A FRAMEWORK FOR PRIVACY-PRESERVING MEDICAL DOCUMENT SHARING

Completed Research Paper

Xiao-Bai Li, Jialun Qin
Department of Operations and Information Systems
Manning School of Business
University of Massachusetts Lowell
Lowell, MA 01854, U.S.A.
xiaobai_li@uml.edu; jialun_qin@uml.edu

Abstract

Health information systems have greatly increased availability of medical documents and benefited healthcare management and research. However, there are growing concerns about privacy in sharing medical documents. Existing approaches for privacy-preserving data sharing deal mostly with structured data. Current privacy techniques for unstructured medical text focus on detection and removal of patient identifiers from the text, which may be inadequate for preserving privacy and data utility. We propose a novel framework to extract, cluster, de-identify and anonymize patient medical documents. Our framework integrates the approaches developed in both data privacy and health informatics fields. The key novel elements of this framework include (i) a meta-learning approach to extract personal and health information from documents; (ii) a recursive partitioning method to cluster patient documents by medical concept; and (iii) a cluster-level value-enumeration method for anonymization. A prototype system has been implemented and evaluated to demonstrate the effectiveness of our proposed framework.

Keywords: Data privacy, electronic medical records, healthcare information systems, information extraction, document clustering, anonymization
Introduction

The advance in the Electronic Medical Record (EMR) technology has enabled organizations to store and share a large amount of personal health data. While secondary use of this data has significantly enhanced the quality and efficiency of healthcare management and medical research, there are growing concerns about patient privacy due to the widespread use of EMR systems. To respond to these concerns, the Health Insurance Portability and Accountability Act (HIPAA) established a set of privacy rules. The HIPAA Safe Harbor rule specifies 18 categories of explicitly or potentially identifying attributes, called Protected Health Information (PHI), that must be removed before the health data is released to a third party (DHHS 2000, 2002). A strict implementation of the HIPAA Safe Harbor rule, however, may be inadequate for protecting privacy or preserving data quality. Recognizing this limitation, HIPAA also provides guidelines that enable a scientific assessment of privacy disclosure risk in order to determine if the data is appropriate for release.

Along the line of the scientific approach, there is a large body of research on privacy-preserving techniques for data sharing and data mining, most of which deal with structured data only (Aggarwal and Yu 2008). Principles such as k-anonymity (Sweeney 2002), l-diversity (Machanavajjhala et al. 2006), and t-closeness (Li et al. 2007) have been proposed to formalize privacy protection requirements. Various methods and algorithms have been developed to anonymize structured data to satisfy the aforementioned privacy principles.

In spite of this abundance of research in data privacy, its application in medical and healthcare domains lags behind in some aspects. Medical data typically contain unstructured or semi-structured components, such as clinical narratives, pathology reports, and discharge summaries. In such cases, identity information is embedded in the textual contents, where anonymization techniques designed for structured data are not readily applicable. Thus, the majority of privacy research in sharing and releasing medical documents has focused on the automatic detection of PHI attributes in the documents (Meystre et al. 2010; Uzuner et al. 2007). The identified PHI values are then simply removed from or replaced with a constant value in the released text. Studies have shown that such simple de-identification strategies lack the flexibility to adequately meet the diverse needs of data users; they can be under-protective (i.e., not satisfying privacy requirements) in some cases and over-protective (i.e., resulting in poor data utility) in others (Meystre et al. 2010; Uzuner et al. 2007). There is a lack of research on how to mask identified PHI (other than simple removal) and how to cope with non-PHI but potentially identifying information to improve privacy protection and data utility.

In this study, we examine privacy protection issues related to sharing patient medical documents. We propose a novel framework to extract, cluster, de-identify and anonymize the medical documents. The proposed framework consists of three major components: (i) information extraction, which extracts PHI and non-PHI attributes that could reveal patients’ identities, as well as health and medical information, from the text; (ii) document clustering, which clusters patient documents based on the similarity of the health and medical information; and (iii) de-identification and anonymization, which remove explicit individual identifiers (called de-identification) and anonymize potentially identifying attributes using a cluster-level value-enumeration method (called anonymization). The proposed framework is implemented into a prototype system called DAST (De-identification and Anonymization for Sharing medical Texts). The main contributions of this research are summarized below.

- We investigate a problem that has not been formally studied in the literature: the under-/over-protective problem associated with current practice of removing the PHI attributes from medical documents. Existing privacy-preserving techniques for releasing structured data are not directly applicable to the problem.
- We propose a meta-learning approach to extract personal information from medical documents. This approach congregates the extraction results from multiple independent machine learning methods to improve extraction performance. Unlike traditional information extraction approaches that aim more at improving extraction precision, our approach focuses more on improving recall because failure to detect PHI fields would result in HIPAA violation.
- The ideas of clustering patient documents by health and medical concepts and then using the cluster-
level enumeration for anonymization are innovative. The proposed document clustering method is novel. And the value-Enumeration method for anonymization is new to the literature.

- We develop a prototype system that integrates three sophisticated modules: information extraction, document clustering, and de-identification and anonymization. We tested and evaluated the prototype system on real-life medical text documents.

The remainder of the paper is organized as follows. In the next section, we briefly review related work in data privacy and health informatics. We then demonstrate the problems with the current approaches using an illustrative example. Next, we present the analysis and design of the proposed privacy-preserving document sharing framework from both privacy protection and data utility perspectives. A prototype system developed based on the framework is then described. We then report the results of the evaluation study based on real-life data. In the last section, we summarize our work and discuss future research.

**Related Work**

There are two types of privacy disclosure widely recognized in the literature (Duncan and Lambert 1989): (a) *identity disclosure* (or *re-identification*), which occurs when a data intruder is able to match a record in a dataset to an individual; and (b) *attribute disclosure*, which occurs when an intruder is able to predict the sensitive value(s) of an individual record, with or without knowing the identity of the individual. Related to the two types of disclosures, the attributes of data on individuals can be classified into three types. We discuss them along with the HIPAA-defined PHI categories, listed in Table 1 (DHHS 2000, pp. 82818-82819).

