Automated De-Identification of Free-Text Medical Records

by

Ishna Neamatullah

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

September 5, 2006

Author_____
Department of Electrical Engineering and Computer Science
September 5, 2006

Certified by_____
Roger G. Mark
Professor of Electrical Engineering
Distinguished Professor in Health Science and Technology
M.I.T. Thesis Supervisor

Accepted by_____
Arthur C. Smith
Professor of Electrical Engineering
Chairman, Department Committee on Graduate Theses

Automated De-Identification of Free-Text Medical Records
by
Ishna Neamatullah

Submitted to the
Department of Electrical Engineering and Computer Science

September 5, 2006

In Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

This paper presents a de-identification study at the Harvard-MIT Division of Health Science and Technology (HST) to automatically de-identify confidential patient information from text medical records used in intensive care units (ICUs). Patient records are a vital resource in medical research. Before such records can be made available for research studies, protected health information (PHI) must be thoroughly scrubbed according to HIPAA specifications to preserve patient confidentiality. Manual de-identification on large databases tends to be prohibitively expensive, time-consuming and prone to error, making a computerized algorithm an urgent need for large-scale de-identification purposes. We have developed an automated pattern-matching de-identification algorithm that uses medical and hospital-specific information. The current version of the algorithm has an overall sensitivity of around 0.87 and an approximate positive predictive value of 0.63. In terms of sensitivity, it performs significantly better than 1 person (0.81) but not quite as well as a consensus of 2 human de-identifiers (0.94). The algorithm will be published as open-source software, and the de-identified medical records will be incorporated into HST's Multi-Parameter Intelligent Monitoring for Intensive Care (MIMIC II) physiologic database.

Thesis Supervisor: Roger G. Mark
Title: Professor of Electrical Engineering
Distinguished Professor in Health Science and Technology

Acknowledgements

I would like to sincerely thank my M.Eng. Thesis supervisor Professor Roger G. Mark for an incredible amount of help during the de-identification algorithm's implementation and testing and for his feedback on the write-up. The current performance of the algorithm also owes greatly to the incremental reviews, feedback and help provided by Gari Clifford, Mauricio Villaroel and George Moody. The de-identification algorithm is a development on a prototype implemented by Margaret Douglass to whom I am grateful for the prototype and her very helpful advice. Finally I would like to thank the National Institute of Biomedical Imaging and Bioengineering for its grant (Grant Number R01 EB001659) that made this de-identification project possible.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem Statement

The goal of our de-identification project was to create an algorithm to autonomously de-identify protected health information (PHI) in medical records from intensive care units (ICUs). Manual de-identification used on large databases tends to be prohibitively expensive, time-consuming and prone to error, making a computerized algorithm an urgent need for de-identification purposes. The project was supervised by Prof. Roger G. Mark at the Harvard-MIT Division of Health Science and Technology (HST), and the de-identified medical records will be incorporated into the Multi-Parameter Intelligent Monitoring for Intensive Care (MIMIC II) physiologic database developed at HST.

1.2 Background

MIMIC II is an annotated database of cardiovascular and related signals and accompanying clinical data from ICUs. The database contains physiologic signals, medical notes, laboratory test reports and related information on ICU patients from hospitals in the Boston area. The project is sponsored by the National Institute of Biomedical Imaging and Bioengineering and two divisions of Philips (Philips Medical Systems, Andover, MA and Philips Research, Briarcliff Manor, NY) with the goal of making the database available to the research community on the PhysioNet physiologic resource website [1].

The main use of the MIMIC II database lies in supporting research in developing intelligent monitoring systems. Its secondary use is in fundamental research in cardiovascular physiology and clinical studies of ICU patients. Cardiovascular and related signals obtained from continual monitoring of patients in MIMIC II will be a vital

resource in waveform analysis and time series modeling of the cardiovascular system. The patient records are currently being reviewed by clinicians to annotate clinically important events, including signal abnormalities (e.g., arrhythmia), disease symptoms (e.g., fever, nausea), disease (e.g., hypertension, hemorrhage) and medical prescription changes. In the future, this annotated database will provide a research resource for the development of intelligent ICU monitoring systems [2].

1.3 Motivation

The release of MIMIC II data [1, 2] for research purposes faces legal hurdles since it poses significant risk of breaching patient confidentiality. The Privacy Rule of the Code of Federal Regulations (45 CFR Parts 160 and 164) stipulates that research groups obtain and use completely de-identified data sets (stripped of all 18 identifiers defined under HIPAA) [3]. The free-text medical files included in the MIMIC II database contain a wide range of confidential patient information that must be de-identified before the database's release to research groups in accordance with federal regulations. Regardless of the actual possibility of identifying an individual patient from released information, there is significant perceived risk of information disclosure. De-identification, supervised by institutional review boards (IRBs) under the Federal Policy for the Protection of Human Subjects (56 Federal Register 28003) [4], thus not only preserves patient confidentiality but also increases public trust in medical research. Confidential information in the text has to be scrubbed in accordance with the United States government's Health Information Portability and Accountability Act (HIPAA) [5], which specifies 18 types of identifiers that must be removed to preserve patient confidentiality. These identifiers, referred to as protected health information (PHI), include names,

geographic locations more precise than a state, elements of dates except year, social security numbers, telephone and fax numbers.

The process of de-identification involves reviewing the entire corpus of medical records to identify all occurrences of PHI and removing them. In an additional optional step the PHI could be replaced by fake, representative information in a process referred to as re-identification. The database could be manually de-identified by clinicians or persons familiar with medical terms. However, manual de-identification trials were found to be very time-consuming (each clinician was able to read only 80,000 words in 4-5 hrs) and expensive in terms of payment to the de-identifiers ($50/hr) [23, 24]. In addition, de-identification results varied greatly from person to person and were prone to error. Large-scale de-identification thus necessitates an automated computerized system that is fine-tuned to the textual structure and content of the medical records and to the research group's specific needs.

We have developed a pattern-matching de-identification algorithm that is usable for any free text but is finely-tuned to our research requirements. Our de-identification efforts centered on developing a specific algorithmic tool to scrub free-text nursing notes, discharge summaries and other text-based reports (e.g. as ECG, radiology reports). We plan to publish the open-source de-identification software on the PhysioNet website, and release the fully de-identified corpus of MIMIC II nursing notes and discharge summaries for use by qualified researchers.

1.4 Thesis Roadmap

In this section we laid out an introduction of the purpose and scope of our de-identification work on MIMIC II text files. Section 2 of this paper provides an overview of current de-identification techniques and previous de-identification research on MIMIC

II text. Section 3 describes our de-identification algorithm in detail, while Section 4 reports on its evaluation. Finally, Section 5 concludes on the MIMIC II de-identification project and discusses avenues for further development.

2 Previous De-identification Work

2.1 De-identification Paradigms and Previous Work

The process of free-text de-identification transforms a piece of text to the same piece of text with some words of the original missing. Current de-identification techniques used in the medical research field can be broadly divided into two paradigms: extraction and concept-match. Under extraction, the original text is implicitly taken as the default result of de-identification, and identifying terms are searched out and removed from that default. Concept-match, on the other hand, implicitly starts off with blank text and fills it in with non-identifying text from the original, thus leaving out all identifying information.

Extraction is widely used for de-identifying medical records, with different research groups developing their own algorithms fine-tuned to their needs. An example of a new and well-documented algorithm is the one developed by Beckwith et al [6] which implements a three-step process to remove potential PHI. Information known *a priori* about the patient, e.g. patient name, medical record number, is removed first. Second, the text is searched for general patterns of dates, addresses names, institution names, etc. The matching text is removed. Finally, a database of known PHI, like proper names and locations, is used to identify and remove these PHI in the text. This type of implementation is fine-tuned to the requirements of the group's research needs, i.e. takes into account the PHI types the group needs de-identified and the *a priori* information available in the group's medical records and databases.

The Health Sciences Library System at University of Pittsburgh Medical Center has developed another de-identification algorithm that uses the extraction paradigm [7]. This approach is very similar to ours in that it uses a set of heuristics and dictionaries of

known PHI to identify occurrences of 17 HIPAA-specified PHI categories. In addition, the algorithm uses the UMLS Metathesaurus to preserve medical terms.

Some algorithms focus on a single PHI category, like the system developed by Thomas SM et al at Regenstrief Institute for Healthcare in Indianapolis [8]. This algorithm uses the knowledge that proper names tend to occur in pairs and are commonly preceded by an affix (Dr., Mrs., etc). Lists of common words and known names are used to augment the search. The method discovered 98.7% of 231 proper names in textual pathology reports. We use a very similar augmented search strategy in our algorithm. A system developed by Miller RE et al [9] identifies proper names of people and institutions in the free-text database of pathology reports using two main techniques: search of known proper names, augmented by context analysis of the proximity to proper name affixes ("Dr.", "Hospital"). Taira RK et al [10] use semantic selectional restrictions to identify patient names in free text. A manually tagged training corpus is used to automatically determine semantic restrictions on the context around names. The algorithm then uses these semantic restrictions to determine fitness of candidate patient names in a testing corpus.

Among other de-identification efforts, Gupta D [11] devised a pattern-matching de-identification engine that utilizes a combination of rules and dictionaries, and the Unified Medical Language System. Sweeney L [12] employed templates and specialized knowledge of the context to replace identifying information in medical records.

Concept-match, like extraction, is aimed at scrubbing medical records, i.e. removing and/or transforming identifying and private information in free text. This method uses a list of words or terms that are neither identifying nor private and can harmlessly appear in the output text. The list generally contains high-frequency words in

the English language (e.g. verbs, articles) and medical terminology (e.g. names of diseases and medical procedures). Here we provide a summary of the implementation of a concept-match algorithm developed by Berman JJ [13]. The algorithm first parses the input text into words, and each word is compared to 2 reference lists: stop words and medical terminology. The word is left unchanged in the text if it matches a stop-word, i.e. a high-frequency word like a common adjective or article. If the word matches a medical terminology, the word or phrase is compared against the Unified Medical Language System (UMLS) database to discover its code. The medical term is then replaced by its UMLS Concept Unique Identifier and by a synonym. All other words or phrases in the text are absent in the allowed lists and are therefore removed and replaced in the text by asterisks. The output text thus consists of UMLS codes, synonyms for the original medical terminology, common English words and gaps, and is likely to be plagued by readability issues. The algorithm proved to be very fast, scrubbing half a million phrases in under an hour. The output text included only standard medical terminology and was considered safe for research use.

Both extraction and concept-match methods of de-identification have significant advantages and disadvantages, as discussed by Berman JJ [14]. In addition to searches of PHI known *a priori*, the extraction method relies on numerous textual pattern searches, e.g. *regex* rules in a text processing language like Perl. These pattern searches are more computationally intensive and time-consuming than simple unit searches. Thus extraction algorithms can become complex and slow for large systems. In addition, medical records of different formats, e.g. x-ray reports and discharge summaries, will require different sets of regular expression rules. Thus a large body of rules will have to be examined and edited every time for code maintenance and updates. The main advantage in using the

extraction method is the readability of its output. Since the default output is the original text and since words are removed only by careful consideration, the output closely matches the original text and there is a somewhat lower risk of false positives than with the concept-match method.

The concept-match paradigm, on the other hand, does not guarantee readability. The original text is not directly copied into the output text, and readability depends heavily on the comprehensiveness of the list of allowed terms. Removal of all non-allowed words invariably changes the readability and/or meaning of the text. Another major issue is the difficulty in compiling a comprehensive list of all non-PHI terms. If auto-coding is used for medical terminology, mistakes or changes in the style of auto-coding may lead to PHI disclosure. Numbers or numerical patterns which may appear in the text, e.g. Social Security Numbers, cannot be accommodated into the allowed list. Thus removal of numerical PHI will require some extraction rules that can recognize the particular patterns of these PHI. On the positive side, the concept-match method generally has very few or no false negatives since only allowed terms appear in the output. In addition, concept-match algorithms tend to be simple and fast.

