subjective, objective, assessment, patient care plan) notes, discharge summaries, and radiology and pathology reports. Because of this, variations in formatting of surgical pathology reports across institutions should not negatively affect the program's performance. For example, some institutions use microscopic description paragraphs, and others do not. Similarly, even though we have tested the software for the reports generated in the UPMC electronic archive, MARS, we believe the software should be able to function with the text reports of any other laboratory or hospital information system, provided the information to be deidentified is fed in an ASCII (American Standard Code for Information Interchange) text format and appropriate provider dictionaries are available.

De-Id is managed by the CRIS, a unit of experienced data analysts with special training in patient confidentiality. As part of its operation, De-Id creates an encrypted linkage file that ties the deidentified documents to the suppressed identifiers. CRIS is the sole custodian of the encrypted linkage file and, therefore, operates as an honest broker service with access to the original, identified reports, if necessary. The University of Pittsburgh IRB and the UPMC-Health System Data Security Committee have endorsed the use of this software tool for providing deidentified reports (http://www.oorhs.pitt.edu/clinicalresearch/index.cfm?location=2.1). The combination of MARS, De-Id, and CRIS is an effective mechanism to provide deidentified clinical information to researchers in an honest broker environment. If the IRB protocol does not include permission for reidentification, the linkage file is destroyed, creating an anonymized data set.

Our assessment of deidentification went beyond a mechanical counting of missed HIPAA safe-harbor identifiers. It included potential identifiers not included in safe harbor, such as names of health care providers (physicians, laboratories, and hospitals), employers, and relatives, and assumed a researcher with standard access to UPMC clinical systems. It also tested the readability and completeness of the deidentified documents. In general, this was an extensive test of the De-Id engine. Given the complexity of the reports, the De-Id program seemed to be an efficient tool to deidentify free text in pathology reports for research, while maintaining patient confidentiality and clinical information. The study indicates that it is possible to automatically deidentify complex textual clinical documents.

The purpose of this article, however, is not to laud the capabilities of De-Id but rather to document the importance of testing and quality assurance in deidentification systems. In initial, limited validation studies by the developers, which did not include pathology reports (or pathologists), De-Id seemed to be an effective mechanism for deidentifying clinical documents; however, the first phase of the evaluation by our team of pathologists revealed a number of serious limitations.

The limitations existed for 2 main reasons: the developers were not familiar with the domain of pathology and some of its nuances, and the system had not been tested extensively on thousands of reports. In particular, the developers did not fully understand the important role of accession numbers in the identification of patients in pathology. Similarly, the system did not deidentify some dental pathologists because the system was not expecting dentists from the dental school to be case pathologists on

DOI: 10.1309/E6K33GBPE5C27FYU

medical cases. This problem was easily resolved by adding to a dictionary (of medical degrees and medical and dental school faculty) in the system. Handling accession numbers proved to be a more difficult problem because of their varied formats and resemblance to clinical data in other report types. However, by the end of the third evaluation, the system was handling accession numbers well. However, we recently detected occasional accession numbers passing undetected through the system, especially those with a limited number of digits (eg, S98-09) that can be confused with laboratory values, immunoperoxidase stains, and other clinical data. Other deidentification errors could be attributed to the fact that this was the first large-scale test of the De-Id system. For example, the initial display of hospital names could be traced to an "immature" rule and a limited dictionary.

Not surprisingly, the De-Id program failed when there was typographic or spelling error. In 1 case, the age of the patient (91 years) was not removed because the word "year" was misspelled as "yea." In addition, there were 5 reports with unique medical or social history, sequence of events, or combinations thereof. Currently, there is no good mechanism for any deidentification system to identify such unique combinations or sequences of events in textual information that has the potential for patient identification. For example, the physician's note might describe a 46-year-old man with Addison disease who received a fatal gunshot wound to the head.[10] It is possible that people could identify this person despite the lack of identifiers such as name, address, or phone number.

Deidentification might not produce perfectly anonymized data that qualifies for release without patient authorization or might not meet the needs of biomedical researchers in these rare occurrences.[11] Similarly, the system deidentifies information by removing explicit identifiers, such as name, address, and phone number, and replacing them with made-up alternatives. It cannot guarantee anonymity if the information is manipulated, matched, or linked to external databases to identify any individual.[12,13]

Even though the De-Id program correctly deidentifies information most of the time, it makes undermarking and overmarking errors. In the foreseeable future, it is unlikely that any computer-based deidentification program will be perfect. A requirement for IRB approval and undertaking that the researchers will not attempt to link the information to identify patients seems reasonable. Specifically, patients with unique combinations of events, diseases, or medical history can present a difficult problem. It is difficult to have a perfect balance between a genuine research need and confidentiality of patients in such unique cases.