The first type is *explicit identifier* (EID), such as name, phone number and social security number, which can be used to directly identify an individual. It is clear from Table 1 that all PHI categories are EIDs except category 2 (locations) and category 3 (dates). The second type is *quasi-identifier* (QID), such as age, gender, race and zip code, which do not explicitly reveal identities but may be linked to external data sources to eventually identify an individual. Sweeney (2002) found out that 87% of the population in the United States can be uniquely identified with three QID attributes – gender, date of birth, and 5-digit zip code – which are accessible from voter registration records available to the public. In the PHI list in Table 1, category 2 (locations) and category 3 (dates) are QIDs. Note that a QID may be an attribute that is not PHI (e.g., gender or race). The third type is *sensitive attributes* (SAs), which contain private information that an individual typically does not want disclosed, such as sensitive diseases, sexual orientation, and personal financial information. None of the items listed in Table 1 is an SA. That is, HIPAA does not provide guidelines on how to protect SA information; instead, the basic idea of HIPAA Safe Harbor rule is to protect privacy by preventing identity disclosure.

**Table 1. Protected Health Information (PHI) Defined by HIPAA**

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Names</td>
<td>Names</td>
</tr>
<tr>
<td>2. Locations</td>
<td>All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial 3 digits of a zip code if the correspond area contains more than 20,000 people.</td>
</tr>
<tr>
<td>3. Dates</td>
<td>(i) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death. (ii) All ages over 89 and all elements of dates (including year) indicating such an age.</td>
</tr>
<tr>
<td>4. Telephone numbers</td>
<td></td>
</tr>
<tr>
<td>5. Fax numbers</td>
<td></td>
</tr>
<tr>
<td>6. E-mail addresses</td>
<td></td>
</tr>
<tr>
<td>7. Social security numbers</td>
<td></td>
</tr>
<tr>
<td>8. Medical record numbers</td>
<td></td>
</tr>
<tr>
<td>9. Health plan beneficiary numbers</td>
<td></td>
</tr>
<tr>
<td>10. Account numbers</td>
<td></td>
</tr>
<tr>
<td>11. Certificate/license numbers</td>
<td></td>
</tr>
<tr>
<td>12. Vehicle identifiers and serial numbers, including license plate numbers</td>
<td></td>
</tr>
<tr>
<td>13. Device identifiers and serial numbers</td>
<td></td>
</tr>
<tr>
<td>14. Web Universal Resource Locators (URLs)</td>
<td></td>
</tr>
<tr>
<td>15. Internet Protocol (IP) address numbers</td>
<td></td>
</tr>
<tr>
<td>16. Biometric identifiers, including finger and voice prints</td>
<td></td>
</tr>
<tr>
<td>17. Full face photographic images and any comparable images</td>
<td></td>
</tr>
<tr>
<td>18. Any other unique identifying number, characteristic, or code</td>
<td></td>
</tr>
</tbody>
</table>
Data Privacy in Structured Data

Data privacy has been an active area of research for decades. Most data-privacy studies assume that data is stored in well-defined relational databases. A major line of privacy research has focused on devising principles to establish the requirements of privacy protection and to form criteria for assessing privacy risks. One of the most popular principles is \( k \)-anonymity (Sweeney 2002), which requires that each individual record in a dataset should be indistinguishable from at least \( k - 1 \) other records with respect to the QID attribute values. These indistinguishable records are usually considered to form a group and the parameter \( k \) is referred to as the minimum group size. The \( k \)-anonymity approach focuses on re-identification risk only and does not consider attribute-disclosure risk. It generalizes different but similar QID values into a higher-level value within a group. Since SA values remain unchanged in \( k \)-anonymity, individuals in a group, who have the same generalized QI values, are subject to high attribute-disclosure risk if their SA values are the same or close. To address this issue, a privacy principle called \( l \)-diversity has been proposed (Machanavajjhala et al. 2006). The \( l \)-diversity principle requires that an SA should include at least \( l \) well-represented values in the \( k \)-anonymized data. Another privacy principle, called \( t \)-closeness (Li et al. 2007), addresses the issue by further considering the overall distribution of the SAs. It requires that, for each group, the distance between the distributions of the SAs in the group and the overall distribution cannot be larger than a threshold value \( t \).

Various anonymization methods have been proposed in the literature, including generalization, suppression, swapping, and noise perturbation. Generalization and suppression either generalizes the original values to a higher level category or removes the values if generalization is inappropriate or impossible (Sweeney 2002). Data swapping involves exchange of attribute values between different records (Dalenius and Reiss 1982; Li and Sarkar 2011). Generalization, suppression and swapping apply to both categorical and numeric data. Noise perturbation adds noise to the original data to disguise their true values (Agrawal and Srikant 2000; Li and Sarkar 2013), which applies mainly to numeric data. All these methods attempt to preserve the utility of the anonymized data, as measured by various data quality metrics.

De-Identification in Medical Document Sharing

Unlike privacy research in the structured data, where numerous techniques have been proposed and developed, privacy protection approaches for sharing medical documents in health information systems have mainly focused on the detection and removal of PHI items from the documents (Meystre et al. 2010; Uzuner et al. 2007). The task of detecting PHI items from text can be viewed as a classification problem, where the terms in a medical document are classified into various PHI categories (e.g., name, date, location) or determined to be non-PHI terms. Often, some information extraction methods in the field of Natural Language Processing are used to perform this task. These methods can be largely divided into two categories: pattern matching and machine learning.

