Technical Report

TR-2007-012

De-identification of medical text

by

L.. Xiong, K. Boronda, M. Graiser, C. Flowers

Mathematics and Computer Science

EMORY UNIVERSITY

De-identification of Medical Text

Li Xiong, Konstantine Boronda, Michael Graiser, Christopher Flowers

Emory University

Introduction

Health informatics is receiving a tremendous amount of attention nationally and locally. While there is an increasing need to release and share health records for purposes such as demographic and public health research and ultimately improve patient care, such data release and sharing must be governed by data privacy requirements. One of the biggest challenges facing health informatics is allowing sharing and dissemination of medical records while maintaining a commitment to patient privacy.

Currently, investigators wishing to use medical records of research purposes have three options: obtain permission from the patients, obtain a waiver of informed consent from their Institutional Review Boards (IRB) or use a data set that has had all (de-identified data set) or most (limited data set) of the identifiers removed. The de-identification process removes explicit personal health information in order to dissociate the individual from his medical record, while still preserving all the medically relevant information about the patient. It provides a necessary and scalable way for sharing medical information in a large scale while preserving privacy of patients. For example, pathologists in the United States examine millions of tissue samples each year and a large proportion of recent samples have reports available in electronic format. Researchers need pathology-based data sets, annotated with clinical information, to discover and validate new diagnostic tests and therapies. The National Cancer Institute proposed the development of the Shared Pathology Informatics Network (SPIN) recognizing the situation. Since the ultimate goal of the network is to provide researchers throughout the country access to tissue specimens, it is absolutely necessary to de-identify the contents of the surgical pathology reports that form the core of the information network.

An overarching complexity for de-identifying medical data is the data heterogeneity. A considerable amount of medical data resides in unstructured text forms such as clinical notes, SOAP (subjective, objective, assessment, patient care plan) notes, radiology and pathology reports, discharge summaries, operative and post-operative reports, and letters to referring physicians. While identifying attributes can be clearly defined in structured data, identifying information is often hidden or have multiple and different references in the text.

This chapter focuses on the de-identification problem for medical text and aims to survey the state-of-the-art approaches for addressing the problem and software tools that are available to use or adopt for health practitioners. The next section will provide broad definitions and background information of the topic. We then present the problem of de-identification for medical text in detail, discuss desired features, relevant techniques, issues and challenges, and evaluation metrics. Following the problem, we survey and

compare a representative set of approaches and systems and make a few recommendations when selecting algorithms and frameworks for de-identification of medical text. Finally we conclude the chapter and point out open problems and suggest a few directions for future research in the area.

Background

Protected Health Information (PHI) is defined by HIPAA as individually identifiable health information. Identifiable information refers to data explicitly linked to a particular individual as well with data that could enable individual identification. Identifiers include obvious ones like name and Social Security number as well as the following indirect ones.

- All geographic subdivisions smaller than a state, including street address, city, county, precinct, Zip Code, and their equivalent geocodes.
- All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
- Voice and fax telephone numbers.
- Electronic mail addresses.
- Medical record numbers, health plan beneficiary numbers, or other health plan account numbers.
- Certificate/license numbers.
- Vehicle identifiers and serial numbers, including license plate numbers.
- Device identifiers and serial numbers.
- Internet Protocol (IP) address numbers and Universal Resource Locators (URLs).
- Biometric identifiers, including finger and voice prints.
- Full face photographic images and any comparable images.
- Any other unique identifying number, characteristic, or code.

Information is considered fully de-identified if all of the identifiers (direct and indirect) have been removed, and there is no reasonable basis to believe that the remaining information could be used to identify a person. A full de-identification is not only practically infeasible but also would render the data useless for many research and analysis purposes. As an alternative, HIPAA makes provisions for a limited data set from which direct identifiers (like name and address) are removed, but not indirect ones (such as age). However, it is prone to data linkage attacks (Sweeney, 2002) that combine subject terms with other publicly available information to re-identify represented individuals. For example, an individual who is the only Caucasian male born in 1925 living in a sparsely populated area could have his age, race, gender, and zip code joined with a voter registry from the area to obtain his name and mailing address.

The Datafly system (Sweeney, 1997) first argued that a de-identified dataset may not be anonymous because of possible re-linking attacks. It proposed approaches for generalizing, substituting, and removing information as appropriate without losing many of the details for field-structured data such as pediatric patient records. Later this idea is formalized into k-anonymity (Sweeney, 2001) that is widely recognized as a privacy model for structured data by requiring a set of k entities to be indistinguishable from each other based on a predefined quasi-identifier set. Many approaches have since been proposed for privacy preserving data publishing focused on structured data.