Readability is of utmost importance for our de-identification project since our de-identified records will be directly used for medical research. The extraction method thus closely fits our research needs. In addition, we plan to de-identify hospital medical records of various formats for our research goals, and expect to change our concept of what constitutes a PHI for these different formats. We can accomplish these changes easily in our extraction algorithm by maintaining a constant body of direct search rules and slightly varying our set of regular expression searches. The extraction method is also

highly suitable for the numerous numerical PHI in our medical records, e.g. dates, ages, telephone/fax/pager numbers, Social Security Numbers.

2.2 Development of a De-identification Gold Standard at HST

A corpus of medical notes was thoroughly de-identified manually to set a gold standard for the de-identification algorithm to be developed in this project. This section of the project was conducted by Douglass M [15] at HST. The corpus consists of 2,646 nursing notes from 148 patients. It consists of approximately 350,000 words and includes 1,747 instances of PHI. Three clinicians from local hospitals were recruited to manually de-identify the same set of notes. The clinicians reviewed the notes and, based on the context of each PHI, labeled and classified all PHI, and suggested replacements.

The results of de-identification varied from clinician to clinician, and the results from all 3 clinicians needed to be compiled to remove the maximum number of PHI. A fourth clinician played the role of an adjudicator, looking over the suggestions of the de-identification results and compiling the 3 versions into one gold standard. A Java-based graphical user interface (GUI), illustrated in Fig. 2-1 from the thesis of Douglass M [15], was used to facilitate the process. This gold standard was used to evaluate the performance of the computerized algorithm developed in this project.

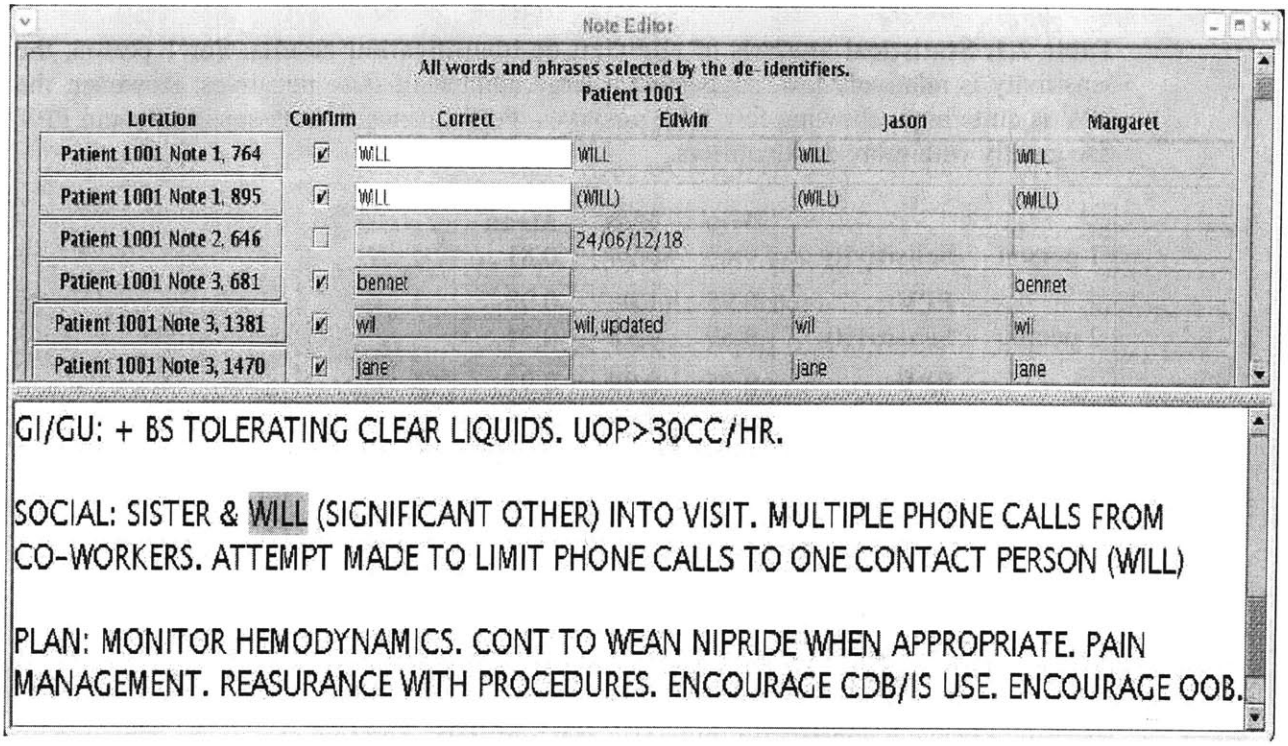

Figure 2-1. Java-based GUI for compiling a de-identification gold standard [15]. The text field in the bottom half presents the context of the highlighted PHI. The first column shows the PHI's location in the medical notes. The last 3 columns show PHI identifications by clinicians; the third column indicates the PHI that should be removed and re-identified based on the consensus of the 3 clinicians. The adjudicating clinician can confirm these PHI by ticking appropriate boxes.

In addition, statistics were calculated to estimate human performance at de-identification. These statistical values provide a standard of performance for the computerized algorithm. The sensitivity (proportion of PHI that are identified) and positive predictive value (proportion of identifications that are PHI) are presented in Table 2.1 for one, two and three human de-identifiers.

Table 2.1. Statistical analysis of clinician de-identification results. For 1 person, the sensitivity is relatively low, suggesting a large number of false negatives. However, the PPV is quite high, showing few false positives. Performance in both sensitivity and PPV rise greatly with more de-identifiers.

		Min	Max	Mean
1 person	**Sensitivity**	0.63	0.94	0.81
	PPV	0.95	1.0	0.98
2 people	**Sensitivity**	0.89	0.98	0.94
	PPV	0.95	0.99	0.97
3 people	**Sensitivity**	0.98	0.99	0.98
	PPV	0.95	0.99	0.97

As seen from the 1-person results, performance varied significantly among different de-identifiers. However, when more people de-identified the same notes, the sensitivity became consistently high at a mean of 0.98. The low sensitivity achieved by the 1-person de-identification suggests that the results contained a high number of false negatives, i.e., many PHI cases were not identified. Such errors may endanger patient confidentiality, and this evaluation suggests that the algorithm should perform significantly better than 1-person results.

2.3 Development of a De-identification Algorithm

Our current de-identification system is based on the algorithm implemented by Douglass M [15] for the MIMIC database. That algorithm used pattern-matching and lists of known PHI, and focused only on nursing notes. Evaluation with a nursing note gold standard revealed a high sensitivity of 0.98 which is on par with 3 people de-identifying. However, the algorithm had a very low positive predictive value of 0.44 which significantly reduced the readability of the de-identified text.

Our goal in continuing to improve this system was to refine the de-identification techniques and include more knowledge of known PHI. The tradeoffs between sensitivity

and PPV had to be carefully assessed to improve the text's readability. We also intended to extend the domain by tuning the algorithm to handle discharge summaries as well as nursing notes.

3 De-identification System

In this section we provide an overview of de-identification requirements, a summary of the de-identification process and techniques, and details about the innards of our system.

3.1 PHI Disclosure Risks

Medical records that have been de-identified by our algorithm still bear the disclosure risks outlined in the Statistical Policy Working Paper 22 [25]. The first type of risk stems from unusual or rare personal information that cannot be strictly categorized as PHI and that often reside in the social history sections of our medical records, e.g. patient's ethnicity. It might be easier to identify the patient from a record with knowledge that the patient is of a rare ethnicity. The second type of disclosure risk is due to the existence of secondary sources of information on the patient. Unusual non-PHI information in the de-identified text can be used in conjunction with news sources to narrow down or often reveal the exact identity of the patient. For example, "the patient's trailer was blown away by a tornado the night before Christmas" is a piece of text that does not contain any terms that are outright PHI, but a news-search could potentially reveal the identity of this relatively unique patient. Information on our research grant, i.e. the grant that has funded our de-identification project, is publicly available and includes the time during which we collected medical data from the participating hospital. This information makes it possible to estimate the time of hospitalization of a patient in the de-identified database. As a result, one can discover potential patient identities by searching news resources for mentions of accidents or health problems that occurred in the area during this time frame.

These disclosure issues are a result of unusual non-PHI text in the record and of the availability of personal information at secondary sources, and as a result cannot be

handled within our de-identification system. Even a perfectly de-identified database is susceptible to disclosures of these types. To minimize the risks of disclosure, the users of our de-identified records are required to sign a data use agreement that limits by whom the records can be used. In addition, we are also excluding all medical documents of VIP patients from our database specifically to minimize the second type of disclosure risk. Despite the risk of inadvertent PHI disclosure, it is not feasible to manually review every de-identified record to ensure removal of all PHI. In fact, Table 2.1 indicates that even a consensus of 3 expert de-identifiers is unable to remove all PHI from large bodies of text. However, we plan to routinely examine subsets of the database for inadvertent disclosure of PHI or patient identity. An important avenue for further work would be to devise an intelligent method to scrub non-PHI information that can indirectly jeopardize patient confidentiality.

3.2 PHI Categories

The Health Insurance Portability and Accountability Act (HIPAA), enacted by the U.S. Congress in 1996, specifies regulations to preserve the confidentiality of protected health information (PHI). In compliance with HIPAA regulations on medical information release, we identify the following PHI categories associated with both living and dead patient(s) [5]:

- Names
- All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census, (1) the geographic unit formed by combining all zip codes with the same three initial digits contains more than 20 000 people; and (2) the initial three digits of a zip code for all such geographic units containing 20 000 or fewer people is changed to 000

- All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death;
- All ages over 89 years 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 years or older
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Web Universal Resource Locators (URLs)
- Internet protocol (IP) address numbers

The MIMIC II database development is a major component of an NIH-funded research program entitled "Integrating Data, Models and Reasoning in Critical Care" (Grant Number R01 EB001659). The data collection is a collaboration effort involving MIT, the Philips company and a local hospital. We have a responsibility to not only protect patient information in our text files but also information specific to the hospitals and providers. Besides, hospital-specific information can narrow down the subset of patients that can be identified from a set of PHI. For example, the subset of 49-year-old patients in Ward A of Hospital X is significantly smaller than that of 49-year-old patients anywhere. To this end we add the following additional PHI categories:

- Any names other than patient names, e.g. provider names
- Any ages over 90

- Hospital names
- Ward names
- Ethnicities/nationalities
- PHI of any other HIPAA category relevant to hospitals or providers

3.3 Text Structure

Creating a general de-identification system for any type of text will be a difficult task. Without clear specifications about what the term "PHI" should constitute and the particular requirements of the text, such a system will have a lower than optimum sensitivity – the proportion of all PHI it can identify - and positive predictive value – the proportion of its identifications that are correct. We have thus restricted our domain to 2 major types of medical records: nursing notes and discharge summaries. These records have a format and structure that is fairly consistent from hospital to hospital; hence our domain will be relevant for any medical research that involves free-text medical records.

Before a de-identification system can be tuned to the particular needs of the text, it is imperative to conduct an analysis of its structure. In this section we present our analysis of nursing notes and discharge summaries in the belief that it will clarify the rationale behind our strategies and will help other de-identification systems for similar records.

3.3.1 Discharge Summaries

A discharge summary is a documentation of the patient's medical condition, history, medical stay and physiologic condition, created at the time of the patient's discharge from the hospital. It consists of clearly demarcated fields, most of which consist of free

text. A sample de-identified discharge summary is presented in Appendix A, and a skeleton follows below.