Pattern matching methods check each term in a document against a list of manually crafted rules and predefined PHI dictionaries to determine the category of the term. The dictionaries typically include PHI-specific lists such as person name list, city list and hospital list, for matching PHI items. They may also include general term lists and medical term lists for matching non-PHI elements. A major advantage of pattern matching methods is that they require little or no annotated training data. However, these methods often require significant work for developers to craft rules and algorithms to account for different PHI categories, and these rules and algorithms require customization to different datasets. In terms of performance, manually crafted rules often miss unexpected or non-standard variations of PHI terms, resulting in low recall (i.e., proportion of relevant items that are retrieved). Therefore, recent studies tend to adopt supervised machine learning methods to categorize PHI terms (Wellner et al. 2007). Popular machine learning methods used for this task include, for example, Support Vector Machines (SVM) (Cortes and Vapnik 1995) and Conditional Random Fields (CRF) (Lafferty and Pereira 2001). Studies have shown that in general machine-learning-based techniques perform better than pattern-matching-based techniques. Many machine-learning-based techniques also incorporate some rules and dictionaries to further improve the performance (Meystre et al. 2010; Uzuner et al. 2007). A disadvantage of machine-learning-based approach is that it requires fairly large corpuses of annotated documents as training data.
The effectiveness of machine-learning-based detection techniques can be further improved using meta-learning, which is a technique where the results from multiple base learning algorithms are combined (through, for example, a weighted voting mechanism) to generate the final results. It has been shown that this approach can help improve machine learning performance on classification problems (Dzeroski and Zenko 2004; Wolpert 1992). Prior research (Uzuner et al. 2007) has shown that performances of different PHI detection techniques vary in different scenarios. Therefore, combining multiple detection algorithms through meta-learning can potentially help improve the overall performance in PHI detection.

Gardner and Xiong (2009) propose an approach that applies the \( k \)-anonymity framework to anonymize PHI values. To our knowledge, this is the only study in the literature that uses a method other than simple removal of PHI attributes for preventing identity disclosure in medical document sharing. Their method groups the documents based on QID attributes such as age, gender, and location, and anonymizes the QID values by generalization. This way of grouping data may not serve well the purposes of sharing medical documents, where data users are typically more interested in patients’ health and medical information such as symptoms, conditions, diagnosis, and treatments. Therefore, it is desirable to have documents grouped based on the health and medical information. In addition, their work does not examine the impact of \( k \)-anonymity on the utility of the anonymized data. Our study attempts to address these issues.

In summary, there is a lack of interaction between the study in de-identification for medical documents and that in anonymization for structured data. To integrate these two research streams, existing attribute detection techniques need to be extended beyond the HIPAA-defined PHI fields. On the other hand, anonymization techniques designed for structured data need to be reinvented to take advantage of the rich semantic information embedded in the textual contents.

An Illustrative Example

To illustrate the problem in more detail and provide an idea about our proposed approach, consider an example of five simplified clinical notes shown in Figure 1.

<table>
<thead>
<tr>
<th>Original Text</th>
<th>De-identified Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Admitted on 4/25/2009, Tufts Med Ctr. The 88 year old man is complaining fever, sore throat, headache, runny nose.</td>
<td>1. Admitted on [2009], [HOSPITAL]. The 88 year old man is complaining fever, sore throat, headache, runny nose.</td>
</tr>
<tr>
<td>3. Admitted on 4-5-2009, patient is a 5 year old female. Having runny nose, headache, vomiting. ... Mass General Hosp.</td>
<td>3. Admitted on [2009], patient is a 5 year old female. Having runny nose, headache, vomiting. ... [HOSPITAL].</td>
</tr>
<tr>
<td>4. Mark is a 17 year old male. Having pain on right side of abdomen, fatigue, dark urine. ... Admitted 8/20/2010, USC Univ Hosp.</td>
<td>4. [NAME] is a 17 year old male. Having pain on right side of abdomen, fatigue, dark urine. ... Admitted [2010], [HOSPITAL].</td>
</tr>
<tr>
<td>5. Visited on 7/19/2010. Female, 44 year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. ... UCLA Med Ctr.</td>
<td>5. Visited on [2010]. Female, 44 year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. ... [HOSPITAL].</td>
</tr>
</tbody>
</table>

Figure 1. Example Clinical Text – Original and De-identified

The left panel of Figure 1 shows the original text, which includes six categories of information: (a) patient name (in some cases), (b) admission or visit date, (c) hospital name, (d) patient age, (e) patient gender, and (f) symptoms. The first three categories are PHI. In compliance with HIPAA requirements, patient and hospital names must be removed, and date expression must be truncated to include the year only (a complete clinical text report would contain other PHI items such as phone number and date of birth, which would be de-identified similarly). The last three categories are not considered PHI by HIPAA and can be released without any change (other non-PHI data in clinical notes would include test results, diagnosis, treatments, etc.). The resulting de-identified notes are shown on the right panel of Figure 1.
Although the de-identified documents are HIPAA-compliant, they can be over-protective or under-protective. For example, truncating the full-date value to year removes important “season” information in the first three records, which could be crucial for detecting a pandemic outbreak. On the other hand, it may not be difficult to identify the 88-year-old man in record 1 using the other information (e.g., race, marital status and geographical area) typically released together with the clinical notes. Thus, existing de-identification approaches could put patient’s privacy at risk and at the same time provide data that is less useful for research.

| 3. | Admitted on (Mar-2009, Apr-2009), patient is a {5, 52, 88} year old female. Having runny nose, headache, vomiting. … {Mass General Hosp, Tufts Med Ctr}. |
| 4. | [NAME] is a {17, 44} year old male. Having pain on right side of abdomen, fatigue, dark urine. … Admitted {Jul-2010, Aug-2010}, {UCLA Med Ctr, USC Univ Hosp}. |
| 5. | Visited on {Jul-2010, Aug-2010}. Female, {17, 44} year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. … {UCLA Med Ctr, USC Univ Hosp}. |

**Figure 2. Example Clinical Text Anonymized with Proposed Approach (Scenario 1)**

To overcome these limitations, the proposed approach first clusters the patient documents based on medical concepts, and then anonymizes the potentially identifying attributes using a cluster-level value-enumeration method. Figure 2 illustrates how the proposed framework may work for the example shown in Figure 1. With the proposed approach, the five records are clustered into two groups based on the patient’s symptoms. The first group includes record 1, 2 and 3, which show flu-like symptoms. The second group includes record 4 and 5, characterized by Hepatitis-A symptoms. After clustering, a full-date value is replaced by a list of month-year values associated with the group. Similarly, individual age and hospital values are anonymized by a list of values. This method of anonymizing clinical documents provides more useful information for public health surveillance and healthcare/medical research without compromising patient privacy. The cluster-level values are listed in natural alphabetic or numeric order, which essentially masks the original values.