De-Identification of Medical Text

The traditional approach for de-identification of medical text is to perform a straightforward global search and replace strategy. In this approach, the patient's identifying information (e.g., name) is given and the algorithm simply searches for all possible combinations of the patient's first, middle, and/or last name within the text report. Such an approach does not handle various false negatives such as nicknames, misspellings, and/or shortened or elaborated forms of a patient's name. Sweeney (1996) reported that on a test database of pediatric letters and physician notes, such an approach located no more than 30-60% of personally identifying information. As a result, we need a computer approach for de-identification of medical text that offers better quality and performance. We discuss a set of desired features as well as issues and challenge in designing a de-identification system in this section.

Desired Features

There are a number of important features that should be considered in designing or selecting a text de-identification system for both health practitioners and computer scientists.

Removal of HIPAA identifiers. Information is considered fully de-identified if all of the HIPAA identifiers (direct and indirect) have been removed or partially de-identified if direct identifiers (like name and address) are removed, but not indirect ones (such as age). Optimally, a de-identification system should allow configuration, so that data may be fully de-identified or partially de-identified. Full de-identification offers maximum privacy, however, this is also a theoretical position as no one will be able to guarantee with absolute certainty that the remaining information could not be used to identify a person. In addition, the resulting data loses a lot of information such as demographic information that may be useful for research. On the other hand, the partial de-identification provides better data integrity and utility. However, it is prone to data linkage attacks (Sweeney, 2002) that combine subject terms with other publicly available information to re-identify represented individuals. For example, an individual who is the only Caucasian male born in 1925 living in a sparsely populated area could have his age, race, gender, and zip code joined with a voter registry from the area to obtain his name and mailing address.

Re-linking. Once a data is de-identified, there should be a way to link the data back to the original data in case there is such a need.

Preservation of non-identifier information. While the de-identification system should remove all HIPAA identifiers (direct or indirect), it needs to preserve any information in the text document that is not an identifier.

Established efficacy. The de-identification system should have accompanying data and evaluation measures that establish the efficacy of the system.

Generality across document types. Many different kinds of medical text reside in healthcare organizations. They include clinical notes, pathology reports, discharge summaries, operative and post-operative reports, and letters to referring physicians. Many of them such as medical discharge summaries are characterized by incomplete, fragmented sentences, and ad hoc language. They use a lot of jargon, many times omit subjects of sentences, use entity names that can be misspelled or foreign words, can include entity names that are ambiguous between identifiers and non-identifiers, etc. Some of them such as pathology reports contain a semi-structured portion with patient information while others such as referring letters consist of completely free text. While specialized de-identification systems can be designed and tuned specifically for a special type of document, a general identification system needs to be flexible and work with all kinds of documents.

Generality across multiple institutions. It is highly likely that individual differences among institutional reporting systems, styles and formats will require adoption effort but this should be minimal.

Maintainability. Related to the above criteria, as new institutions deploy the system, the system should be designed to minimize the burden of making small adjustments.

Integration into current clinical system. The system should be easily integrated into the current clinical systems so that data can be de-identified and moved to data repositories for research purposes as it becomes available from the clinical systems.

Techniques

The problem of de-identification of medical text involves two sub-tasks: 1) The identification of personally identifying references within medical text, and 2) the masking, coding, or replacing of these references with values irreversible to unauthorized personnel. Once the de-identification is done, the last step is to effectively evaluate the de-identification for quality control and further improvement.

Extracting identifiers from medical text can be seen as an application of named entity recognition (NER) (also known as entity identification (EI) and entity extraction) problem. It is a subtask of information extraction that seeks to locate and classify atomic

elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

NER systems can be roughly classified into two categories, each having their advantages and disadvantages. The first uses *grammar-based or rule-based* techniques. It relies heavily on hand-coded rules and dictionaries. Depending on the type of identifying information, there are common approaches that can be used.

- For identifiers that is a closed class with an exhaustive list of values such as geographical locations, and names, common knowledge such as lists for area codes, common names, common first names, nicknames, words that sound like first names (Soundex) can be used for lookups. Local knowledge such as first names of all patients in a specific hospital can be also used for specific dataset.
- For identifying information that follows certain syntactic pattern (e.g., phone numbers, zip codes), regular expressions can be used to match the patterns. Common recording practices (templates) with respect to personal information can be utilized to build rules.
- For many cases, a mixture of semantic information including context such as prefix for a person name, syntactic features, dictionaries, and heuristics need to be considered.

Hand-crafted grammar-based systems typically obtain better results, but at the cost of months of work by experienced domain experts. In addition, the rules that are used for extracting identifying information will likely need to change for different types of records (radiology, surgical pathology, operative notes) and across organizations (hospital A formats, hospital B formats). The software will become increasingly complex and slow with growing rules and dictionaries.