NAME:	UNIT NUMBER:
ADMISSION DATE:	DISCHARGE DATE:
DATE OF BIRTH:	SEX:
SERVICE:	
HISTORY OF PRESENT ILLNESS:	
PAST MEDICAL HISTORY:	
ALLERGIES:	
MEDICATIONS:	
SOCIAL HISTORY:	
PHYSICAL EXAMINATION:	
LABORATORY DATA:	
HOSPITAL COURSE:	
DISPOSITION:	
DISCHARGE STATUS:	
CONDITION ON DISCHARGE:	
DISCHARGE DIAGNOSIS:	
	<NAME OF ATTENDING PHYSICIAN>
DICTATED BY:	

Figure 3-1. Discharge summary format. Each discharge summary begins with some structured patient information fields. The body of the summary is segmented into different content categories, but the text is not structured or formatted. The summary ends with the full names of the attending physician and the person dictating the text.

The beginning of each discharge summary lists patient information in a formatted pattern. This section is rich in PHI which we identify using the field-names that are commonly used. For example, upon matching the term "date of birth:" our algorithm targets the following characters for a search of date patterns, and thus effectively removes the patient's birth date. Similarly, the name of the person dictating the record is removed by identifying the phrase "dictated by:" and removing the words that follow it.

Among the free-text fields that occur in the body of the discharge summary, the History of Present Illness and Social History fields are especially risky. The first field is rich in the names of hospitals and various dates related to the patient's medical history; while the latter is likely to contain references to family members and locations. Our algorithm does not use different or more aggressive strategies for either of these fields. Based on our performance statistics and manual evaluation, we deem that our system is adequately removing these PHI. However, future work could consist of fine-tuning a de-identification to use more aggressive techniques for the riskier discharge summary fields.

The History of Present Illness section contains PHI, e.g. date of a heart attack or of a hospital admission, that is vital to understanding the patient's medical condition. The same is not true for the types of PHI more frequent in the Social History field. PHI like the name of the patient's spouse is not likely to be useful to the ultimate users of our de-identified product – the medical researchers. This fact coupled with the risk inherent in the Social History has led us to the decision to discard this field altogether. Thus, our efforts to make discharge summaries available on MIMIC II will sacrifice the integrity of the records for a significantly reduced risk of PHI disclosure.

Fig. 3-2 presents a breakdown of the PHI categories and their average frequencies in discharge summaries.

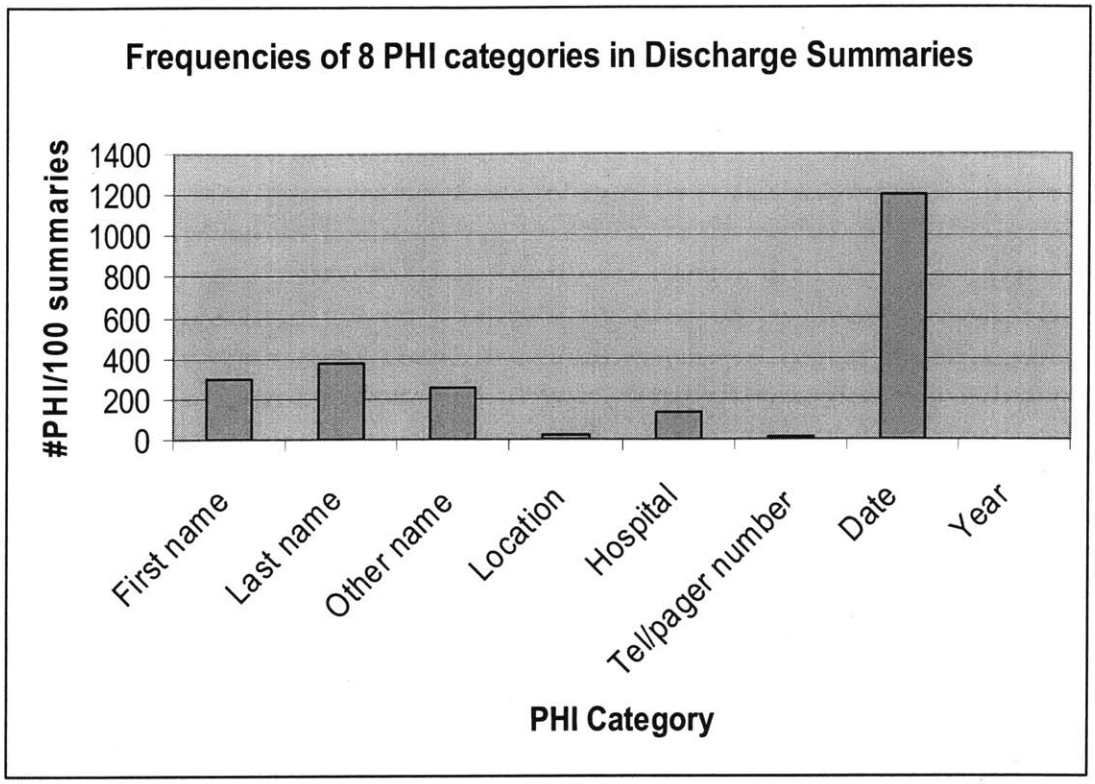

Figure 3-2. Frequencies of different PHI categories in discharge summaries. In the subset of discharge summaries we analyzed, dates were the most common PHI. Patient, provider and hospital names were also fairly common.

3.3.2 Nursing Notes

A nursing note is a periodic documentation of the patient's progress in the hospital. The nursing note is much less structured than the discharge summary or any other medical record, and its body consists entirely of free text. Appendix B presents a sample de-identified nursing notes, while a skeleton follows in Fig. 3-3.

```
<HEADER>
<BODY>
<ENDING SEGMENT>
```

Figure 3-3. Nursing note format. Each nursing note begins with a header that contains date PHI, and ends with a segment that does not contain PHI. The bulk of the note lies in the free-text area that occurs between these 2 segments. For our purposes we strip off the header and ending segment, and run our algorithm on the body of the nursing note.

The de-identification algorithm deals only with the body of the nursing note which is basically a chunk of free text. Even though a nursing note does not have clearly demarcated fields, the nurse (responsible for taking the notes) might create separate sections within the note. The SOCIAL section contains the social history or information of the patient, and is the most "dangerous" section for our purposes. Similar to our procedure with discharge summaries, we remove this section wherever it occurs in the note. The section normally spans one paragraph and is started off by a sub-heading, e.g. "SOCIAL", "SOCIAL/DISPO", or "FAMILY". When the algorithm identifies such a sub-heading, it removes all text from the current paragraph.

Fig. 3-4 presents a breakdown of the PHI categories and their average frequencies in nursing notes.

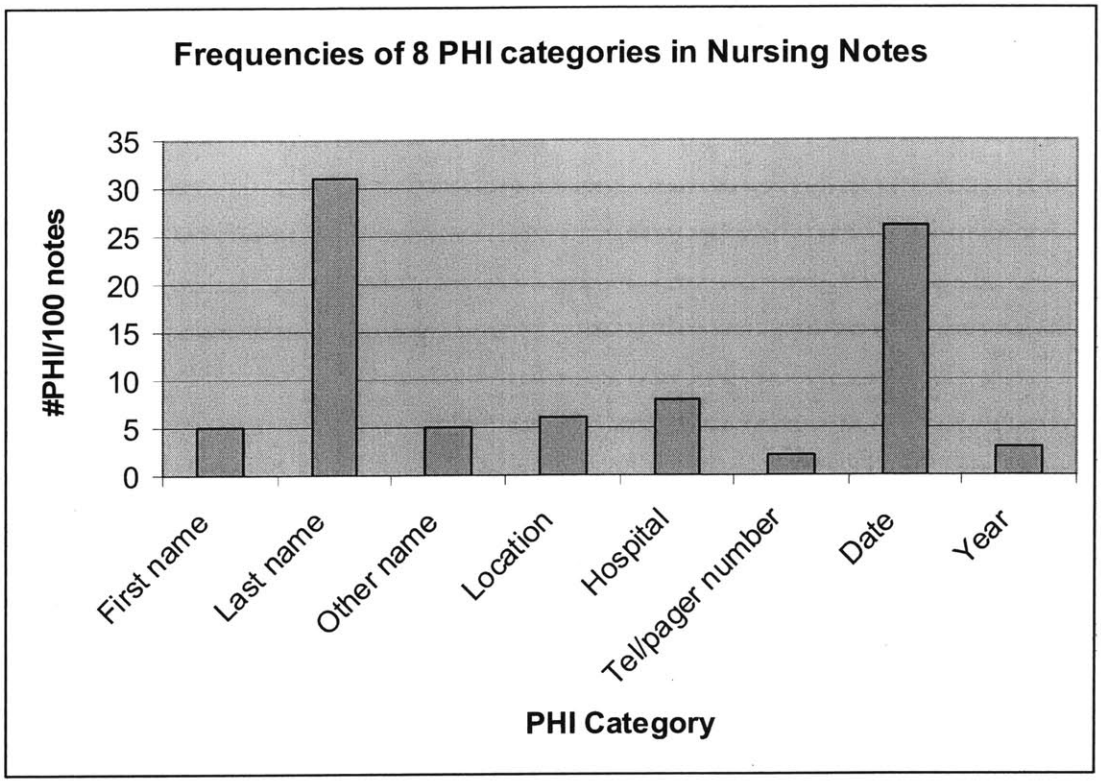

Figure 3-4. Frequencies of different PHI categories in nursing notes. Dates and last names are very common, and these notes are also expected to contain most of the other major PHI categories. However, comparison with Fig. 3-2 will show that PHI frequencies in nursing notes are greatly lower than those in discharge summaries. Discharge summaries are thus the significantly riskier of the two types of medical records.

3.4 Design Overview

The current de-identification algorithm is implemented in Perl, a high-level programming language that supports many types of string manipulation, on a Linux platform. De-identification involves parsing the entire text to identify PHI, classifying each item of PHI based on the HIPAA categories, and replacing it with a PHI category tag in a process called re-identification. This process is illustrated in the flowchart in Fig. 3-5.

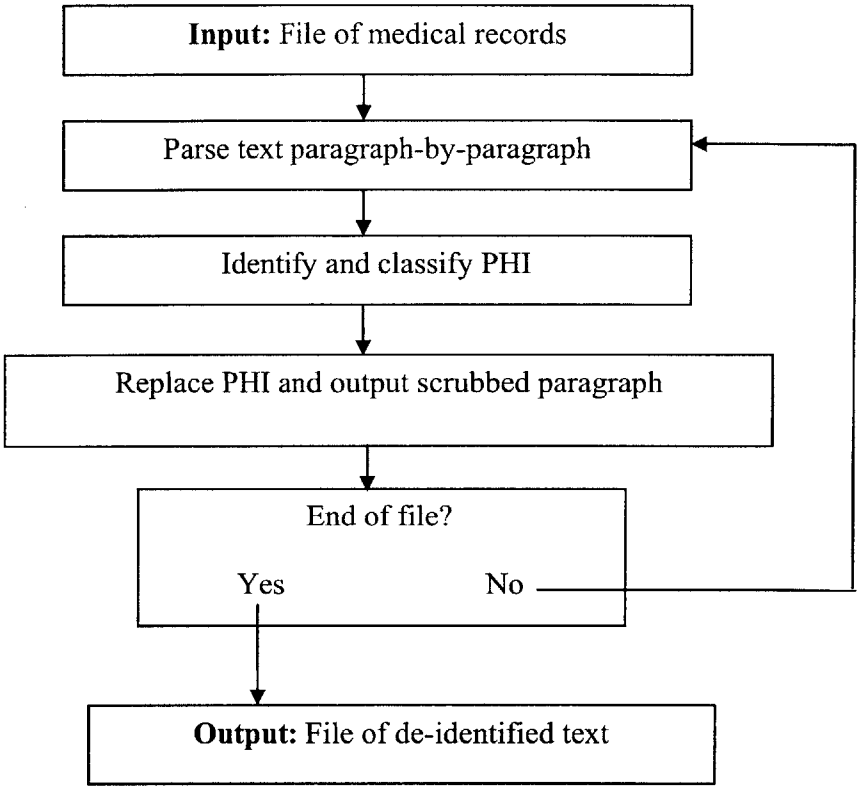

Figure 3-5. Flowchart of the de-identification algorithm's operation

The techniques used to de-identify the text and identify its PHI can be broadly categorized into 3 groups. One technique involves using lists of known PHI, e.g. patient and hospital names, to directly remove PHI that are known *a priori*. The main de-identification efforts, however, center on pattern-matching. Most categories of PHI have a particular format, e.g. telephone numbers are generally of the format (nnn) nnn-nnnn and its variants. Additionally, PHI often occur in certain textual contexts, e.g. a last name is likely to be preceded by "Mr.", "Dr", etc. Our algorithm uses comprehensive format and context information relevant to each PHI category to match patterns that are likely to divulge PHI. Finally, another strategy related to the use of known PHI lists is the use of

lists of potential PHI, e.g. common words that can be used as names, to guide the pattern-matching process. For example, when the algorithm matches the textual pattern "Dr", it uses a list of potential first and last names to definitively determine if the words following "Dr" are names. Details of these techniques follow in Section 3.4.