**The Proposed Framework**

In the proposed framework, we classify the information in medical documents into three categories of attributes and process them differently: (a) **Explicit identifiers** (EID), which will be removed or replaced with a constant in the released documents. (b) **Quasi-identifier** (QID), which includes, in the context of medical documents, some PHI attributes such as date of birth, admission/discharge date, hospital, and zip code; also included are some non-PHI attributes such as age, gender, race and marital status. To prevent re-identification, we apply the cluster-level value-enumeration method to anonymize QID values. (c) **Health and medical details** (HMDs), such as symptoms, test results, disease, medications and procedures. We will follow the common practice to keep the HMDs unchanged because they are critical for clinical analysis and healthcare research. In our framework, HMDs will be the basis for clustering medical documents. The proposed framework consists of three functional modules as shown in Figure 3.

**Information Extraction**

Module 1 reads the original documents, detects and extracts three types of data attributes from the textual contents. This module extends the scope of existing medical de-identification systems by extracting not only the PHI categories defined by HIPAA but also additional QIDs (mainly patient demographic information such as age, marital status, ethnicity) not covered by HIPAA, as well as HMDs such as patient symptoms, diagnosis, and treatments. The additional QIDs are later analyzed in terms of their privacy risks and are subject to anonymization in module 3. The HMDs are extracted as input for module 2 to perform document clustering.
The effectiveness of this module is critical since the performance of the other modules depends on it. For PHI detection, we propose using a meta-learning approach which combines multiple PHI detection and extraction algorithms to ensure superior performance. Figure 4 shows the design of Module 1, which contains three components: (1.1) Feature Extractor, (1.2) Base Classifiers, and (1.3) Result Wrapper.

The Feature Extractor (Component 1.1) breaks medical documents into terms and extracts three categories of features for each term: local features regarding the term itself (e.g., term length, part-of-speech, etc.); global features regarding the term's position in the document (e.g., header, body text, heading, etc.); and external features regarding term information gained from external resources (e.g., belonging to a proper noun list, belonging to a medical concept lexicon, etc.). Component 1.2 consists of a set of independent term classifiers called Base Classifiers (e.g., SVM-based classifier, CRF-based classifier, rule-based classifier, etc.), which can be taken directly from existing tools (Wellner et al. 2007; Savova et al. 2010). These base classifiers classify the terms extracted by Component 1.1 into one of the four categories with proper labels: EID, QID, HMD, or Irrelevant. The results of base classifiers are then fed into Component 1.3 to produce final combined results. For example, in the combined result for record 2 in Figure 1, “Mrs. Brown” will be classified as an EID and assigned a tag “PATIENT” to indicate it is a patient name. Similarly, “52 year old” and “female” will be recognized as QIDs and are tagged with “AGE” and “GENDER” respectively. Words such as “joint pain,” “sore throat,” and “fever” will be classified as HMDs.

As mentioned earlier, recall is of particular importance for PHI extraction. The Result Wrapper (Component 1.3) uses a “union” strategy to combine result sets from the base classifiers to improve the performance in recall. For example, if result set 1 generated by base classifier 1 is {Mrs. Brown: PATIENT; Mrs. White: PATIENT} and result set 2 generated by base classifier 2 is {Mrs. Black: PATIENT; Mrs. White: PATIENT}, then the result wrapper will generate a result set that is the union of the two base result sets {Mrs. Brown: PATIENT; Mrs. White: PATIENT; Mrs. Black: PATIENT}. In this way, the result wrapper will find every term that is identified as PATIENT by any base classifier, resulting in high recall.
The result wrapper uses a “popular vote” mechanism to resolve conflicts between base classifiers. For example, if two base classifiers recognize “White” as a PATIENT name and one base classifier considers it as a RACE, “White” will be classified as a PATIENT name. The wrapper also considers different privacy risks represented by different attribute types. For example, EID attributes present higher risk than QID attributes. If “White” is classified as a PATIENT (EID) by one base classifier and as a RACE (QID) by another base classifier, then the wrapper will assign it to PATIENT to minimize disclosure risk. However, if “White” is classified as a RACE by two classifiers but as a PATIENT by only one classifier, whether it will be assigned to PATIENT or RACE depends on the “voting weight” between the two types.

**Disclosure Risk and Data Utility with Value Enumeration**

Data privacy studies in the healthcare and medical domain typically concentrate on the risk of identity disclosure, which is consistent with the privacy principles underlying HIPAA. Similarly, the approach taken in this study also focuses on identity disclosure. That is, we consider limiting the risk of re-identifying an individual from the anonymized data released. A medical document typically contains rich information about symptoms, test results, diagnosis, medications and procedures, which are all sensitive. As such, it is very difficult, if not impossible, to protect against attribute disclosure; i.e., to develop privacy mechanism that masks sensitive medical information. This is perhaps one of the reasons why the HIPAA Safe Harbor rule focuses exclusively on re-identification issue.