The second category uses *statistical learning* approaches. Machine learning methods have been applied to the NER problem with remarkable success. They typically require a large amount of manually annotated training data that has pre-labeled identifiers. It can then use a list of feature attributes to train a classification model and classify the terms in new text as either identifier or non-identifier. Typical feature sets include terms themselves, their local contexts (such as words before and after), and dictionary related features. The most frequently applied techniques were the Maximum Entropy model, hidden markov model, and support vector machines. It can be ported to other languages, domains or genres of text much more rapidly and require less work overall.

Once the identifying information is detected, they need to be removed, replaced, or recoded to maintain the anonymity of the patients. De-identified data will be subjective to linking attack where adversary can infer some information concerning some patients by linking the document to other sources.

Issues and Challenges

Applying these techniques for extracting identifiers from medical text presents a number of challenges.

Misspelling. Misspelling could be challenging to detect, especially if the misspelling results in a valid word or even worse a valid word with a specific medical meaning.

Foreign address and foreign names. These are challenging since many of the words will not be found in English language lists that are typically used in dictionary-based information extraction techniques. In addition, there are a variety of address formats used throughout the world. This posts an issue for rule-based systems that rely on known templates for addresses.

Medical eponyms. The system should not remove medical eponyms such as Barrett's esphagus or medical terms that include locations such as Philadelphia translocation.

Indirect identifying information. Information is considered fully de-identified if all of the identifiers (direct and indirect) have been removed, and there is no reasonable basis to believe that the remaining information could be used to identify a person. Indirect identifying information such as unique medical or social history, sequence of events, or combinations thereof are inherently difficult to find and remove since they do not contain specific identifiers, but can effectively limit the number of possible individuals greatly. 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. It is possible that people could identify this person despite the lack of identifiers such as name, address, or phone number.

Preservation of identifier relationships. When multiple identifiers are present within a particular document, they create implicit relationships that are needed to understand the meaning of the document. For example, dates can be used in text in such a way that when they are removed, it is impossible to determine the sequence of events (which thing happened first, which thing happened next). Names suffer from the same problem (who did what). Optimally, a de-identifier system should replace all instances of the same identifier within a single document with an identical and unique de-identification token. In the case of dates, these should include information about the temporal offset from the original date. In this way, the semantic meaning inherent in these relationships is not lost.

Evaluation Metrics

Quality of de-identification. The quality of de-identification is directly related to how well the extraction component detects the identifying information. Evaluation of the quality of de-identification is typically carried out by having domain experts examine the de-identified dataset along with original dataset. Their task is to find the text that should have been de-identified but was not (under de-identified or under-marked) and clinical text that was inadvertently removed as if it were identifying text but should not (over-marking).

There is a large variation in the measures that are used to evaluate de-identification systems. We introduce a few common measures used in the medical informatics and machine learning community and define them in our de-identification context. For simplicity, we assume the medical text consists of a set of identifiers (positive) and non-identifiers (negative). The identifiers or non-identifiers could be a word or a phrase. A de-identification system will mark each term either as an identifier (positive) or a non-identifier (negative). Table 1 shows a confusion matrix which depicts how a de-identification system may mark the text. Each term increments one cell in the confusion matrix. The false negative corresponds to the under-marking error where the system misses an identifier that should have been de-identified. The false positive corresponds to the over-marking error where the system removes a non-identifier as an identifier.

Give the confusion matrix, a few metrics are commonly used and they are summarized in Table 2. In general, all three measures (Precision, Recall, and Specificity) or the entire confusion matrix should be presented if one does not know the appropriate measures to use for the distribution of their data set. For our task of de-identification, our main concern is to detect all identifiers (to preserve privacy) and not to over-mark non-identifiers (to preserve data integrity and utility), so we recommend a minimum of precision and recall/sensitivity. Arguably, recall is more important than precision. Low recall indicates that many identifiers remain in the documents and that there is high risk to patient privacy. Low precision means that words that do not correspond to identifiers have also been removed. This hurts the integrity of the data but does not present a risk to privacy.

Table 1. Confusion Matrix

		Detected Identifiers	
		Positive	**Negative**
Known Identifiers	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 2. Evaluation Metrics

Measure	**Formula**	**Intuitive Meaning**
Precision	TP / (TP + FP)	The percentage of marked identifiers that are real identifiers
Recall / Sensitivity	TP / (TP + FN)	The percentage of identifiers that are correctly marked
Specificity	TN / (TN + FP)	The percentage of marked non-identifiers that are real non-identifiers
Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of identifiers and non-identifiers that are correctly marked

Data Utility. Another aspect that has not been a traditional focus of de-identification of text is data utility. We argue that more attention should be given to maximizing data utility when masking or recoding the detected identifying information.