Two features of the de-identification algorithm proved essential in its rapid prototyping, development and consistency in performance. First, the algorithm is modular with each non-interacting module recognizing the patterns of a different PHI category. As a result, the algorithm's performance at recognizing dates, for example, does not affect its performance at name recognition. This modular design has allowed rapid, iterative changes to the algorithm based on repeated testing. Second, each portion of text is examined by all the non-interacting modules. Hence, the order in which different modules parse the text is immaterial. The modules can be laid down in any linear order for the algorithm to consistently de-identify the text, thus keeping the algorithm easy to design, modify and parallelize.

3.4.1 Execution

The de-identification algorithm resides in a single Perl file. All relevant files and folders that are included with the source code must reside in the same directory as the algorithm file. Running the algorithm requires only 1 input: the raw text file. The system can be run very easily by executing the following Perl execution command.

=> perl deid.pl <filename>

A command-line interface then takes over the function of interacting with the user. The system queries the user about the PHI categories that are to be de-identified and

the dictionaries/lists to be used in the process, as presented in Appendix C. De-identification then produces the following output:

- Filename.res: The raw text with PHI replaced
- Filename.phi: Text locations of the identified PHI
- Filename.info: Additional information about PHI and suspected PHI
- Printout of performance statistics if user has asked for comparison with a gold standard

The previous version of the algorithm was too slow for mass de-identification. We replaced some lookup tables with hashes, an adjustment that significantly increased execution speed. Currently the algorithm de-identifies around 22 nursing notes (approximately 4,000 words) per minute, and this speed is sufficiently high for our purposes. The MIMIC II database currently houses several types of records for 17,000 patients, and de-identifying the database in its entirety still requires time on the order of days. Distributing the task among different computing clusters has proved invaluable in speeding up the process.

3.4.2 Usability

We have already laid out an overview of the algorithm and its execution. In this section we will address the usability of the system.

A user interacts with the system for two main tasks: running the de-identification algorithm on some text file, and computing performance statistics by comparison with a gold standard. The system's user interaction strategy is based on these two central tasks, and aims to minimize queries to the user and to simplify user interaction.

3.4.2.1 User task 1: Running algorithm

The software is run and communicates with the user through a simple command-line interface. The interface is minimalist, and the number of queries to the user is kept to a minimum. A sample run is presented in Appendix C.

As we mentioned in our system overview, the algorithm is partitioned into separate modules for the different PHI categories. We have implemented a switch for each module that can be turned on or off by the user before the algorithm executes. This measure has been taken in an effort to make the de-identification system more universal, and to allow user control and freedom. Users of our system may not want de-identification of the whole suit of PHI categories. For example, if the exact dates of a patient's medical events are important, the user may switch off the date identification module in our algorithm. At the outset, our interface lists the PHI categories and asks the user if the default suit of PHI categories should be used. It asks about each category separately only if the user answers negatively to this question. Thus, only users who require customized de-identification specifications are taken through the series of questions about which modules to turn on. Similarly, users can also specify which lists and dictionaries to use during de-identification.

Claredi, a company that delivers health information technologies, has released a library of codes that scrub different PHI categories online [16]. Most of these PHI categories, e.g. National Association of Boards of Pharmacy Number, Treatment Codes, do not occur in our medical records and our algorithm does not have modules to de-identify them. However, it is likely that other research groups using our algorithm might

need to de-identify documents containing these PHI. Rather than implementing new code scrubbers to tackle these PHI, any set of these independent code sets from Claredi can be plugged into our main algorithm to suit different de-identification requirements.

The system status is visible at all times during a run of our algorithm since de-identification of large text files could become a lengthy process. After de-identifying an individual record, our interface prints out a message on the record's identifier. As a result, user frustration and confusion is avoided.

3.4.2.2 User task 2: Computing performance statistics

The software contains an independent module that computes performance statistics in comparison to a de-identification gold standard. This module requires that a gold standard – i.e. a list of all PHI locations in the given text - be available for the file being processed. The first step in computing performance statistics is to actually run the de-identification algorithm on the file. The results of the run are then compared against the gold standard in the module in question to output performance statistics. This performance statistics calculation module is automatic and acts as a black-box, thus allowing the user to easily verify the system's performance at the same time that a file is being de-identified.

Since the first step in the process is the same as the first user task, the system uses the same command-line user interface. In addition to accomplishing the goal of running the algorithm, the user can specify an option for computing performance statistics after the algorithm has been run. This constitutes a simple question in the series of queries the user has to answer. A sample interaction follows.

```
**********************************************************************
De-Identification Algorithm: Identifies Protected Health Information (PHI) in Discharge
Summaries and Nursing Notes
**********************************************************************
Comparison with Gold Standard: Press '0' for no comparison or '1' for statistical
comparison with existing Gold Standard, and then hit return.
If unsure, press '0'.
=>1
```

Figure 3-6. Command-line interaction between the de-identification system and the user to determine whether performance statistics are to be calculated. In this case, the use responds in the affirmative.

The software then automatically compares de-identification results and the gold standard, and outputs the following statistics about its performance:

- Number of false positives
- List of false positives
- Number of false negatives
- List of false negatives
- Sensitivity
- Positive Predictive Value (PPV)

3.5 De-identification Algorithm

As previously described, the algorithm parses through the text and attempts to de-identify the relevant PHI categories. The following sections describe the modules used to identify and replace each type of PHI.

3.5.1 Algorithm Modules

As previously mentioned, the algorithm consists of independent modules that deal with the different PHI categories. Each module uses any combination of the 3 general de-

identification techniques we outlined in Section 3.3: removal of known PHI, pattern-matching, and the use of lists of potential PHI to complement the pattern-matching process. Here we break up our discussion of the modules into separate subsections.

3.5.1.1 Names

Names directly identify a patient or provider, and are the single most risky PHI category. To ensure maximum name extraction, we have fine-tuned this module to the information in our database. The MIMIC II database contains the full patient and provider names associated with each medical record. We are thus able to extract the full or partial names of patients and providers specific to each record. This record-specific extraction almost guarantees suppression of patient name information in the text, while minimizing false positives.

Medical records often contain names of patient's relatives, nurses, etc, that can be used to narrow down the set of identifiable patients. Thus the above technique is not sufficient by itself. We use context information and lists of possible PHI to extract the remaining names. Most last and full names are preceded by titles like Mr., Mrs., Dr, etc. If the algorithm identifies any such title that normally precedes a name, it considers the following 3 words to check for potential first names, last names and initials. It matches each potential name and all other words in the text with a large list of first and last names compiled from US Census reports. If a match occurs, the word is classified as a name PHI. Even if a match is not found, the algorithm uses context information and information about whether the word is common to determine if it is a name. Additionally, the algorithm identifies any individual uncommon words that match potential names in our name lists.

3.5.1.2 Dates

Discharge summaries and nursing notes are rich in dates. HIPAA stipulates that all date PHI pertaining to patients except years, e.g. birth, date, admission, discharge dates, be scrubbed during de-identification. It is difficult for an automatic algorithm to determine whether a date pertains to a patient; we therefore remove and replace all dates from our text. Dates generally follow the specific formats listed in Table 3.1; the table also indicates if the algorithm considers context during identification. The algorithm tries to match any of the formats in the text, and considers contextual information in some cases before identifying the text as date PHI.

Table 3.1. Date formats identified in text. Dates can include any combination of day, month and year, and separate years. We use knowledge of the common formats to extract dates by pattern-matching. Context information is taken into account only in the case of separate years, since this category is more likely to yield false positives.

Type of date	Date format	Context information
month/day/year	*mm/dd/(yy)yy*	No context checked
day/month/year	*dd/mm/(yy)yy*	No context checked
year/day/month	*(yy)yy/dd/mm*	No context checked
year/month/day	*(yy)yy/mm/dd*	No context checked
day month, year	*dd month, (yy)yy*, e.g. 2 Jan(uary), (19)96	No context checked
month day, year	*month dd, (yy)yy*, e.g. Jan(uary) 2, (19)96	No context checked
year	*yy, yyyy*	Identified as PHI only if preceding or following text includes name of medical event, e.g. CABG 1996

Years are not considered PHI according to HIPAA regulations. However, years associated with other medical information can reveal when the patient experienced a landmark medical event. For example, the mention of 'CABG 1996' in a nursing note divulges that the patient had a Coronary Artery Bypass Graft in 1996. Given this information it is possible to narrow down the subset of patients at a hospital who had this procedure in 1996. Thus, we remove all instances of years in addition to the HIPAA-specified date formats to make our de-identification standards more stringent.

We replace all PHI other than dates with a PHI category tag, e.g. the name 'John Brown' is replaced by <Full name>. Dates, however, are necessary to track the patient's stay at the hospital and the evolution of his/her medical condition. For example, it is important for a medical researcher to know the duration of the patient's stay in the hospital, the delay after a bypass surgery that the patient was re-admitted, etc. Our algorithm therefore automatically re-identifies dates, preserving the day of the week and season. Each date is shifted by a patient-specific random number of days that is consistent for the patient throughout all his medical files. The mapping between patient ID and PID-specific date shifts is kept in an encrypted file that is not for release. Our date replacement scheme uses patient-specific date shifts to reduce the risk of an accidental release of a constant hospital-wide date shift. A malicious reader of the text, equipped with such a constant date shift, would be able to recover the original dates for all patients in the hospital. However, in our scheme, release of the PID-specific date shift would divulge only the dates in that particular patient's files.

The original date format in the text is preserved: e.g. mm/dd/yyyy is replaced by a shifted mm/dd/yyyy date, whereas an individual year is replaced by a shifted individual year. As a result, the patient's progress can be tracked by checking the days of the week.

In addition, we preserve the season which may have an impact on the patient's condition and treatment, and may be interesting to the medical researcher perusing the de-identified text.

A problem is posed by textual references to significant calendar events that might disclose the dates of the events. For example, knowledge that a certain patient was admitted to the hospital on Christmas Eve will significantly reduce the subset of matching patients. Although HIPAA does not specify such textual references to dates, we deem these important in preserving date information. The date module of our algorithm thus searches for patterns that match the names of significant calendar events, notably Christmas, Thanksgiving, Easter, Hanukkah, etc, which are then scrubbed.

3.5.1.3 Locations

HIPAA specifies that the de-identification process scrubs all location identifiers more geographically precise than the state-level. Our algorithm is customized to the de-identification requirements of MIMIC II discharge summaries and nursing notes. A thorough examination of these files revealed that they generally do not contain street addresses or street-level location data. The vast majority of location references constitute names of cities, counties, highways, lanes, etc. Since the MIMIC II database contains patient information from only local hospitals, neighboring locations are more likely than others to appear as PHI.

We employ the following strategies to remove the maximum number of location PHI.