Besides the HIPAA Safe Harbor rule, the generalization method can be used to protect against re-identification disclosure in our context. Figure 5 shows the generalized data with the minimum group size $k = 2$, where Age, Visit Date and Hospital are considered the QID attributes. The generalization hierarchy designed for this problem by the data owner allows the values of Visit Date to be generalized to a “season” of the year and the values of Hospital to be generalized to a region (Sweeney 2002). It can be observed that generalization causes significant information loss. Comparing to the results in Figure 2 where the value-enumeration method is used to anonymize the data, it is clear that value enumeration provides more detailed information than generalization. Note that the $k$-anonymity principle defines the re-

---

*Figure 4. Design of the Information Extraction Module*
identification risk based on the group size. Therefore, the risks of the patients in Figures 2 and 5 are the same; that is, the re-identification risk is 1/3 for record 1, 2 and 3 in the first group, and is 1/2 for record 4 and 5 in the second group.

| 1. Admitted on [Spring 2009], [Boston Area Hosp]. The [5-88] year old man is complaining fever, sore throat, headache, runny nose. |
| 3. Admitted on [Spring 2009], patient is a [5-88] year old female. Having runny nose, headache, vomiting. … [Boston Area Hosp]. |
| 4. [NAME] is a [17-44] year old male. Having pain on right side of abdomen, fatigue, dark urine. … Admitted [Summer 2010], [LA Area Hosp]. |

**Figure 5. Example Clinical Text Anonymized with Generalization**

We now define some terms necessary to analyze the disclosure risk in the value enumeration method.

**Definition 1.** The **grain** of a QID attribute refers to the level of the QID attribute values to be used for value enumeration.

Information contained in a grain is typically more detailed than that allowed by the HIPAA Safe Harbor rule or that generated by \( k \)-anonymity. For example, in Figure 2 the grain is “month-year” for the Visit Date attribute and hospital name for the Hospital attribute (and it is the age value itself for the Age attribute). They are more detailed than the corresponding attribute values in the right panel of Figure 1 (HIPAA Safe Harbor) and in Figure 5 (\( k \)-anonymity). Like generalization hierarchy in \( k \)-anonymity, the grain is determined by the data owner.

**Definition 2.** Given a dataset of \( d \) QID attributes, \( A_1, \ldots, A_d \), let \( M_j (j = 1, \ldots, d) \) be the number of distinct grain values of the \( j \)th QID attribute. Let \( a_{m_j} (m_j = 1, \ldots, M_j) \) be a distinct grain value of \( A_j \). A **profile** refers to a combination of the grain values of the \( d \) QID attributes, expressed as \( \{A_1 = a_{m_1}, \ldots, A_d = a_{m_d}\} \).

For instance, the number of distinct grain values of Visit Date in Figure 2 is four (while the number of distinct values of Visit Date in the original data in Figure 1 is five). One possible profile for record 1 in Figure 2 is \{Visit Date = ‘Mar-2009’, Hospital = ‘Mass General Hosp’, Age = 5\}.

Let \( M_j^g \) \((j = 1, \ldots, d)\) be the number of distinct grain values of the \( j \)th QID attribute in group \( g \). Then the number of possible profiles for each record in group \( g \) is \( \prod_{j=1}^d M_j^g \). We call this quantity the **profile size**.

That is, the profile size refers to the number of possible combinations of the grain values of all QID attributes. For example, the profile size for each record in the first group in Figure 2 is \( 2 \times 2 \times 3 = 12 \). It is easy to see that in order not to increase the re-identification risk defined by the group size, the profile size must satisfy a constraint as follows:

**Profile size constraint:** The profile size must be greater than or equal to the group size; i.e.,

\[ \prod_{j=1}^d M_j^g \geq k_g, \]  

(1)

where \( k_g \) is the number of records in group \( g \), which is no smaller than the minimum group size \( k \).

Figure 6 shows a scenario generated using the cluster-level value enumeration that satisfies constraint (1). This dataset provides much more detailed information about patients than that in Figure 2 – the month-year and hospital information is given as a single value. If a doctor receiving this data is interested in getting more detailed information about a particular patient, the doctor can contact the hospital and request an approval (e.g., patient consent) to access the original record.
1. Admitted on \{Apr-2009\}, \{Tufts Med Ctr\}. The \{5, 52, 88\} year old man is complaining fever, sore throat, headache, runny nose.
2. [NAME] is a \{5, 52, 88\} year old female. Visited on \{Mar-2009\}. Having joint pain, sore throat, fever. ... \{Mass General Hosp\}.
3. Admitted on \{Apr-2009\}, patient is a \{5, 52, 88\} year old female. Having runny nose, headache, vomiting. ... \{Mass General Hosp\}.
4. [NAME] is a \{17, 44\} year old male. Having pain on right side of abdomen, fatigue, dark urine. ... Admitted \{Aug-2010\}, \{USC Univ Hosp\}.
5. Visited on \{Jul-2010, Aug-2010\}. Female, \{17, 44\} year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. ... \{UCLA Med Ctr\}.

Figure 6. Example Clinical Text Anonymized with Proposed Approach (Scenario 2)

In terms of disclosure risk, however, the records in this dataset appear to have higher re-identification risks than those in Figure 2, even though the risk based on \(k\)-anonymity definition is the same in both scenarios (\(1/3\) for the records in the first group and \(1/2\) for those in the second group). The problem is that the re-identification risk defined by \(k\)-anonymity does not consider the background knowledge a privacy intruder may have. Suppose the intruder knew that the visit date of a patient is in March 2009. Then the intruder can uniquely identify record 2. A similar re-identification can be made using known hospital information. To reduce the risk due to such background knowledge, we introduce a parameter, called \textit{profile size multiplier} and denoted \(r\), and revise the profile size constraint (1) as:

\[
\prod_{j=1}^{d} M_{j}^{g} \geq r_{g} \cdot \tag{2}
\]

Clearly, a larger \(r\) value implies a higher privacy protection level. Given the grains of the QID attributes, the problem of value enumeration can now be formulated as:

\[
\min_{g} \prod_{j=1}^{d} M_{j}^{g} , \tag{3}
\]

subject to \(\prod_{j=1}^{d} M_{j}^{g} \geq r_{g} \), \(\forall g\).