Efficiency and Scalability. Finally, a system should be efficient and scalable in order to work with large volume of data.

Solutions and Recommendations

In this section we will survey the state of the art approaches for addressing the text de-identification problem as well as the software tools that can be used or adopted. The list is not intended to be comprehensive. The objective is to highlight the representative approaches for practitioners and summarize the techniques and limitations of current work for researchers. A summary table of the surveyed systems is provided in Table 3 for a quick comparison. We discuss representative systems and classify and compare them along the following dimensions as we discussed in previous section.

- Document types that the approach is designed for, whether they are general text or specific document types such as pathology reports.
- The information that is removed or anonymized, whether they are a specific subset of HIPPA identifiers, and/or indirect identifying information,
- Algorithms and techniques that are used for extracting identifying information.
- Algorithms and techniques that are used for removing, masking or replacing the identifying information.
- Datasets that the approach is tested against or adopted for and evaluation results in terms of quality of de-identification and performance when reported.

Scrub System

The scrub system (Sweeny, 1996) is one of the earliest scrubbing systems that locates and replaces HIPAA-compliant personally-identifying information for general medical records.

Methods. The system is an example of the rule-based and dictionary-based systems. It utilizes a set of detection algorithms competing in parallel to label text as being a name, an address, a phone number, and so forth. Each detection algorithm recognizes a specific entity such as first name, street address, and date. The entities may overlap. For example, there is a detection algorithm for first name, last name, and full name. The precedence is based on the number of entities that constitute the algorithm's assigned entity. E.g. Location detection algorithm constitutes city, state, and country and detecting location may make it possible to identify a city, state and country. Therefore location has a higher precedence than the latter three detection algorithms. Each algorithm tries to identify occurrences of its assigned entity, and reports a certainty score. The algorithm with the highest precedence and greatest certainty above a minimal threshold prevails.

Table 3. State-of-Art De-identification Systems for Medical Text

Approaches/ Systems	Status	Document Types	De-identified Information	Methods	Dataset	Results
Scrub system (Sweeny, 1996)		General	HIPAA identifiers	Rule and dictionary based	3198 letters to referring physicians	Precision: unknown Recall: 99% - 100%
Semantic Lexicon system (Ruch, 2000)		General	Explicit identifiers	Rule based	Training: 20% Test: 80% 40 rules	Precision: 100% Recall: 96.8% Performance: 3 weeks to write all the rules
Name De-identifier - Semantic Selectional Restrictions (Taira, 2002)		General	Name only	Statistical - Maximum entropy classifier	Training: 1350 reports (907 names) Test: 900 reports.	Precision: 99.2% Recall: 93.9% Performance: 0.5 minute/5Kb reports
Name De-identifier - Augmented Search and Replace (Thomas, 2002)		Pathology reports	Name only	Dictionary based	1001 pathology reports 231 names	Precision: unknown Recall: 98.7%
Concept-Match System (Berman, 2003)	Open source	Pathology reports	Everything except terms in a non-identifying word list	Dictionary based (white list)	JHARCOLL (567921 pathology phrases)	Precision: unknown Recall: near 100%
De-Id (Gupta, 2004)	Commercial	General	HIPAA identifiers and other identifying information	Rule and dictionary based	Phase 2: 1000 reports Phase 3: 3000 reports	Precision: unknown Recall: unknown
De-identification nursing notes (Douglas, 2005)	Open source	Nursing notes	HIPAA identifiers	Rule and dictionary based	747 nursing notes (22 patients)	Precision:43.5% Recall: 92%
HMS Scrubber (Beckwith, 2006)	Open source	Pathology reports	HIPAA identifiers	Rule and dictionary based	Test: 1800 pathology reports	Precision 42.8% Recall: 98.3% Performance: 47 cases per minute
SVM-based System (Sibanda, 2006)		Discharge summaries	HIPAA identifiers	Statistical – SVM classifier	Re-identified dataset: 2784 identifiers Authentic: 4194	Precision: 92-97.7% Recall: 92.1-98.2%

					identifiers	
Iterative Method (Szavas, 2006)		Discharge summaries	HIPAA identifiers	Iterative and Statistical – decision tree classifier	I2b2 dataset: 671 reports (14314 identifiers)	Precision: 98.1% Recall: 96.7%

If the detected entity was a date, the replacement date may involve lumping days to the first of the nearest month or some other grouping. If the detected entity was a first name, the typical strategy is to perform a hash-table lookup using the original name as the key. The result of the look-up is the replacement text. This provides consistent replacement. In terms of the replacement content, several strategies are available including the user of orthographic rules that replace personally identifying information with fictitious names. The system also suggested a few strategies to combat the problem of reverse scrubbing where a person with inside knowledge is able to identify the real person from scrubbed materials. One of them is to group fields together such as lumping dates by week.