- Using lists of known locations
- Identifying locations from textual context

For the first strategy, we compiled 2 lists of locations: a list of words that are unambiguous locations (e.g. Chicago), and a list of words that are ambiguous locations (e.g. Anchorage which can also be a common noun). These lists contain the names of major cities in America, major cities and locations in the world, places in the Massachusetts area, etc. The algorithm matches each word or phrase in the text against these lists to identify possible location PHI. Our location lists are sufficiently comprehensive to identify mentions of most city or county-level locations in the records. In addition, the algorithm checks if the spelling in the text is approximately close to the spelling of a location in the lists. Free text is hand-typed and contains frequent misspellings. Using this approximate-matching technique makes it possible to identify misspelled locations, e.g. "Chicage" which is approximated-matched to "Chicago". For re-identification, the algorithm currently replaces the location with a *<Location>* tag.

There is a risk in the above technique of missing locations that are not included in our lists of known locations. In addition to separate mentions of geographical subunits like cities, counties, etc, the records also contain more specific references to where patients may be from. These references can include names of lanes, highways, etc, and are too numerous to list exhaustively. This issue necessitates a more general de-identification technique of analyzing textual context to identify location references.

When mentioned in free text, locations are generally preceded or followed by words that indicate locations. This is analogous to a hospital name, e.g. "Mount Sinai" being followed by the term "Hospital". Some of these contextual terms are presented in Table 3.2. When the algorithm identifies such a contextual term, it checks the neighboring words. If these neighboring words turn out to be ambiguous locations or

uncommon words, they are identified as location PHI. This technique successfully removes almost all locations that were not identified by our first list-search technique.

Table 3.2. Context terms that identify locations. Certain nouns, commonly integrated with location names, reveal potential locations in their neighboring text. We use either a preceding or a following term to identify any potential location. It is noteworthy that words that exist in the list of unambiguous locations are extracting without undergoing this context check.

Terms following locations	Terms preceding locations
Street	Cape
Parkway	Fort
Town	Lake
Ville	Mount
Harbor	Los

3.5.1.4 Hospital-Specific Information

It is crucial to remove hospital identifying information before releasing medical records to the general public. Using this type of information, a malicious individual can narrow down the set of matching patients from anywhere in America to a specific hospital. Our algorithm thus removes all hospital names. It also removes ward names which are in some cases hospital-specific and can reveal the names of the hospitals themselves.

The hospital name recognition module functions very similarly to the location recognition techniques. First, the module matches each word in the text to a comprehensive list of neighboring hospitals/clinics/medical facilities. Second, it searches for context information that is likely to be preceded or followed by a hospital name, e.g. "hospital", "rehab center", etc. These 2 techniques make our hospital name recognition very sensitive.

Since MIMIC II records are obtained from only one local hospital, all ward names occurring in the records are specific to that hospital. We have compiled a comprehensive

list of all ward names in the hospital, and identify ward names in the records by simply performing a match against this list.

3.5.1.5 Telephone, Fax and Social Security Numbers

Patient or provider identities can be easily tracked down from telephone, fax, or Social Security Numbers (SSNs) released in their files. In fact, the risk factor associated with released SSNs can be ranked with the risk associated with released full patient names. Telephone numbers, fax numbers, and SSNs follow specific formats, some of which are presented in Table 3.3. To check whether these formats match important non-PHI information in the text, we have conducted an analysis of other numerical patterns, e.g. heart rates, blood gas data. Fortunately, the listed formats are specific to telephone/fax numbers and SSNs, and can be used to eliminate these PHI categories with a low rate of false positives.

Table 3.3. Telephone/fax and Social Security Number formats identified in text. These numerical PHI can potentially be confused with other non-PHI terms, e.g. medical data. We use certain known formats to extract these numerical PHI to reduce the number of false positives.

Telephone/Fax numbers	Social Security Numbers (SSNs)
(nnn) nnn-nnnn	nnn-nn-nnnn
nnn-nnn-nnnn	nnn-nnnnnn
nnn nnn nnnn	nnnnn-nnnn
nnn-nnnn	nnnnnnnnn
nnn nnnn	

The modules of our algorithm that handle these PHI categories parse the text line-by-line. Upon matching one such format, the module labels the matched text as PHI and tags it with its PHI category. Thus, currently, telephone/fax numbers are replaced in de-identified text as <Telephone/Fax number> and SSNs as <Social Security Number>. During our later re-identification efforts, the numerical values in the PHI will be replaced

by randomly-generated numbers that fit the format of the PHI. For example, the phone number (333) 333-3333 in the original text might be re-identified as (777) 777-7777.

3.5.1.6 Ages over 89

The patient population at any hospital over the age of 89 is significantly low. Given the release of some additional PHI (e.g. the first name of such a patient), it will be relatively easy to track the identity of a patient in this limited subset. For example, there may be more than 10 'Bobs' registered as patients in a large hospital, but there may be only 1 'Bob' who is over the age of 89. We have considered it important to remove this PHI category not just when they are related to patients but wherever they may occur in the text.

A simple-minded approach would be to identify any number over 89 that occurs in the text. This method results in a large number of false positives since many numerical medical data take the format of nn and nnn. We have devised 2 techniques that accurately identify almost all ages over 89 without contributing to the false positive rate. The relevant module of our algorithm searches for either numerical or text patterns that fall within an age range. Therefore, it will identify an age expressed either as '95', 'ninety-five' or 'ninety five'. We limit the age range to 90-125. The upper limit is introduced as a sanity check since it is highly unlikely that a patient's age will exceed 125. A larger number is likely to be some other medical term. Additionally we do a contextual analysis for each age identified. An overview of the characteristics of our nursing notes and discharge summaries reveals the following terms that generally surround ages.

Table 3.4. Context information that identify patient age. Age can be confused with numerical medical data. It is common knowledge that references to age either precede or follow some semantic context. We therefore identify numbers within a meaningful range as age only when they occur in the neighborhood of the context in this table.

Textual context preceding age	Textual context following age
Age (e.g. patient is of age 90.)	year old, year-old, -year-old
He is (e.g. he is 120.)	years old, years-old, -years-old
She is (e.g. she is 120.)	years of age, yrs of age
Patient is (e.g. patient is 125 today.)	y.o., yo,

Therefore, the module identifies numbers in the range of 90-125, and identifies only those that are either preceded or followed by some textual context that indicates it is an age. This scheme has been successful in identifying almost all ages without mistaking other numerical information as PHI. The identified age is then replaced by a general <Age over 89> tag that aggregates all ages over 89 into a single group to preserve confidentiality. This replacement still presents the age information in a way that is likely to be helpful in understanding the patient's specific condition.

3.5.1.7 Other PHI Categories

Emails and Uniform Resource Locators (URLs) are rare in MIMIC II medical records. However there is a likelihood of these categories cropping up in latest and future records. Our algorithm thus de-identifies both these PHI categories.

3.5.2 Misspellings and Abbreviations

Misspellings and typos pose one of the most challenging problems for a de-identifying system. Free-text medical records are created by a nurse or medical practitioner jotting down facts and thoughts about the patient or by a transcriber manually typing handwritten records. This technique of free-text creation poses 2 problems which we elaborate on in this section. However, these problems are likely to exist for any free-text

database, and the lessons we learned in implementing our system are likely to be helpful in the design of other de-identification systems.

First, due to a lack of rigid format or quality specifications, the records contain misspellings and typos which make it difficult to identify PHI and also which lead to false identification of non-PHI. Misspellings can thus lower both the sensitivity and positive predictive value of the system. For example, a naïve de-identification system will not pick up "Chicage" which might be a misspelled version of "Chicago". On the other hand, a common word like "organ" misspelled as "morgan" will lead to a false identification of the word as a name.

The second problem inherent in free-text records is the use of medical acronyms and general abbreviations. Almost all of these abbreviations are textual, not numerical, and are likely to be mis-identified as names by a de-identification system. As a result, the positive predictive value of a de-identification system can suffer significantly. For example, the frequently occurring medical term "moving all extremities" is recorded as "MAE". "Mae" is a potential first name, and a system that is agnostic about this medical abbreviation will mistakenly label each occurrence of it as a name. We have compiled a list of medical terms and abbreviations common to our medical records, and eliminated these terms from our lists of known first and last names. As a result, any of these terms appearing independently in the text will not be tagged as PHI. Our job was made easier by the fact that there were very few terms that were both legitimate medical terms and names.

3.5.3 Un-identified PHI Categories

Two major PHI categories not handled by our algorithm are biometric identifiers (finger and voice prints) and photographic images. Patient biometric identifiers uniquely reveal

the identity of the patient, and their release is highly risky. Photographic images may not divulge any other personal information about the patient, but reveal a crucial aspect of identity - what the patient looks like. Currently MIMIC II does not include files that contain these 2 PHI categories. As a result, our de-identification efforts are focused on textual PHI as presented in this section. However, we recommend that any other files that incorporate images or biometric identifiers be thoroughly scrubbed of these types of PHI.

3.5.4 Lists of Known PHI

The algorithm uses several lists of known PHI and of dictionary words. Each list is stored in its own text file that is separate from the main algorithm. These lists can be easily modified or replaced, making the system amenable to updates and use for other types of text files.

3.6 Re-identification

Re-identification has the potential to significantly reduce risks associated in disclosure of de-identified data. Readers will find it nearly impossible to distinguish between a PHI that is left un-identified by the algorithm and re-identified information. Whatever false negative PHI we may have will be scattered among numerous fake PHI in the re-identified files, thus reducing risks of disclosure. Additionally, re-identification will thwart attempts to recover the individual's identity by investigating the intersection of all identifying facts, since the fake facts will no longer refer to real individuals. By replacing only identified PHI, re-identification avoids altering the readability and information content of the data.

In our current re-identification scheme, after running our de-identification algorithm on the MIMIC II data, we replace each PHI with a tag specifying its PHI

category. This tag preserves the information content of the original PHI without potentially harming patient confidentiality. For most PHI, it is sufficient to know the PHI category to parse the meaning of the sentence it occurs in. For example, the re-identified phrase "Patient was sent to <Hospital> at <Location> today" is readable, contains all the useful information in the original phrase, and preserves protected information. Thus, in our current scheme, we replace each PHI with a tag that labels its PHI category, as summarized in Table 3.5. Dates form the only PHI category that fall outside this replacement scheme. Dates are shifted by a patient-specific random amount; the original dates are then substituted by these replaced dates with the date format preserved. For example, "2006/4/5" might be replaced by "2009/6/24", and "3/2/2005" might be substituted by "6/21/2012". The rationale and details of the date substitution technique are presented in Section 3.4.1.2.

Table 3.5. Re-identification scheme. We are currently unable to replace PHI with representative fake PHI. Instead we use a PHI category tag to replace each identified PHI. The mapping between the PHI category and its tag is presented in this table.

PHI Category	Example	Re-Identification Tag
Full name	John Doe	[**Full name**]
Last name	Doe	[**Last name**]
First name	John	[**First name**]
Location	Cambridge, Memorial Drive	[**Location**]
Hospital name	Mount Sinai	[**Hospital**]
Ward name	Ward E	[**Ward**]
Full date	2006/3/4	[**2010/4/2**]
Partial date	3/21	[**5/3**]
Year	2006, '97	[**2010**], [**2001**]
Age over 89	93	[**Age over 89**]

An avenue for future work will be to refine this PHI replacement strategy. Each PHI could be substituted by a fake representative term from the same PHI category. For example, the name "Jane Brown" could be re-identified as "John Doe", while the phone

number "666-666-6666" could be replaced by "111-111-1111". This technique should ensure that recurrent PHI within the same file is replaced by the same fake PHI to preserve coherence. In order to prevent confusion, it should prevent different PHI from mapping to the same fake PHI.