In 1 case, the name of a specific research protocol was not suppressed, but there was no patient identifier. It is possible that keeping the names of clinical trials would not substantially compromise patient confidentiality, and it is possible that such information might be useful to researchers who want to include information about patients who have participated in specific trials. Clinical trials might turn up information that could be beneficial to a patient enrolled in the study. In those cases, it would be impossible to identify the people involved if the data had been anonymized. Loss of this possible benefit for research subjects is an example of the "practical, scientific, and ethical problems"[14] associated with deidentification and anonymization.

The downside of applying De-Id is the inadvertent removal of clinical information interpreted as identifiers. Most of the inappropriately deidentified text (overmarkings) involved overlap between medical terms and address information (eg, the "MI" in Lansing, MI, vs an abbreviation for myocardial infarction) and patient names that are also medical terms but not included in the UMLS (eg, Hickman catheter). Similar problems would occur if the patient's name were the same as the medical term (eg, Mr Gleason who has prostate cancer). As these problems are identified, De-Id is augmented to address them.

The De-Id program occasionally removed small amounts of clinical information during the deidentification process. We usually were able to find such errors because of context and the irrelevance of the tags to adjacent text. Researchers using deidentified reports who are not so familiar with our conventions might find it difficult to "fill in the blanks." If the surrounding text suggests that the missing information is important to the research study, the researcher can submit a request to the honest broker (CRIS) for the missing information. It is possible that there may be a systematic underestimation of overmarking errors because of the familiarity of evaluating pathologists with local conventions in this study. Because the evaluators did not have access to the actual, identified reports, other important items might have been missed.

The De-Id engine has a complete UMLS library (terms and concept unique identifiers) and searches for UMLS terms. It also has a list of stop words. Therefore, it is possible to modify the De-Id engine as an autocoder that would replace UMLS terms with UMLS concept unique identifiers (or preferred terms). This feature could be used to further deidentify data and standardize terminology in reports between institutions.

By the end of the third evaluation, the De-Id program was specifically and reliably deidentifying patient, specimen, date, and provider and hospital identifiers, while maintaining a complete and highly readable clinical report. The UPMC Center for Pathology Informatics has continued to monitor the output. However, it is important to remember that De-Id is a complex, rule-based software program that uses multiple

DOI: 10.1309/E6K33GBPE5C27FYU

dictionaries and UMLS, operates on a large and varied collection of complex clinical reports from multiple hospitals, and is designed to work over a long period. Changes in the rules or dictionaries could cause unexpected changes in performance. Seemingly unrelated changes to the De-Id configuration have, in fact, caused a failure in the accession number algorithm after the evaluation was completed. This problem was identified and corrected quickly, but it underscores the importance of ongoing quality assurance systems in rule-based deidentification software.

Collaboration between pathology domain experts and system developers and continuous quality assurance were needed to optimize the performance of De-Id. Some quality assurance processes could be automated. For example, one could run the same 10,000 reports through the system each month and look for discrepancies in the output. However, because many of the errors identified in this project resulted from domain nuances (eg, accession numbers) and human errors (eg, spelling) not expected by the developers and because some of the problems in the future will result from changes in domain constructs and report content, there will remain a requirement for pathologists to formally (and manually) evaluate the output of deidentification engines. Finally, it should be evident that quality control consumes resources and, hence, represents a cost to the organization performing it. This cost, like the costs of hardware, developers, and honest brokers, should be calculated in the cost of deidentification and, by extension, clinical research in general.

To our knowledge, there are no reports of quality assurance mechanisms for complex deidentification engines in the literature. This article describes the initial quality assurance evaluation by the UPMC Center for Pathology Informatics and CRIS of the De-Id engine in the domain of surgical pathology reports. By the end of the study, the deidentification engine was successfully deidentifying surgical pathology reports, removing HIPAA safe-harbor identifiers but overmarking no important clinical information. This article describes a useful model for other testing and quality control of deidentification engines at other institutions and for identifying the unique challenges presented by pathology reports.

From the [1]Centers for Pathology Informatics and Oncology Informatics, Department of Pathology, University of Pittsburgh Medical Center–Presbyterian Shadyside; and [2]Clinical Research Informatics Service, Center for Biomedical Informatics, School of the Health Sciences, University of Pittsburgh, Pittsburgh, PA.

Supported in part by Integrated Advanced Information Management Systems grant 1-G08-LM06625 from the National Library of Medicine to the University of Pittsburgh.

Address reprint requests to Dr Gilbertson: Dept of Pathology and Center for Pathology Informatics, University of Pittsburgh, 5150 Center Ave, Pittsburgh, PA 15232.