Dataset and evaluations. A subset of pediatric medical record system consisting of 275 patient records and 3198 letters to referring physicians are used for evaluation. The reported recall achieves 99% - 100%. However, the precision is not reported.

Semantic Lexicon based System

The medical document anonymization system with a semantic lexicon (Ruch, 2000) is another scrubbing system that locates and removes personally-identifying information in patient records. It removes explicit personally-identifying information such as name, address, phone number, and date of birth.

Extraction techniques. Rules based. It differentiates strict identity marker (IDM), tokens always directly followed or preceded by identifiers, e.g. Ms., Dr… and tokens likely to refer to general persons (doctors, professors …) and not necessarily followed or preceded by identifiers. Phone numbers and dates use explicit markers and well defined patterns (exhaustively listed).

Replacement. The system replaces each character of any confidential items by an 'x', respects the case and punctuation. Tractability is not allowed in the system, i.e. P.Nertens and W.Keuster are both replaced by X.Xxxxxx. The advantage of such a replacement strategy is that the reverse scrubbing is forbidden or more difficult. However, tractability is lost which may be necessary for studies on genealogy

Dataset and Evaluations. The evaluation involved tuning the system on 20% of the corpus (training set) and writing more than 40 rules to reach 100% success rate for the training set. It was then tested on 80% of the corpus (test set). It reported that 96.8% (452 out of 467) identifiers are correctly removed and 0 tokens removed which are not identifiers. This converts to a 100% precision and 96.8% recall. However, it is a very small dataset. In addition, it is at the cost of three weeks to write all the rules. It is not clear how such a system would perform with large volume of data.

Name De-identifier - Augmented Search and Replace

Thomas (2002) described a method for removing names in pathology reports by using an augmented search and replace method.

Extraction Methods. The extraction technique is mainly dictionary based. The system takes advantage of the fact that the vast majority of proper names in pathology reports occur in pairs. In rare cases where one proper name is by itself, it is preceded or followed by an affix that identifies it as a proper name (Mrs., Dr., PhD). The tool was based on this observation and was largely based on publicly available data sources. They compiled a Clinical and Common Usage Word (CCUW) list as well as a fairly comprehensive proper name list and used an augmented search and replace method.

Dataset and Evaluations. They reported 98.7% of 231 proper names in the narrative sections of pathology reports are correctly identified. Three single proper names were missed out of 1001 pathology reports (0.3%, no first name/last name pairs). Precision is not reported.

Name De-identifier - Semantic Selectional Restrictions

Taira (2002) described a method for removing names in general medical text based on statistical models. Unfortunately, the system was limited to the removal of patient names. Other identifiers, such as those described by HIPAA, remained in the medical documents.

Document types. The system is tested against reports from pediatric patients generated at the UCLA Clark Urology Center. The categories of the reports included: 1) letters and reports to referring physicians, 2) discharge summaries, 3) clinical notes, and 4) operative/surgical reports.

Extraction methods. The proposed algorithm is based on estimating the fitness of candidate patient name references to a set of semantic selectional restrictions. The semantic restrictions place tight contextual requirements upon candidate words in the report text and are determined automatically from a manually tagged corpus of training reports. Maximum entropy classifiers are used to provide a probabilistic measure of the belief of a given candidate token to a given semantic restriction.

Dataset and Evaluation. The training files consists of 1350 random reports from pediatric patients generated at the UCLA Clark Urology Center. A researcher reviewed each report within the training corpus, manually tagging all references to patient names and the local contexts in which the names were used. Within the training set, a total of 907 patient name instances were tagged, all located within the unstructured (i.e., non-header) portions of the report. The test set includes 900 random reports. The best overall performance reported has a precision score 99.2% and recall score 93.9%. The time to process a 5Kb report was about one-half minute on a 2GHz personal computer.

Concept-Match Scrubber

The concept-match scrubber (Berman, 2003) provides an alternative way to traditional de-identification systems by extracting and removing every word from the text except words from an approved list of non-identifying words. It is designed for pathology reports.

Extraction methods. The concept-match based medical data scrubbing system uses a white-list based approach, in other words, it extracts every word as identifying information except words from an approved list of non-identifying words such as the Unified Medical Language System (UMLS). When a medical term matching a standard nomenclature term is encountered, the term is replaced by a nomenclature code and a synonym for the original term. When a high-frequency "stop" word, such as a, an, the, or for, is encountered, it is left in place. When any other word is encountered, it is blocked and replaced by asterisks.