4 Evaluation

4.1 Performance Criteria and Algorithm Testing

A text parser was developed to compare the results of de-identification. The gold standard is a corpus of fully de-identified nursing notes and has an associated record of all PHI locations identified by 3 clinicians. The testing algorithm parses the de-identified text to analyze these PHI locations. The text at each PHI location is compared with that in the original text to determine whether the algorithm has successfully identified and replaced the PHI. This comparison is used to determine the proportion of PHI de-identified (sensitivity) and the number of false negatives. In addition, the non-PHI locations of the Gold Standard are compared with the corresponding locations in the algorithm's result. This comparison is used to detect if the algorithm has identified non-PHI words, and to estimate the positive predictive value (PPV) with the number of false positives. The action of the algorithm on each PHI is categorized as true positive, true negative, false positive or false negative.

4.2 Comparison with Gold Standard

4.2.1 Procedure

We tested the algorithm's performance during its iterative development by comparing it against the gold standard de-identified database. The gold standard is a corpus of 2,646 medical notes from 148 patients that has been thoroughly de- and re-identified by the consensus of 3 experts. It consists of approximately 350,000 words and includes 1,747 instances of PHI. This database is considered to have a perfect sensitivity of 1.0 and positive predictive value of 1.0.

In this comparison, we first run the de-identification algorithm on the same raw dataset as the gold standard. We then run comparisons between the list of PHI identified in the algorithm's output and the list of PHI in the gold standard. We determine the number of false negatives (FN) by counting up those PHI that were identified in the gold standard (i.e. that are unambiguously PHI) but not identified by the algorithm. We determine the number of false positives (FP) by counting up those PHI that were identified by the algorithm but were not identified in the gold standard (i.e. are unambiguously non-PHI). PHI identified by both gold standard and the algorithm are counted as true positives (TP). Using the values for FN, FP and TP, we compute the following performance statistics:

- Sensitivity = TP / (TP + FN)
- Positive Predictive Value (PPV) = TP / (TP + FP)

4.2.2 Consistency Issues

The de-identification system stores its identified PHI as a list of PHI locations, with each PHI being represented by its location (starting and ending locations) in the text. This representation introduced some consistency issues in the comparison procedure. For some PHI types, the gold standard considers different parts of the same PHI (e.g. the area code and the local extension of a telephone number) to be different PHI. Therefore, it lists more than one PHI location for what is essentially a single PHI. For example, the telephone number "617 225 6598" is identified as 3 PHI locations corresponding to "617", "225" and "6598".

The algorithm, on the other hand, maintains the integrity of multi-part PHI. In our example, it would store the telephone number as a single PHI location corresponding to "617 225 6598". Thus, comparison of PHI locations in the gold standard and the

algorithm's lists could lead to inconsistencies, and could thus raise the number of false negatives and false positives artificially. This issue never overcasts the sensitivity or PPV of the algorithm, and is safe from the standpoint of performance evaluation. We addressed this issue by standardizing the PHI identification scheme to closely match the one used in the gold standard.

4.2.3 Results

We compared the algorithm's performance against the gold standard numerous times during the iterative development. Final results obtained at the end of the algorithm's development and testing cycle indicate a sensitivity of 0.87 and a PPV of 0.63. A general discussion of the algorithm's performance follows in Section 4.3.3. Table 4.1 presents the results of the algorithm and of human de-identification.

Table 4.1. Sensitivity and positive predictive value of human and algorithm de-identification.

		Min	*Max*	*Mean*
1 person	***Sensitivity***	0.63	0.94	0.81
	PPV	0.95	1.0	0.98
2 people	***Sensitivity***	0.89	0.98	0.94
	PPV	0.95	0.99	0.97
3 people	***Sensitivity***	0.98	0.99	0.98
	PPV	0.95	0.99	0.97
Algorithm	***Sensitivity***	-	-	0.87
	PPV	-	-	0.63

The algorithm has a sensitivity of around 0.87. This level of sensitivity was arrived at after evaluating tradeoffs between false negatives and false positives in the algorithm. An analysis of the false negative PHI categories revealed that the majority of escaped PHI was of medium to low risk and that zero patient names escaped in our

evaluation sample. Section 4.3.3 presents a summary of our findings. As a result, we consider our current level of sensitivity sufficient for our de-identification purposes. Our PPV of 0.63 is significantly higher than previous versions of the algorithm. Subjective evaluation by medical experts has found de-identified text to be adequately readable.

4.3 Interim Manual Evaluation

4.3.1 Procedure

In January 2006, after 6 months of development, we performed an interim in-house manual evaluation of the de-identification algorithm. Our goal was to identify the PHI that escaped the de-identification algorithm, and thereby to estimate the false negative rate (which indicates the algorithm's sensitivity) and analyze a breakdown of escaped PHI types. We focused on finding false negatives because minimizing the number of escaped PHI is crucial in protecting patient confidentiality. The evaluation was performed on random samples of de-identified nursing notes that were uploaded to the MIMIC II server.

We recruited 11 lab personnel, who were familiar with the organization of the nursing notes, as evaluators. The nursing notes are written in free-form English, but frequently include medical terminology. Our lab personnel were familiar with such medical terminology and proved to be capable of parsing the nursing notes to discover escaped PHI. We declared a flat reward scheme of $1/escaped PHI identified. We noted that each evaluator would likely spend about 1 hour and earn in the range of $10-30. The expected reward thus fell in the range of $10-30/hour. It can be argued that such a low hourly rate may not have been attractive enough to obtain the highest performance of the lab personnel. However, the de-identification project is an important part of the lab's overall activity, and hence we expected our evaluators to do their best possible job.

We assigned approximately 130 nursing notes/22,000 words to each evaluator, which added up to 1,836 notes in total. Additionally, we provided information on the following:

- Definition of a PHI
- PHI categories, and examples, that should be identified
- Format of the nursing notes
- Instructions on how to discover escaped PHI
- Sample nursing note and same note after evaluation

The evaluators were requested to analyze the entire sample of 130 nursing notes at one sitting. They were instructed to read the notes carefully and highlight every word or phrase that they suspected to be PHI. They were requested to repeat the process if they so wished. Another reviewer acted as a central supervisor of the evaluation process and double-checked whether the PHI identified by our evaluators were actually PHI. Hence, we were able to minimize the possibility of the evaluation process adding to the estimated false negative rate of the de-identification algorithm.

For each evaluator's sample, we counted the exact number of nursing notes, words and escaped PHI as identified by the evaluator. Using this information, we determined the following statistics that are indicative of the algorithm's sensitivity:

- False negative rate = # of escaped PHI in sample / # of words in sample
- False negative per nursing note = # of escaped PHI in sample / # of nursing notes in sample

4.3.2 Limitations

Due to time limitation, we were not able to double-check every nursing note to discover additional PHI that escaped both the de-identification algorithm and our human evaluators. We could not have more than one evaluator read each sample for the same reason. As a result, we were not able to minimize the number of false negatives in the evaluation. Therefore, there is a risk that the evaluation process underestimates the false negative rate of the de-identification algorithm.

Performance of the different evaluators was variable. High performers could include evaluators who were very experienced in reading nursing notes, and those that spent significantly longer times reading the notes. As a result, the results of the evaluation had an evaluator-specific variance that we could not calculate due to time limitations.

As previously mentioned, this evaluation does not deal with the false positive rate of the algorithm. False positives in the nursing notes do not jeopardize patient confidentiality, but do have an adverse effect on readability. A manual expert evaluation of false positives would be useful in illuminating the readability issue. This evaluation would involve discovering what the algorithm tagged as PHI and contextually judging whether each is actually PHI. However, in our de-identified nursing notes, PHI identified by the algorithm are simply removed and are not available for perusal. We were thus unable to analyze the false positive rate. With this false positive information absent, we were unable to accurately determine the algorithm's sensitivity, but which we estimated from the false negative rate.

4.3.3 Quantitative Results

Table 4.2 presents the summarized results of the manual evaluation of de-identified nursing notes compared with results of human de-identification. Table 4.3 breaks down the algorithm's false negatives based on their PHI categories.

Table 4.2. False negative rate for algorithm and human de-identification.

	Min	Max	Mean
Algorithm			0.05% or 5 FN/10,000 words
1 person	0.19% or 29 FN/10,000 words	0.03% or 3 FN/10,000 words	0.1% or 10 FN/10,000 words
2 people	0.06% 6 FN/10,000 words	0.01% or 1 FN/10,000 words	0.03% or 3 FN/10,000 words
3 people	0.01% or 1 FN/10,000 words	0.005% or 0.5 FN/10,000 words	0.01% or 1 FN/10,000 words

Table 4.3. Categorization of algorithm false negatives by PHI type.

PHI type	# False negatives	# FN per 100,000 words
Full name	4	1
Last name	14	5
First name	31	11
Location (not street address)	7	2
Hospital/rehab/nursing home name	26	9
Full date	2	1
Partial date	9	3
Year	8	3
Age over 89	3	1

We categorize the false negatives into 3 risk categories to gain a better sense of the risk to patient confidentiality. This classification is presented in Table 4.4.

Table 4.4. Risk classification of algorithm false negatives.

Risk category	PHI type	# False negatives	% of all FN
High	Full names, last names	18	17
Medium	First names, hospital names, locations, full dates	66	63
Low	Partial dates, years, age over 89	20	19

The low risk category consists mainly of partial dates and individual years. These short numerical patterns are often confused with non-PHI terms. In order to reduce false identifications, the algorithm does not identify partial dates, years and ages very aggressively. This is a likely explanation of the false negatives in this risk category. It is to be noted that neither category is specified by HIPAA regulations and pose very little risk to confidentiality.

Most false negatives fall under the medium risk category. These mainly include first names, locations and a few full dates and hospital names. The main problem with these PHI categories are the incompleteness of our known PHI lists. We plan to make relevant de-identification rules more stringent and include more known PHI for direct retrieval.

The highest risk category includes full and last names of providers. Within the set of nursing notes evaluated, not a single patient last or full name escaped de-identification. To remedy the provider name false negatives, we plan to enrich our list of known provider names.

It is noteworthy that nursing notes, on which we evaluated our algorithm, are generally less stylized than discharge summaries. As discussed in Section 3.3, PHI in discharge summaries tend to occur mainly in a few sections, e.g. "Patient Name", "Social History". Our algorithm takes into account the section heading in identifying PHI, e.g.

any words under "Name" is automatically identified as name PHI. Nursing notes, on the other hand, include no such segmentation since all PHI occur in a single body of text, thus making de-identification more difficult. Names in discharge summaries overwhelmingly follow common name patterns, e.g. "Dr Brown", "Jane Doe". In contrast, nursing notes have many informal mentions of patient and doctor names that do not follow these standard formats, e.g. "Jane complained of breathing difficulties at night". Due to these factors, we strongly believe that our algorithm will prove to have a higher performance in terms of sensitivity and PPV on discharge summaries than our current report values (that are based on nursing notes). Further performance evaluation will be undertaken in the lab once a gold standard de-identified corpus of discharge summaries has been compiled.

4.3.4 Qualitative Results

Readability is an important performance variable that determines how informative and usable the de-identified records are in medical research. The smaller the number of false positives, the higher the positive predictive value and the more readable the text. However, readability is a subjective quality of the text, and it is not evident how one can formulate a universally-accepted measure for it. Persons who are medically trained or are familiar with medical reports, as the end users are likely to be, might find it easier to guess information blotted out by false positives than naïve users. Thus they would find de-identified text more readable than lay readers. In addition, readability is likely to be significantly variable throughout each record since certain portions of text, e.g. physiologic information with different numerical formats, have textual patterns that are more prone to false identification. However, we have independent sources for these

physiologic data in MIMIC II and false positives affecting physiologic variables are not likely to hinder research.