Finding a global optimal solution for this problem is computationally prohibitive because the process involves computation of all possible profiles in any possible group. We discuss how to approximate the optimal solution using an efficient clustering algorithm in the next section.

1. Admitted on \{Mar-2009, Apr-2009\}, \{Tufts Med Ctr\}. The \{5, 52, 88\} year old man is complaining fever, sore throat, headache, runny nose.
2. [NAME] is a \{5, 52, 88\} year old female. Visited on \{Mar-2009, Apr-2009\}. Having joint pain, sore throat, fever. ... \{Mass General Hosp\}.
3. Admitted on \{Mar-2009, Apr-2009\}, patient is a \{5, 52, 88\} year old female. Having runny nose, headache, vomiting. ... \{Mass General Hosp\}.
4. [NAME] is a \{17, 44\} year old male. Having pain on right side of abdomen, fatigue, dark urine. ... Admitted \{Jul-2010, Aug-2010\}, \{USC Univ Hosp\}.
5. Visited on \{Jul-2010, Aug-2010\}. Female, \{17, 44\} year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. ... \{UCLA Med Ctr\}.

Figure 7. Example Clinical Text Anonymized with Proposed Approach (Scenario 3)

The anonymized data in Figure 6 is obtained by setting \(r = 1\), while Figure 2 results from setting \(r = 4\). Figure 7 shows a third scenario, which is obtained by setting \(r = 2\). It is easy to see that given an \(r\) value, the solution is not necessarily unique. In this situation, the data owner can prioritize the order of anonymization for QID attributes. For example, Figure 7 is obtained by anonymizing the Visit Date attribute rather than the Hospital attribute.
**Document Clustering**

Module 2 of the proposed framework clusters patient documents based on medical concepts, represented by HMD terms extracted from Module 1. There exist a variety of document clustering techniques, as described in a survey by Carpineto et al. (2009). However, none of them can be adopted directly for our purpose. We aim at clustering the data such that the documents within a group are more similar with respect to a medical concept. At the same time, we attempt to cluster the data with an appropriate group size for each group to strike a balance between disclosure risk and information loss – an overly large group size would result in too many QID values enumerated, rendering the information useless; whereas with an overly small size, too few QID values would be listed, causing a high disclosure risk. Therefore, the clustering algorithm should be able to control the size of each cluster. Some well-known clustering techniques, such as $k$-means clustering, require the number of groups (clusters) to be specified as an input, and thus are not appropriate for our clustering purpose.

We consider a powerful document clustering algorithm, based on a state-of-the-art technique called Non-negative Matrix Factorization (NMF) (Lee and Seung 1999). Given a collection of $n$ documents with a total vocabulary of $m$ terms, let $\mathbf{X}$ be an $m \times n$ term-document matrix. The NMF factorizes $\mathbf{X}$ into two non-negative matrices, $\mathbf{W} = [w_{ip}] \ (i = 1, \ldots, m; p = 1, \ldots, c)$ and $\mathbf{H} = [h_{jq}] \ (q = 1, \ldots, c; j = 1, \ldots, n)$, such that $\mathbf{X} \approx \mathbf{WH}$, where $c << \min(m,n)$ is a pre-specified parameter. The NMF problem can be formulated as:

$$
\min_{\mathbf{W}, \mathbf{H}} \| \mathbf{X} - \mathbf{WH} \|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - (\mathbf{WH})_{ij})^2,
$$

subject to $\mathbf{W}$ and $\mathbf{H}$ being non-negative; i.e., $w_{ip} \geq 0$, $h_{jq} \geq 0$, $\forall i, p, q, j$.

The NMF has a very appealing interpretation. Each of the $c$ columns of $\mathbf{W}$ is a basis vector representing a semantic feature, i.e., a set of words denoting a particular concept or cluster, with each element $w_{ip}$ indicating the degree to which term $i$ belongs to cluster $p$. Matrix $\mathbf{H}$ describes the contribution of the documents to these concepts or clusters, with each element $h_{jq}$ representing the degree to which document $j$ is associated with cluster $q$. Unlike other matrix-decomposition-based clustering methods that generally have negative entries in the component matrices, NMF ensures that all elements are non-negative. This is consistent with the non-negativity nature of the original term-document matrix, and also avoids subtractive linear combination of base semantic features, an operation that is hard to interpret.

It is computationally prohibitive to find a global optimal solution for the NMF (Lee and Seung 2001). However, studies have shown that the NMF-based methods outperform, in terms of clustering quality, many other popular document clustering methods, such as $k$-means clustering, spectral clustering, singular value decomposition, and graph-based clustering (Kuang et al. 2012; Xu et al. 2003). The NMF is also computationally efficient, running either faster or comparable to the other popular algorithms.

The NMF, however, cannot be used directly for our clustering purpose because it requires the number of clusters, $c$, as an input. In our clustering task, the number of clusters cannot be pre-specified because each cluster must satisfy the requirement on the minimum number of records, as well as the profile size constraint (2). To address this issue, we propose a recursive NMF procedure, as a component of Module 2 in our framework. Essentially, this procedure performs a recursive binary partitioning of the data, using NMF for each binary split, until each group cannot be further partitioned due to the minimum group size and profile size requirements. A sketch of the recursive binary NMF algorithm is given in Figure 8.

1. Given a term-document matrix $\mathbf{X}$, partition $\mathbf{X}$ into two clusters (i.e., $c = 2$) using the NMF model (4).
2. Partition the subset of the data in each cluster into two sub-clusters using NMF again.
3. Repeat Step 2 for each of the two sub-clusters. Stop the process if the sub-cluster cannot satisfy the minimum group size or profile size requirement.