Dataset and Evaluation. A surgical pathology text corpus was obtained from JHARCOLL, a public domain collection of more than half a million different phrases extracted from actual pathology reports (surgical pathology, cytopathology, and autopsy). Although not explicitly reported, the concept match method almost guarantees a 100% *recall*. No identifiers will be present in the output because the only words in the output come from an approved list of terms. But it suffers a low *precision* because of a large number of over-marked non-identifiers. In addition, the output of the algorithm will be hard to read. On the *performance* note, all 567921 pathology phrases in JHARCOLL were scrubbed in 2968 seconds (nearly 200 phrases per second). The speed is fast enough to make the program usable even for a very large corpus of text.

De-Id

DE-ID was developed at the University of Pittsburgh, where it is used as the de-identification standard for all clinical research approved by the Institutional Review Board (IRB). De-Id's main function is to locate identifiable text, as defined by the safe-harbor method or the limited data set in the document in question. De-ID Data Corp acquired the global rights to De-ID Software in 2004 and, after a period of product development and refinements, launched De-ID for commercial and academic license in March 2005.

The National Cancer Institute (NCI) has licensed DE-ID to be used as a de-identification component of some of the software applications developed in the Tissue Banks and Pathology Workspace of the NCI-sponsored cancer Biomedical Informatics Grid (caBIG). The NCI will use DE-ID internally and sublicense the software to sixteen (16) NCI-designated Cancer Centers participating in caBIG across the United States as part of a program to advance the development of the caBIG network while considering longer term, strategic options.

Document Types. De-Id is designed to work with archives of all types of clinical documents. 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, the reported evaluation (Gupta, 2004) only includes pathology reports. It supports multiple formats of input documents including XML formatted documents and tab-delimited documents (spreadsheets or relational databases).

De-identified Information. It includes the 17 (full-face photographic images not included) HIPAA-specified identifiers as well as potential identifiers not included in safe harbor, such as names of health care providers (physicians, laboratories, and hospitals), employers, and relatives.

Extraction Methods. De-Id implements a set of rules and dictionaries designed to identify the identifying information. Examples of these rules include the examination of document headers for patient and provider names, use of the Unified Medical Language System (UMLS) Meta-thesaurus 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.

Replacement Methods. De-Id replaces identifiable text with de-identified but specific tags. Identifiers found multiple times in the report are replaced consistently with the same tag to improve continuity and readability of the report. Dates are replaced by tags, but date tags from 2 different dates retain the time interval between them.

Dataset and Evaluation. Gupta (2004) described the initial quality assurance evaluation of the De-Id engine in the domain of surgical pathology reports and describes a useful model for other testing and quality control of de-identification engines at other institutions and for identifying the unique challenges presented by pathology reports. Textual surgical pathology reports were selected randomly from UPMC and processed by the De-Id engine. The de-identified reports were distributed for evaluation to 4 pathologists. The de-identification engine was evaluated 3 times. Problems identified at each pass were discussed between evaluations to improve de-identification of pathology reports. A new set of 1,000 reports was reviewed in a similar manner in the second evaluation and an additional 300 reports were evaluated in the third evaluation. Only # of over-marking errors and itemized # of under-marking errors and total number of reports are reported. No precision or recall is reported or can be derived without the knowledge of total # of marked identifiers.

Re-linking. De-ID automatically creates a linkage file when a dataset is processed. The linkage file is stored in an encrypted format and only accessible by password. This is to assure the study ID remains consistent across data sets, but different admissions and/or multiple reports can be easily identified.

HMS Scrubber

HMS Scrubber (Beckwith, 2006) is an open source, HIPAA compliant, de-identification tool tailored for pathology reports.

Extraction Methods. It performs a three-step process for removing potential identifiers. The pathology reports are first converted to an XML format. This format includes a header portion and a textual portion. The header portion contains demographic information about the patient such as name, medical record number, date of birth and social security number; and information about the pathology report, such as the accession number and the pathology department. The first step of the extraction is to take advantage of identifying information that may be present in the header of the file such as such as name, medical record number, pathology accession number, etc and remove them from the textual portion of the reports. The second step is to perform a series of pattern matches to look for predictable patterns likely to represent identifying data; such as dates, access numbers and addresses as well as patient, institution and physician names that can be found by markers such as Dr, MD etc. The pattern searches are implemented as regular expressions. Finally, a database of proper names derived from publicly available census lists and geographic locations derived from a gazetteer file which contains US place names are used. At one institution, the names were augmented with the names of pathologists who were active during the period from which the reports were drawn.

Replacement. The scrubber replaces any suspect words with a series of X's and a notation as to what type of information was presumably removed.