In light of these issues, devising an objective measurement of readability will be a valuable avenue for future work. For our current purposes, we obtained subjective feedback from medically trained lab personnel about the readability of our de-identified nursing notes and discharge summaries. The overall assessment was that the text is sufficiently readable for medical research purposes. As mentioned before, a high incidence of false positives occur due to physiologic information that is mistaken to be dates. These physiologic data, e.g. blood gases, are usually available in the other medical files on the MIMIC II database. Thus despite the readability issue, relevant information can often be retrieved from other sources to fill in the gaps left by the de-identification system.

5 Conclusion

5.1 Thesis Summary

To summarize, we have developed and evaluated a comprehensive de-identification system to preserve patient and provider confidentiality in free-text discharge summaries and nursing notes. Medical records are increasingly being used in medical research. These records have to go through a process of thorough de-identification before they can be transferred from hospitals to interested research groups. For our de-identification purposes, we deemed it crucial to fine-tune our system to two major types of free-text medical records in the MIMIC II database: discharge summaries and nursing notes. Currently our algorithm uses pattern matching augmented by lists of known and potential PHI. We conducted comparative evaluation against a de-identification gold standard and also manual evaluation by medical experts. The algorithm achieved a sensitivity of approximately 0.87 - which is significantly higher than one person but lower than a consensus of two people de-identifying - and a positive predictive value (PPV) of 0.63. We deem the sensitivity to be sufficient for our current purposes of information release and use, and the moderately high PPV preserves the readability of the text. The system has been used to de-identify medical records of 17,000 patients included in the MIMIC II database. We expect continued work on the system to improve the sensitivity and PPV statistics, and to develop more extensive methods of algorithm evaluation.

5.2 Current De-identification Status of MIMIC II Database

The purpose of developing our de-identification software was to scrub MIMIC II free-text files of identifying information before posting them for limited research use. With our current system we have de-identified 412,509 nursing notes and 1,934 discharge

summaries on MIMIC II. We anticipate adding thousands more medical records to the database, all of which will undergo the de-identification process.

As presented in previous sections, our system's performance has a sensitivity rating significantly better than one person de-identifying. In any de-identification system, there is a significant likelihood that the software may come across PHI that are absent in the extensive dictionaries of known PHI and that are also not identified by the rules. To reduce the risk of inadvertent PHI exposure, we are currently releasing our de-identified text files only to selected research groups who are required to sign data use agreements.

De-identification systems like ours can have widespread use in information sharing for research purposes. Our system is sufficiently generalized to handle text files of any format, albeit with varying performance, and may be useful in other research groups' research efforts. In the spirit of open-source software, we intend to make the source-code of our system, excluding some MIMIC II-specific PHI lists, online for full public use. As presented in Appendix C, while running the software, the user is asked to specify which PHI categories are to be identified. Thus, it is possible to identify the full range of PHI categories or any subset of it. This makes our system usable by different research groups for slightly varying de-identification goals.

In the version of our software that will be released to the public, we will exclude references to the doctor and patient names extracted from MIMIC II that are specific to our files. Releasing these names would completely thwart the purpose of our system. Additionally, we will exclude the re-identification module of our software in the released version since this module provides a mapping between each PHI in the text file and the fake representative term we replace it with. On the other hand, we will include

information essential to our system's functioning, including dictionaries of common words and SNOMED terminology, lists of first and last names, etc.

De-identification is an ongoing project at HST and the algorithm described in this document is constantly undergoing incremental changes. For more up-to-date documentation of the algorithm, consult the User Manual and the Developers' Guide [26, 27].

5.3 Future Work

5.3.1 Evaluation

Our current software is optimized for performance on MIMIC II nursing notes and discharge summaries. The gold standard we currently use to determine performance statistics includes only nursing notes. As textual medical records become more and more important in medical research, de-identification and evaluation of different types of records will become valuable. It would be interesting to evaluate our current de-identification algorithm on other medical records, e.g. medical test reports. An avenue for future work would be to extend our current gold standard to include these other record types.

Readability is an important factor in de-identified text that is unfortunately difficult to evaluate objectively. Formulation of a universal way to assess readability would be invaluable in evaluating a de-identification algorithm's tradeoffs between false positives and false negatives.

5.3.2 Statistical Training Approach

Our rule-based pattern-matching method de-identifies free text with sufficiently few false negatives, i.e., it identifies most PHI in the text, but with significantly more false

positives. Extraction of non-PHI terms as PHI reduces the readability of the nursing notes and might remove valuable medical information that makes the notes useful in research. A new approach using statistical training could be designed with an aim to increase both sensitivity and positive predictive value.

Like other more-structured text, free-form medical records are grammatical and have a general structure. For example, a first name (which is a noun) is likely to be followed by a last name (which is also a noun). The general syntactic structure of these records could thus be used to statistically determine if a given word is PHI. Such a de-identification system could use a Hidden Markov Model (HMM), and would perform 2 steps: training the HMM based on the syntactic structure of a medical record training set, and using the model to de-identify a separate test set of records.

Bibliography

[1] PhysioNet. Available: http://www.physionet.org

[2] Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring. *Computers in Cardiology*, 29:641–644, 2002.

[3] Standards for Privacy of Individually Identifiable Health Information (45 CFR Parts 160 and 164). Dec 28, 2000; last amended April 17, 2003.

[4] Federal Policy for the Protection of Human Subjects (Federal Register 56 28003). Code of Federal Regulations.

[5] Health insurance portability and accountability act of 1996.
Available: http://privacyruleandresearch.nih.gov/pdf/HIPAA_Booklet_4-14-2003.pdf

[6] Beckwith B, Mahaadevan R, Balis U, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Medical Informatics and Decision Making*, 6(1), 2006.

[7] De-Identification Tool for Patient Records used in Clinical Research. Health Sciences Library System for the University of Pittsburgh *and* UPMC. University of Pittsburgh Medical Center.
Available: http://www.hsls.pitt.edu/about/news/hslsupdate/2004/june/iim_de_id

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

[9] Miller RE, Boitnott JK, Moore GW. Web-based free-text query system for surgical pathology reports with automatic case-identification. *The Johns Hopkins Autopsy Resource.*

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

[11] Gupta D, Saul M, Gilbertson J. Evaluation of a de-identification software engine: Progress towards sharing clinical documents and pathology reports. *Am J Clin Pathol*, 121(2):176-186, 2004.

[12] Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Symp*, 333-337, 1996.

[13] Berman JJ. Concept-Match Medical Data Scrubbing. *Archives of Pathology and Laboratory Medicine*, 127(6): 680-686, 2003.

[14] Berman JJ. Comparing de-identification methods (comments on an article). *BMC Medical Informatics and Decision Making*, 6(1), 2006.

[15] Douglass M. Computer-Assisted De-identification of Free-text Nursing Notes. MIT Press, 77 Mass. Ave., Cambridge, MA, USA, 2005. MEng Thesis.

[16] HIPAA Code Sets. Claredi. Available: http://www.claredi.com/hipaa/codesets.php?PHPSESSID=8fcbc0bf053e7db681d6add673665849

[17] Committee on the use of humans as experimental subjects. Available: http://web.mit.edu/committees/couhes

[18] U.S. Census Bureau. 1990 census name files, 1999. Available: http://www.census.gov/genealogy/names/

[19] U.S. Census Bureau. Census 2000 urbanized area and urban cluster information, 2004. Available: http://www.census.gov/geo/www/ua/uaucinfo.html#lists

[20]Atkinson K. Spell checking oriented word lists. Revision 6, 2004. Available: http://prdownloads.sourceforge.net/wordlist/scowl-6.tar.gz.

[21] Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. *Methods Inf Med*, 32(4):281–291, 1993.

[22] Levine J. De-identification of ICU Patient Records. MIT Press, 77 Mass. Ave., Cambridge, MA, USA, 2003. MEng Thesis.

[23] Douglass M, Clifford GD, Reisner A, Long WJ, Moody GB, Mark RG. De-Identification Algorithm for Free-Text Nursing Notes. *Computers In Cardiology*, S6.2, 2005.

[24] Douglass M, Clifford GD, Reisner A, Moody GB, Mark RG. Computer-Assisted Deidentification of Free Text in the MIMIC II Database. *Computers In Cardiology*, M6.2, 2004.

[25] Statistical Policy Working Paper 22 - Report on Statistical Disclosure Limitation Methodology. Federal Committee on Statistical Methodology. Available: http://www.fcsm.gov/working-papers/wp22.html

[26] Neamatullah I. De-Identification of Free-Text Medical Records: User Manual. Harvard-MIT Division of Health Sciences and Technology. Sept 5, 2006. Available: http://mimic.mit.edu/wiki/uploads/DeidUserManual.doc

[27] Neamatullah I. De-Identification of Free-Text Medical Records: Developers' Guide. Harvard-MIT Division of Health Sciences and Technology. Sept 5, 2006. Available: http://mimic.mit.edu/wiki/uploads/DeidDevelopersGuide.doc

Appendix A Sample De-Identified Discharge Summary

"Name: [**Name 1**], [**First Name 2**] [**Last Name 3**] [**Unit Number 4**]

Admission Date: [**2012-08-12**] Discharge Date: [**2012-09-10**]

Date of Birth: [**1952-05-24**] Sex: F

Service: BLUM-MEDIC

HISTORY OF PRESENT ILLNESS: This patient was transferred from the Cardiac Care Unit on [**2012-08-01**]. For a description of the events taken place in the Cardiac Care Unit, please see Dr. [**Last Name 5**]'s discharge note for the Cardiac Care Unit.

Briefly, the patient is a 68 year old female with a history of hypertrophic cardiomyopathy diagnosed by echocardiogram in the year [**2010**], who broke her left hip on [**2012-08-11**], and was admitted to the Orthopedic Service for open reduction and internal fixation. However, prior to sickle cell, the patient experienced shortness of breath secondary to congestive heart failure and was transferred to the [**Hospital 6**] where she was aggressively diuresed with
Lasix. She experienced worsening respiratory distress. Cardiology advised a blood transfusion and aggressive administration of negative inotropics (Lopressor 100 four times a day and Verapamil 80 three times a day currently), and close monitoring of intervascular volume (goal euvolemic) in order to minimize the patient's outflow tract gradient. She was subsequently transferred to the Cardiac Care Unit on [**2012-08-15**], where she was intubated for respiratory decompensation and underwent open reduction and internal fixation of the left hip on [**2012-08-16**], which required 11 units of packed red blood cells, 7 units of fresh frozen plasma, and 2 units of plasma.

During her stay in the Cardiac Care Unit, the patient's main issues have included the following: 1) Aggressive use of negative inotropics and close maintenance of euvolemic to minimize outflow tract gradient; 2) prolonged ventilation with extubation [**2012-08-29**]; 3) brief oliguria secondary to ATN following surgery, now resolved; 4) prolonged high-grade fevers to 103.0 F./104.0 F., thought to be secondary to Zyprexa induced NMS, then C. difficile infection; 5)

persistent obtunded state secondary to toxic metabolic encephalopathy.

PAST MEDICAL HISTORY:
1. Hypertrophic obstructive cardiomyopathy diagnosed by echocardiogram in [**2010**].
2. Hypertension.
3. Schizo-affective personality disorder.
4. Depression.
5. Anxiety disorder.
6. Breast cancer.

MEDICATIONS PRIOR TO TRANSFER:
1. Lovenox 30 mg subcutaneously twice a day.
2. Protonix 40 mg intravenously q. day.
3. Calcium carbonate 1000 mg four times a day.
4. Miconazole.
5. Nystatin swish and swallow.
6. Metoprolol 100 p.o. four times a day.
7. Verapamil 80 mg p.o. three times a day.
8. Fentanyl patch 25 micrograms.
9. Tylenol 500 to 1000 mg p.r.n.
10. Bumetanide 1 gram intravenously q. day.
11. Risperidone 1 mg p.o. p.r.n.
12. Ativan 0.5 mg three times a day.

ALLERGIES: No known drug allergies.