**Figure 8. Recursive Binary NMF Algorithm**

The computational complexity of the traditional NMF algorithm is of $O(\text{cnt})$, where $t$ is the number of iterations that is controllable by the user (Xu et al. 2003; the default value in our system is $t = 50$, a
commonly used value). For the proposed recursive binary NMF algorithm, each binary split takes $O(nt)$ time; and it requires $c-1$ such splits to get $c$ clusters. Therefore, the time complexity of the proposed NMF algorithm is also of $O(cnt)$.

The Prototype System

We have developed a prototype system, called DAST (De-identification and Anonymization for Sharing medical Texts), based on the proposed framework described above. The system is implemented in Java and text data is organized using XML. Screenshots of the DAST prototype are shown in Figures 9, 10 and 11. The prototype system includes six main menus. The File and Edit menus have functions similar to those in most application systems (e.g., to open, save, edit and print a file). The Extract, Cluster, and Anonymize menus represent respectively the three modules described earlier in Figure 3. There is also a Help menu to provide documentation and user support.

We illustrate how the text data are processed in DAST using a clinical note example (record 2 in Figure 1). Figure 9 displays the original text marked up with XML tags. Figure 10 shows the results of information extraction on the same record, including the extracted EID, QID and HMD values. The EID values will be removed or replaced with a constant (de-identification). The HMD terms will be used for clustering text documents. The QID values will be converted into grain-level values for the value enumeration. Figure 11 shows the final results after de-identification and value enumeration.
Experimental Evaluation

We use the data provided by the Informatics for Integrated Biology and the Bedside (i2b2) project for this study. The i2b2 project has obtained multiple sets of clinical documents from healthcare organizations and made them available for research (https://www.i2b2.org/NLP/DataSets). We use three of the datasets for the experimental evaluation. The first set (889 records) is related to a clinical-text de-identification challenge competition. The second set (1237 records) was initially used for evaluating document classification techniques. The third set (871 records) was used for a challenge competition to extract medical concepts, assertions and relations. All three datasets are medical discharge summaries; so they are called Discharge-1, Discharge-2, and Discharge-3, respectively. The elements of information contained in these datasets include patient name, admission and discharge date, age, gender, hospital, symptoms, test results, diagnosis, disease, medications, and so on. To protect patient privacy, all PHI values had already been replaced with surrogate values by the data providers, which are realistic semantically and consistent with the original format. For example, “Mr. Anderson” may be replaced by “Mr. Taylor” and “6/21/2003” replaced by “9/16/1997.”

Information Extraction

In the DAST prototype, we implemented three base classifiers. The first is an SVM-based classifier that relies mostly on the local features to classify a term as part of a certain PHI or non-PHI. The second is a CRF-based pattern recognizer that utilizes both local and global features to classify a term. The third is a rule-based classifier that mainly uses lexical cues (e.g., prefixes such as “Dr.”, etc.) and record structure information (e.g., the “cc:” and “signed by” sections in a record) to recognize PHI. On top of these three base classifiers, we implemented a result wrapper using the union and popular vote strategies discussed earlier to combine the outputs of the three base classifiers and produce the final results. We assigned three voting weights to three types of attributes, 3.0 for EID, 2.0 for QID, and 1.0 otherwise, which represent the levels of risk involved in these attributes.

Out of the three datasets, Discharge-1 is the only one that has all PHI marked in the text for testing extraction performance purposes. In fact, Discharge-1 contains two sets of records: a training set of 669 records and a testing set of 220 records. We focused on the extraction of three PHI categories in this experiment: Patient Name (EID), Admission Date (QID), and Age (QID). In the three base classifiers, the CRF-based classifier performed the best, with a recall of 86.64% and a precision of 91.36%. The SVM-based classifier was the second, with a recall of 82.72% and a precision of 88.93%. The rule-based classifier performed the worst, with a recall of 37.78% and a precision of 79.36%. The final results generated by the wrapper showed a recall of 90.94% (about 4.3 percentage points higher than the CRF classifier) and a precision of 89.85% (about 1.5 percentage points lower than the CRF classifier). We believe this tradeoff is worthwhile because, as explained earlier, recall should be considered more important than precision in PHI detection.
Privacy Risks

All three datasets were used in the experiments for evaluating privacy risk and data utility. After extracting the QID attributes from the text, we found that there were very few zip code and/or location values in the text that we could use for the experiment. Therefore, we focused on two QID attributes for this experiment: Age and Admission Date. The grains for the two attributes are age in year and admission month-year. We set the minimum number of records in a cluster, $k = 5$, and used three profile size multipliers: $r = 1$, $2$, and $3$, which means that the profile size (i.e., the number of possible combinations of the grain values of the two QID attributes) for each record is at least $5$, $10$ and $15$, respectively.

Based on the HIPAA Safe Harbor rule, Admission Date values should be truncated into admission year (AdminYear) while Age values can remain unchanged. When the data is processed in this way, it is HIPAA-compliant. However, there are many unique Age and Age-AdminYear combination values in the HIPAA-compliant data. Table 2 shows the number and percentage of these unique values for each dataset. Therefore, if a data snooper knew the Age and admission year (or even only Age) of a patient in a dataset, it is not difficult for the snooper to re-identify the patient and subsequently reveal all the information of the patient. With the data anonymized using the proposed value-enumeration method, it is not possible to uniquely identify any record using the value of Age or Age-AdminYear combination. Therefore, the disclosure risk with the data anonymized using our approach is substantially lower. Even with lower disclosure risk, our anonymized data provides more accurate time period information (month-year) than HIPAA-compliant data (year only), which is critical to many medical and healthcare studies such as epidemiologic research.