Dataset and Evaluation. Pathology reports from three institutions were used to design and test the algorithms. The software was trained on training sets until it exhibited good performance (the exact number of reports used as training set is not reported). 1800 new pathology reports were used as a test set. It reported a *recall* of 98.3% (3439 of 3499 unique identifiers in the test set were removed). Although achieving results similar to previous systems designed for pathology reports, the result also highlighted the wide variance in number of identifiers between external consult and in-house cases as well as between different institutions. It also reported a *precision* of 42.8% (4671 over-scrubs) that is primarily related to the large number of words contained in names and places table. The reported *performance* is 47 cases per minute which makes it suitable for high volume applications compare to other approaches.

SVM-based System

Sibanda (2006) proposed a SVM-based system for de-identifying medical discharge summaries using statistical SVM-based classification method.

Document types. The system is designed and test again discharge summaries. The authors argued that the medical discharge summaries are characterized by incomplete, fragmented sentences, and ad hoc language. Entity names can be misspelled or foreign words, can include entity names that are ambiguous between PHI and non-PHI, etc. In addition, discharge summaries do not present in formation in the form of relations

between entities, and many sentences contain only one entity. Given such situations, the local context is more important than global context in learning the identifiers.

Extraction Methods. Sibanda and Uzuner (2006) built an SVM-based system, that given a target word (TW), would accurately predict whether the TW was part of PHI. The feature set focuses on immediate context of the target word, paying particular attention to cues human annotators found useful for de-identification. The feature set included a total of 26 features, 12 of which were dictionary-related. Information gain showed that the most informative features were the TW, the bigram before and after the TW, the word before and after the TW.

Dataset and Evaluations. The evaluation used two corpora. One was previously de-identified where many PHI and some non-PHI had been replaced with the generic placeholder and then re-identified by replacing the placeholders with appropriate, fake PHI or non-PHI terms. The second corpora is a set of authentic discharge summaries. It reported a precision of 97.5% on authentic discharge summaries and 92-97.7% on re-identified dataset including the corpus containing ambiguous data. The recall is 95% on authentic discharge summaries and 92.1-98.2% for the re-identified dataset.

Recommendations

Although not directly comparable, we could expect statistical approaches outperform the traditional rule and dictionary based approaches. However, it relies on well annotated training dataset. When evaluating the quality of de-identification, systems that offer a high precision but slightly lower recall are best suited for limited data use agreements because of the maximum data integrity. On the other hand, the methods that provide a high recall but a low precision may be the best option for preparing data for public distribution.

Even though the De-Id program correctly de-identifies information most of the time, it makes under-marking and over-marking errors. In the foreseeable future, it is unlikely that any computer-based de-identification 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.

All the tools should be used together with classical and legal procedural barriers
It must be kept in mind that de-identification processes do have limitations and it is safest if de-identified reports are restricted to use by investigators who have signed data use agreements that include prohibitions on attempting to re-identify the subjects of such reports. In such a setting, the risk to patients and research subject is low and is consistent with the statutory protections described in HIPAA and the Common Rule.

Conclusion and Future Trends

While there have been extensive developments of data de-identification and data privacy technologies in general, there are still gaps to be filled. This section will discuss a set of open problems that still remain with the text de-identification problem and data de-identification problem and suggest directions for future research.

Quality of de-identification. All the results show that it is extremely difficult, if not impossible, to automatically remove all possible identifiers from medical text while leaving the remainder intact. Misspelled identifiers and foreign names and addresses remain challenging for rule and dictionary based system. Use of an automated spell-checker prior to scrubbing could be tested to see if it improves identifier removal. Ambiguous terms remain to be challenging. For example, De-Id (Gupta, 2004) reported that names that are medical terms but are not listed in the Unified Medical Language System (UMLS), such as Hickman catheter, are one of the main reasons attributed to their over-marking errors. Addresses that contain commonly used words, such as MI (state of Michigan) can be also confused as an abbreviation for myocardial infarction.

We believe statistical methods are better positioned to address such challenges by incorporating contextual information learned from a much more massive corpus of data. Disambiguation techniques in natural language processing will be a promising direction to pursue. They will obviate the need of rule and dictionary based system to look up almost endless lists of personal identifiers and compile rules of exceptions, which may conflict.

Indirect identifying information. It is still inherently difficult to find and remove indirect identifying information. Currently, there is no good mechanism for any de-identification system to identify such unique combinations or sequences of events in textual information that has the potential for patient identification. Relationship extraction techniques can be explored in addition to entity extraction techniques to detect relationships and implicit identifying information.

Preservation of identifier relationships. When multiple identifiers are present within a particular document, they create implicit relationships that are needed to understand the meaning of the document. Relationship extraction techniques again can be utilized to extract relationships among entities and their attributes. More intelligent replacing strategies are also needed to preserve the semantic meaning inherent in these relationships.