PHYSICAL EXAMINATION: Upon arrival, the patient was afebrile with a temperature of 97.2 F.; pulse of 88; blood pressure of 128/72; tachypneic at a rate of 36, saturation of 98% by 40% face mask. In general, this is an obtunded female laying in bed, with very shallow rapid breathing. Her pupils were equal, round and reactive to light. Her eyes were anicteric. Oropharynx is benign. Anterior lung fields were clear bilaterally. Her heart was regular rate and rhythm. III/VI harsh systolic murmur best heard at the apex. Her abdomen was soft, obese, nontender, nondistended, with positive bowel sounds with no evidence of hepatosplenomegaly. Extremities were cool and dry with one plus distal pulses bilaterally; no edema. Foley catheter and rectal bag were in place. Left inguinal incision was clean, dry and intact. She intermittently answered the questions, acknowledged the presence of visitors. She would track movements in the room but otherwise was very unresponsive.

LABORATORY: She had a white blood cell count of 8.6, hematocrit of 28.4, platelet count of 439. Coags were a PT

of 13.5, PTT of 25.6 and an INR of 1.3. Chem-10 included the sodium of 142, potassium 3.5, chloride 103, bicarbonate 31, BUN of 21, creatinine 0.6, glucose of 127. Calcium 9.1, phosphorus 3.3, magnesium 2.2.

STUDIES: Echocardiogram on [**2012-08-21**] showed elongated left atrium, moderate symmetric left ventricular hypertrophy, severe resting left ventricular outflow tract, normal right ventricular size and wall motion, three plus mitral regurgitation (increased compared to [**2012-08-13**] study), ejection fraction of 70%.

She was Clostridium difficile positive on [**2012-08-23**].

Chest x-ray on [**2012-08-28**] which showed a new left lower lobe consolidation consistent with either pneumonia or atelectasis.

LENI's performed on [**2012-08-27**] were negative for deep venous thrombosis.

CT scan of the abdomen / pelvis on [**2012-08-23**], which showed post-traumatic changes around the left pelvic fracture. There were no signs of abscess. Overall appearance is of stranding and fluid amongst the left hip muscles, consistent with resolving hematoma. Right lower lobe pneumonia versus contusion bilaterally with small pleural effusion. Renal scarring.

SUMMARY OF HOSPITAL COURSE:
1. Cardiac: The patient continued to be managed aggressively with high dose Lopressor and Verapamil. There was careful attention paid to strict maintenance of intakes an outputs. She continued to receive her Bumex 1 gram every morning. The overall strategy for her cardiac status was, as noted in the Cardiac Care Unit note, to administer aggressive negative inotropics with optimization of patient's intervascular volume. Again, the patient has no history of coronary artery disease. She did not have any episodes of arrhythmia during her hospital course. She did, however, continue to have a low baseline troponin leak. Her troponin was always greater than 1.6 every time it was measured. It was never higher than 3.5.

It was believed that this mild troponin leak was secondary to constant strain given the patient's hyperdynamic state.

2. Pulmonary: The patient was initially transferred to the

Floor on 50% face mask. She was weaned to room air successfully on [**2012-09-04**]. She had required Bi-PAP in the Cardiac Care Unit, but it was never required on the floor. She did, however, have one episode of tachypnea early on in her course on the floor on [**2012-09-01**], with respiratory rates in the mid-40s. She continued to have O2 saturations in the high 90s during this time with normal blood pressure and slightly tachycardic. This episode self resolved. She had subtle EKG changes, specifically ST abnormalities which were present on her EKG before. During this setting of tachypnea, her troponin rose from her baseline one to two range to a maximum of 3.5. Chest x-ray taken at this time showed persistent mild congestive heart failure.

3. Neurologic: Upon arrival to the floor, the patient was initially weaned off her Ativan and her Fentanyl patch was discontinued. Initially, the patient's neurologic status consisted of a very sparse verbal output, not really following basic commands, very limited attention span, and persistent cogwheel rigidity. It was believed on the part of Neurology and Psychiatry consultations, that her obtundation was secondary to toxic metabolic encephalopathy. She had an EEG performed that showed global slowing.

A head CT scan did not show any focality. Over the course in the floor, she did however, continue to improve in her neurological status to the point where she was able to recognize individuals walking into her room. Her verbal output increased. She was able to answer in full sentences. She had a longer attention span. She was able to move her upper and lower extremities more spontaneously. Over the last couple of days, she has been complaining of some depression. Her psychiatric medications were stopped while she was in the Cardiac Care Unit. Those medications should be restarted at her rehabilitation center once her neurologic status stabilized.

4. Gastrointestinal: The patient was initially continued on tube feeds by nasogastric tube upon transfer to the floor. She was transitioned to p.o. feeds on [**2012-09-06**] after passing a bedside speech and swallow evaluation. Her present diet consists of clear liquids and pureed foods with Boost for supplementation. She should be continued on this diet, advanced slowly, and observed for any signs of aspiration. She is also receiving Colace and Sennas laxatives.

5. Renal: The patient had no acute renal issues while on the Floor. She had good urine output with stable BUN and

creatinine levels. She did, however, have a continuation of a metabolic alkalosis and respiratory alkalosis after immediate transfer from the Cardiac Care Unit but these acid based disorders resolved over the course of her stay on the floor.

6. Hematologic: The patient's hematocrit was stable around 30. Her goal hematocrit was 30 and above to avoid exacerbation of her cardiac hyperdynamic state.

7. Infectious Disease: The patient experienced high grade fevers to 103.0 F., 104.0 F., in the Cardiac Care Unit and she spiked once to a temperature of 101.0 F., while on the floor. She was pan-cultured and the cultures were all negative. She completed a 10 day course of Flagyl for her C. difficile infection. She also completed a four day course of Vancomycin for a slightly purulent left inguinal wound. The wound was also debrided and is appearing clean, dry and intact currently.

8. Musculoskeletal: The patient is status post open reduction and internal fixation. She has been receiving Physical Therapy approximately three times a week for range of motion exercises, sitting in chair and increasing her weight bearing status. Her Physical Therapy should be continued with the Rehabilitation Center. She is also receiving Lovenox subcutaneously twice a day and her anti-coagulation should be continued for another two and a half weeks for a total of six weeks.

9. Prophylaxis: The patient was given Protonix for peptic ulcer disease and Pneumoboots with subcutaneous Lovenox for deep venous thrombosis prophylaxis.

CONDITION AT DISCHARGE: Stable.

DISCHARGE STATUS: The patient is being discharged to [**Hospital 7**] for rehabilitation on [**2012-09-10**].

DISCHARGE INSTRUCTIONS:
1. Her diet currently is clear liquids with pureed foots and Boost for supplementation.
2. She should continue to receive Physical Therapy for rehabilitation of her left hip.

DISCHARGE MEDICATIONS:
1. Enoxaparin 30 mg subcutaneously q. 12 times six weeks total (started on [**2012-08-17**]).

2. Pantoprazole 40 mg intravenously q. 24 hours.
3. Calcium carbonate 1000 mg p.o. four times a day.
4. Metoprolol 100 mg p.o. q. six hours.
5. Verapamil 80 mg p.o. q. eight hours.
6. Bumetanide 1 mg intravenously q. day.
7. Albuterol nebulizer solution, one nebulizer inhaler q. two hours p.r.n.
8. Ipratropium bromide nebulizer, one nebulizer inhaler q. six hours p.r.n.
9. Aspirin 325 mg p.o. q. day.
10. Colace 100 mg p.o. twice a day.
11. Senna 2 tablets p.o. q. h.s. p.r.n.
12. Miconazole powder 2%, one application topical four times a day p.r.n.
13. Nystatin oral suspension 5 ml four times a day p.r.n.

The patient should be restarted on her psychiatric medications when her neurologic status stabilizes. Her psychiatric medications upon admission included:

14. Celexa 20 mg q. day.
15. Paxil 20 mg q. day.
16. Zyprexa 5 mg twice a day plus 20 mg q. h.s.

DISCHARGE DIAGNOSES:
1. Hypertrophic obstructive cardiomyopathy.
2. Congestive heart failure.
3. Status post left open reduction and internal fixation of hip.
4. Schizo-affective personality disorder.
5. Depression.
6. General anxiety disorder.
7. Status post breast cancer.

[**First Name 8**] [**Initial **]. [**Last Name 9**], M.D. 12-207

Dictated By: [**First Name 10**] [**Last Name 11**], M.D.

MEDQUIST36

D: [**2012-09-10**] 11:23
T: [**2012-09-10**] 11:27
JOB#: [**Zip Code 12**]
(more)

Appendix B Sample De-Identified Nursing Note

10 | 2001/08/07 16:34:00 | MICU Nursing Progress Note 7a-7p:

Neuro: Alert and oriented x3. Ativan 1mg po prn for expressed anxiety. Moving all extremities w/o difficulty. OOB to BSC and to chair with one assist tolerating increased activity well.

CV: HR 90-105 ST. No ectopy. K+ 3.9. Denies cardiac compliants. SBP 149-159. SBP 171 with increased activity to chair. Cont to encourage pt to drink secondary tachycardia.

PULM: Trach with ventilatory support on SIMV (home settings). Sats 99-100%. RR 14-22 on vent. Sxn'd for whitish thick secretions. Conts on Resp treatments q4hr. Pt placed on passe muir valve and humidified oxygen. Speech clear. Trach site c/d. Pt tolerating wean well. RR 20-25. Sats 99%. Pt denies SOB. Pt appears to be breathing comfortably. Abx changed per sensitivities to ceftaz and pipercillin iv. Levofloxacin po d/c'd.

GI: Abd soft NT +BS. Excellant appetite. Large soft BM on BSC.

GU: Voiding spontaneously via urinal cyu. -fluid status.

SKIN: R elbow with stage 2 breakdown, team aware. Cleansed with NS and DSD appiled.

ID: afebrile

PROPH: Protonix po and hep SC.

DISPO: Full Code

A: Improving resp status.
tolerating time of vent.
Conts with tachycardia

P: Cont to increase time off vent as tolerated.

Pulm toileting
Ativan po prn.
Provide support.
Await discussion with team regarding dc plans to home.

Appendix C Command-Prompt Interaction with User

```
**************************************************************************
De-Identification Algorithm: Identifies Protected Health Information (PHI) in Discharge
Summaries and Nursing Notes
**************************************************************************

Comparison with Gold Standard: Press '0' for no comparison or '1' for statistical
comparison with existing Gold Standard, and then hit return.
If unsure, press '0'.
1

Date shifting: Press '0' to preserve all dates.
To shift dates, enter the amount of forward shift in weeks.
200

**************************************************************************
PHI categories filtered:
1. Social Security Numbers (SSN)
2. Uniform Resource Locators (URL)
3. Email addresses
4. Telephone/fax numbers
5. Provider/unit/medical record numbers
6. Ages over 90
7. Locations and hospital names
8. Dates
9. Years
10. Names
**************************************************************************

De-identify all PHI categories?
Please press 'y' or 'n', and then hit return.
If unsure, press 'y'.
n

1. De-identify Social Security Numbers (SSN)?
Please press 'y' or 'n', and then hit return: y

2. De-identify Uniform Resource Locators (URL)?
Please press 'y' or 'n', and then hit return: y

3. De-identify email addresses?
Please press 'y' or 'n', and then hit return: y

4. De-identify telephone/fax numbers?
Please press 'y' or 'n', and then hit return: y
```

5. De-identify provider/unit/medical record numbers?
Please press 'y' or 'n', and then hit return: *y*

6. De-identify ages over 90?
Please press 'y' or 'n', and then hit return: *y*

7. De-identify locations and hospital names?
Please press 'y' or 'n', and then hit return: *y*

8. De-identify dates?
Please press 'y' or 'n', and then hit return: *n*

8. De-identify individual years?
Please press 'y' or 'n', and then hit return: *y*

9. De-identify names?
Please press 'y' or 'n', and then hit return: *y*

Starting de-identification...