Appendix 1
An Example of a Deidentified Surgical Pathology Report

E_O_R
S_O_H
Counters Record Type
228,228 SP
E_O_H
[Record deidentified by De-Id v 3.3]*

PATIENT HISTORY

The patient is a **AGE<in 60s>-year-old male with elevated PSA levels.
OSS **SLIDE-NUMBER 12/00, **PLACE
PRE-OP DIAGNOSIS: Elevated PSA
POST-OP DIAGNOSIS: Same
PROCEDURE: Prostate biopsies
1. Left apex.
2. Left body.
bjs

FINAL DIAGNOSIS

PART 1: PROSTATE, LEFT APEX, NEEDLE BIOPSY (OSS **SLIDE-NUMBER 12/00)
A. INVASIVE MODERATELY DIFFERENTIATED PROSTATIC ADENOCARCINOMA WITH A COMBINED GLEASON SCORE OF 3 + 3 = 6.
B. THE CARCINOMA INVOLVES ONE OUT TWO (1/2) CORE FRAGMENTS AND COMPRISES APPROXIMATELY 5% OF THE PROSTATE TISSUE EXAMINED.
C. NO EVIDENCE OF PERINEURAL INVASION IS SEEN.
PART 2: PROSTATE, LEFT BODY, NEEDLE BIOPSY (OSS **SLIDE-NUMBER 12/00) BENIGN PROSTATE TISSUE WITH NO EVIDENCE OF HIGH GRADE PROSTATIC NEITHER INTRAEPITHELIAL NEOPLASIA NOR CARCINOMA SEEN.
mb **INITIALS<QQQ/QQQ>

COMMENT

All the foci of prostatic carcinoma found small in size and constitute less than 5% of the material submitted.
Mb **NAME<VVV: **NAME<WWW Q. XXX>, MD, PhD
Fellow/Chief Resident: **NAME<UUU Q. TTT>, MD
** **NAME<SSS RRR QQQ PPP **
OOO VVV: **NAME<WWW Q. XXX>, MD, PhD
**DATE<6/25/00> 11:50
OUTSIDE ACCESSION **SLIDE-NUMBER 8 CONSULT SLIDES **SLIDE-NUMBER 8 CONSULT BLOCKS OUTSIDE **NAME<SSS> RECEIVED: Y
CONSULT MATERIAL DESCRIPTION:
Received for consultation from **NAME<YYY Q. ZZZ>, DO, are eight (8) consult slides labeled **SLIDE-NUMBER and eight (8) consult blocks labeled **SLIDE-NUMBER from **PLACE, **ADDRESS, PA along with an accompanying surgical pathology report.
bjs

MICROSCOPIC

Microscopic examination substantiates the above diagnosis.
mb
TC1 BC1

BC, billing code; bjs, initials of transcriptionist; counter, serial case number (given by the program to the cases requested by researchers, satisfying certain criteria); DO, doctor of osteopathy; E_O_H, end of header; E_O_R, end of report; mb (also Mb), initials of the pathologist; OSS, outside slide accession number; POST-OP, postoperative; PRE-OP, preoperative; PSA, prostate-specific antigen; S_O_H, start of header; SP, surgical pathology; TC, tissue code.
* De-Id version 3.31, Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA.

DOI: 10.1309/E6K33GBPE5C27FYU

Appendix 2
Instructions to Reviewers*

Purpose

The purposes of this study are to evaluate the ability of the De-Id computer program to (1) deidentify pathology reports and (2) retain necessary clinical information in those reports.

Instructions

We will be providing you textual pathology reports from MARS (MARS C Medical Archival System, Pittsburgh, PA) that have been processed by the De-Id computer program in an attempt to remove information that could reveal the identity of patients. Please evaluate each report, using the following steps:

1. Read the entire report looking for patient or specimen identifiers:
   a. While reading the report, please underline in red ink and label with a "P" any text that reveals (or helps reveal) the identity of the patient. This could include NAME, ADDRESS, CITY, STATE, ZIP CODE, PHONE NUMBER, MEDICAL RECORD NUMBER, SOCIAL SECURITY NUMBER, EMPLOYER, JOB TITLE, EMAIL ADDRESS, ACCOUNT NUMBERS, RELATIVES, ETC.
   b. While reading the report, please underline in red ink and label with a "D" any dates that could identify a patient or specimen. This includes Birth Date or Accession Date. NOTE: we are looking for dates that include a date/month/year or a month/year. Mention of a year (without a month or day) should not be flagged as this has been deemed not to identify a patient. Similarly, the mention of a time period not associated with a specific day/month/year or month/year (for example, "38 days after the operation" should not be flagged as this has been deemed not to identify a patient. THE PATIENT'S AGE IN YEARS AND HIS OR HER BIRTH YEAR DO NOT IDENTIFY A PATIENT unless the patient is more than 90 years old.
   c. While reading the report, please underline in red ink and label with an "M" any information, including Name, Address, web site, Phone Number, etc, of any physician, health care worker, or institution associated with the patient's care.
   d. While reading the report, please underline in red ink and label with an "S" any text that identifies the specimen, including outside accession numbers and referring hospitals.
2. Read the entire report looking for clinical information suppressed by the deidentification program:
   a. Please underline in red ink any clinical information (other than dates and patient age) that you believe was replaced by the De-Id program. Writing next to the underlined token, please provide a description of the type of clinical information that you believe is missing. In addition, rate from 1 (least) to 10 (most) the importance of this suppressed information in understanding the message of the report if you feel that you can make such a judgment. For example, if you think that "S100 antigen" was suppressed to "XXX antigen" and you think this could seriously impact the meaning of the report, circle the XXX and write "S100, 9."
3. After reading and annotating the record, please provide an assessment of the following probability:
   a. Assume that a clinical researcher does not previously know the medical history of the patient described in this deidentified report, with the exception of knowing the patient's gender and race and knowing that the patient was seen at UPMC during 2000. What is your probability that the identity of the patient could be determined by the researcher (who has no previous information on this case) from the remaining information in this report? In answering this question, it might be useful to consider that 100 clinical researchers at UPMC had read this record, and then assess the expected number (n) of them who would correctly identify the patient from just the information in the record. The fraction n/100 is the probability being assessed.
   b. In giving your assessment, please circle one of the following probabilities provided on the patient-record assessment form:
   0 to 1/1,000 (expect that up to 1 researcher in 1,000 would be able to identify this patient)
   1/1,000 to 1/100
   1/100 to 10/100
   10/100 to 50/100
   50/100 to 90/100
   90/100 to 100/100
   (Technical note: Each probability range listed above as "x to y" means [x, y], except when y = 1, in which case it means [x, y].)
   Alternatively, you may provide a numerical probability (from 0 to 1, inclusive) in the line provided at the end of the report.
4. At the end of the report, please provide any comments that may be helpful in explaining your probability assessment.
5. At the end of the report, please feel free to write any additional comments.

UPMC, University of Pittsburgh Medical Center, Pittsburgh, PA.
* De-Id, software engine, Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA.

Acknowledgments: We thank James Harrison, MD, PhD, and Rebecca Crawley, MD, for reviewing the reports; Gregory Cooper, MD, PhD, for input regarding the study design; Paul Hanbury for programming assistance; and Julie Nys for secretarial assistance.

References

1. Edelstein L. *The Hippocratic Oath: Text, Translation, and Interpretation From the Greek*. Baltimore, MD: Johns Hopkins Press; 1943.
2. Standards for privacy of individually identifiable health information: final rule. 67 *Federal Register* 53181 (2002) (codified at 45 CFR §160 and §164).
3. Protection of human subjects: common rule. 56 *Federal Register* 28003 (1991) (codified at 45 CFR §46).
4. Taira RK, Bui AAT, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. *Proc AMIA Symp*. 2002:757-761.
5. Berman JJ. Concept-match medical data scrubbing: how pathology text can be used in research. *Arch Pathol Lab Med*. 2003;127:680-686.

DOI: 10.1309/E6K33GBPE5C27FYU

6. Yount RJ, Vries JK, Councill CD. The Medical Archival Retrieval system: an information retrieval system based on distributed parallel processing. *Inform Process Manage*. 1991;27:379-389.
7. Berman JJ. A tool for sharing annotated research data: the "Category 0" UMLS (Unified Medical Language System) vocabularies. *BMC Med Inform Decis Mak*. 2003;3:6.
8. Thomas SM, Mamlin B, Schadow G, et al. A successful technique for removing names in pathology reports using an augmented search and replace method. *Proc AMIA Symp*. 2002:777-781.
9. Sweeney L. Replacing personally-identifying information in medical records: the Scrub System. *Proc AMIA Annu Fall Symp*. 1996;333-337.
10. Kurtzman NA, Nichols J. President Kennedy and Addison's disease. *JAMA*.1967;201:1052.
11. Ferris TA, Garrison GM, Lowe HJ. A proposed key escrow system for secure patient information disclosure in biomedical research databases. *Proc AMIA Symp*. 2002:245-249.
12. Sweeney L. Guaranteeing anonymity when sharing medical data: the Datafly System. *Proc AMIA Annu Fall Symp*. 1997:51-55.
13. Malin B, Sweeney L. Re-identification of DNA through an automated linkage process. *Proc AMIA Symp*. 2001:423-427.
14. Behlen F, Johnson SB. Multicenter patient records research: security policies and tools. *J Am Med Inform Assoc*. 1999;6:435-443.

DOI: 10.1309/E6K33GBPE5C27FYU