Easy deployment and continued quality assurance. It is important to have systems and techniques that can be easily deployed with minimal tuning. In addition, mechanisms and interfaces for continued monitoring and improvement of a de-identification system will also prove extremely useful, especially when expanding the types or sources of data to be scrubbed.

Benchmark. Lack of publicly available and standardized dataset has been one of the biggest barriers to de-identification research. It's a chicken and egg problem that researchers need clinical data to design and evaluate de-identification algorithms but authentic clinical data is very difficult to obtain for privacy reasons without a good existing de-identification system. In addition, without a common benchmark, it is extremely difficult to evaluate and compare various approaches. Uzuner et al. (2006) took an important step towards this direction in their i2b2 workshop on natural language processing challenges for clinical records. They generated and released a set of fully-de-identified medical discharge summaries that is re-identified with surrogate and ambiguous personal identifying information to the research community. One of the grand challenges for the workshop is automatic de-identification of the released data. The availability of the clinical records will be a major resource for both medical informatics community and natural language processing community.

Data utility. While maximizing data utility has been an active research topic for data anonymization research for structured data, the implications of maximizing data utility for text data are not clearly understood. Application-driven metrics need to be designed and employed to maximize the data utility of de-identified dataset for targeted applications and users. An interesting research direction is to apply models such as k-anonymity (Sweeney, 2002) to medical text data. It presents interesting research challenges due to the difficulty of mapping the relevant identifying attributes to a single entity.

Scalability. Most of the existing systems did not report performance results in terms of efficiency and scalability. It is our belief that de-identification solutions need to be not only effective but also scalable for the ever-growing medical data.

References

Beckwith, B. A., Mahaadevan R., Balis U. J., & Kuo F. (2006). Development and evaluation of an open source software tool for de-identification of pathology reports. *BMC Medical Informatics and Decision Making*.

Berman J.J. (2003). Concept-match medical data scrubbing: How pathology text can be used in research. *Arch Pathol Lab Med* 2003, 127:680-6.

caBIG workspace University of Pittsburg DeID developer project form. https://cabig.nci.nih.gov/workspaces/CTMS/Documents/CTMS_Documents/UofPitt_DeID_Developer_Project_Form.pdf

Miller R., Boitnott J. K., & Moore G.W. (2001). Web-based free-text query system for surgical pathology reports with automatic case de-identification. *Arch Pathol Lab Med* 2001, 125:1011.

Farkas R., Szarvas G., Ivan S., Andras K., & Busa-Fekete R. (2006). An iterative method for the de-identification of structured medical text. In *AMIA I2B2NLP workshop proceedings*, 2006

Douglass, M., Clifford, G.D., Reisner, A., Moody, G.B., & Mark R.G. (2004). Computer-assisted de-identification of free text in the MIMIC II database. *Computers in Cardiology*, 2004, Sept. 2004, 341- 344

Douglass M., Clifford G. D., Reisner A., Long W. J., Moody G. B., & Mark R. G. (2005). De-Identification algorithm for free-text nursing notes, *Computers in Cardiology*, 32:331-334; IEEE Computer Society Press, September 2005.

Fielstein E. M., Brown S. H., & Speroff T. (2004). Algorithmic de-identification of VA medical exam text for HIPAA privacy compliance: preliminary findings. *MEDINFO* 2004.

Gupta D, Saul M, Gilbertson J (2004). Evaluation of a de-identification (De-Id) software engine to share pathology reports and clinical documents for research. *Am J Clin Pathol* 2004, 121:176-86.

Ruch P, Baud RH, Rassinoux AM, Bouillon P, & Robert G. (2000). Medical document anonymization with a semantic lexicon. *Proc AMIA Symp*. 2000, 729-33

Sibanda T., Uzuner O. (2006). Role of local context in de-identification of ungrammatical, fragmented text. *Proceedings of the North American Chapter of Association for Computational Linguistics/Human Language Technology (NAACL-HLT 2006)*.

Sweeney L. (1996). Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp 1996, 333-7.

Sweeney L. (1997). Guaranteeing anonymity when sharing medical data, the Datafly System. Proc AMIA Annu Fall Symp 1997, 51-5.

Sweeney L. (2002). k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002

Taira RK, Bui AA, Kangarloo H (2002). Identification of patient name references within medical documents using semantic selectional restrictions. Proc AMIA Symp 2002, 757-61.

Thomas SM, Mamlin B, Schadow G, McDonald C (2002). A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp 2002, 777-81.

Uzuner O., Szolovits P., & Kohane I. (2006). I2b2 workshop on natural language processing challenges for clinical records. *Proceedings of the Fall Symposium of the American Medical Informatics Association (AMIA 2006)*, Washington, DC, 2006.