Table 2. Unique Values in HIPAA-Compliant Data

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Unique Age Value</th>
<th>Unique Age-AdminYear Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Percentage</td>
</tr>
<tr>
<td>Discharge-1</td>
<td>22</td>
<td>2.5%</td>
</tr>
<tr>
<td>Discharge-2</td>
<td>12</td>
<td>1.0%</td>
</tr>
<tr>
<td>Discharge-3</td>
<td>17</td>
<td>2.0%</td>
</tr>
</tbody>
</table>

Data Utility

Data utility can be evaluated in many different contexts. Following Rizvi and Haritsa (2002), we evaluate the performance in data utility in the context of association rule mining, or more specifically large itemset mining. This is reasonable because many medical document sharing applications can be viewed as finding large itemsets and associations in the documents. We focused on mining associations between the medical concepts and the admission month, which represents seasonal health information that is essentially not available if the HIPAA Safe Harbor rule is followed strictly. We compare the mining results using data generated by the DAST system with that by a traditional HIPAA-compliant de-identification technique.

We converted the document datasets into structured medical-term-association tables. Each row of the table corresponds to a document in the original dataset. Each column represents a medical term extracted from Module 1 of the DAST system and there are a few hundred columns in each dataset. These medial terms indicate the medical concepts and topics covered by the dataset. If a medical term $i$ appears in document $j$, then the $(j, i)$ element in the medical-term-association table will be assigned a value 1; otherwise, it will be zero. Additional columns were added to the table to represent the admission months.

Similarly, we converted the anonymized datasets, both the HIPAA-compliant datasets and those anonymized by the DAST system, into medical-term association tables. Because a record anonymized with value enumeration contains multiple admission month values, we randomly choose one from all the enumerated months for that record. The selection process was designed so that month values appeared in a cluster more often are more likely to be selected. For example, suppose a cluster contains 6 records and three month values: “Jan” in three records, “Feb” in two records, and “Mar” in one record. Then for each record in this cluster, “Jan” will have a 1/2 chance being selected and assigned to the record, “Feb” will
have a 1/3 chance, and “Mar” will have a 1/6 chance. There are no month values for the HIPAA-compliant data. So, the month values in the data were generated randomly based on the month-value distribution in the original data.

For each dataset, an association rule mining algorithm was run on the three tables (i.e., the original table, the HIPAA-compliant table, and the value-enumerated table) to discover associations between medical terms and the admission months. We compared the large itemsets discovered from an anonymized table with those from the original table based on an error measure defined below (Evfimievski et al. 2004; Rizvi and Haritsa 2002):

\[
\text{Error rate in support count} = \frac{1}{|L|} \sum_{l \in L \setminus \{\cdot\}} \left| \frac{\hat{F}_l - F_l}{F_l} \right|,
\]

where \(L\) represents the set of all large itemsets with support count larger than the minimum support threshold (which was set to 10 in this experiment); \(F_l\) and \(\hat{F}_l\) are the frequency count of the \(l\)th large itemset from the original dataset and the anonymized dataset (HIPAA-compliant or value-enumerated), respectively. Clearly, a small error value indicates that the association result based on the anonymized data is close to that on the original data, which is desirable.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>(r = 1)</th>
<th>(r = 2)</th>
<th>(r = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DAST</td>
<td>HIPAA-Compliant</td>
<td>DAST</td>
</tr>
<tr>
<td>Discharge-1</td>
<td>0.1336</td>
<td>0.4661</td>
<td>0.1764</td>
</tr>
<tr>
<td>Discharge-2</td>
<td>0.2124</td>
<td>0.4713</td>
<td>0.2535</td>
</tr>
<tr>
<td>Discharge-3</td>
<td>0.1458</td>
<td>0.4256</td>
<td>0.1813</td>
</tr>
</tbody>
</table>

* The results are statistically significantly different at \(\alpha = 0.00001\) for all pairwise comparisons except for Discharge-2 when \(r = 3\), where the significance level is \(\alpha = 0.003\). The experimental procedure was repeated 10 times on each dataset considering the effect of the random selection of month values. The average error rates over the 10 runs are reported in Table 3. It is clear that the average error rates on the data generated by DAST are much smaller than those by the HIPAA-compliant method. The results of the two methods are statistically significantly different at \(\alpha = 0.00001\) for all pairwise comparisons except for Discharge-2 when \(r = 3\), where the significance level is \(\alpha = 0.003\). Therefore, our proposed approach preserves data utility significantly better than the HIPAA-compliant approach. The results also show that as \(r\) increases the error rate associated with DAST increases. This is expected because a larger \(r\) value means a larger profile size (i.e., the number of possible combinations of the grain values of the two QID attributes), resulting in more enumerated month-year values for a record. On the other hand, a larger \(r\) value implies a higher privacy protection level, as discussed earlier.

**Conclusions and Future Research**

In this study, we investigate the issues related to privacy protection in sharing unstructured medical documents. We propose a novel and innovative privacy protection framework that integrates and improves technologies developed in the data privacy and health informatics communities to reduce privacy disclosure risks in sharing medical documents while preserving the utility of the data. We also demonstrate the validity and effectiveness of the proposed framework by developing and evaluating a prototype system. Our experiments show that the proposed method significantly outperforms the traditional HIPAA-compliant method. We should mention that the proposed framework is designed for sharing a set of patient documents, not for sharing an individual document. This is the same setting as that assumed in data privacy literature on the structured data (Aggarwal and Yu 2008; Sweeney 2002).

We plan to pursue future research along several lines. We will explore more effective ways to further improve the performance of machine learning and document clustering techniques in our framework. Due to privacy concerns, it is very difficult to obtain real-life medical documents with true PHI data. Although we have used three datasets in the experimental evaluation, they are all medical discharge summary
documents. We will collect different types of medical or healthcare documents and conduct more comprehensive studies on the usability and performance of the proposed approach. This work studies the problem where each individual has only one record. It is also interesting to examine the privacy problem where an individual may have multiple records in the dataset. This is an under-studied problem even in the data privacy research on the structured data. This topic deserves an in-depth investigation in future research.

Acknowledgments

Xiao-Bai Li’s research was supported in part by the National Library of Medicine of the National Institutes of Health under Grant Number R01LM010942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


