HHS Public Access

Author manuscript

Inf Syst Res. Author manuscript; available in PMC 2018 March 19.

Published in final edited form as:

Inf Syst Res. 2017 ; 28(2): 332–352. doi:10.1287/isre.2016.0676.

Anonymizing and Sharing Medical Text Records

Xiao-Bai Li[a] and **Jialun Qin**[a]

[a]Department of Operations and Information Systems, Manning School of Business, University of Massachusetts Lowell, Lowell, Massachusetts 01854

Abstract

Health information technology has increased accessibility of health and medical data and benefited medical research and healthcare management. However, there are rising concerns about patient privacy in sharing medical and healthcare data. A large amount of these data are in free text form. Existing techniques for privacy-preserving data sharing deal largely with structured data. Current privacy approaches for medical text data focus on detection and removal of patient identifiers from the data, which may be inadequate for protecting privacy or preserving data quality. We propose a new systematic approach to extract, cluster, and anonymize medical text records. Our approach integrates methods developed in both data privacy and health informatics fields. The key novel elements of our approach include a recursive partitioning method to cluster medical text records based on the similarity of the health and medical information and a value-enumeration method to anonymize potentially identifying information in the text data. An experimental study is conducted using real-world medical documents. The results of the experiments demonstrate the effectiveness of the proposed approach.

Keywords

privacy; information extraction; document clustering; anonymization; data analytics

1. Introduction

The advances in health information technology have enabled organizations to store, share, and analyze a large amount of personal health and biomedical data. As a result of widespread implementation of electronic health records (EHR) systems, stimulated intensely by the Health Information Technology for Economic and Clinical Health Act (HITECH Act 2009), there has been an explosion of health data being generated and collected by health organizations. In tandem with this unprecedented growth of available electronic data, secondary use of the data has increased rapidly across a wide variety of health domains. Empowered by health informatics and analytic techniques, secondary use of health data can aid clinical decision making; expand knowledge about diseases, genetics, and medicine; improve patients' healthcare experiences; reduce healthcare costs; and support public health

Contact: xiaobai_li@uml.edu, http://orcid.org/0000-0001-8009-8439 (X-BL); jialun_qin@uml.edu (JQ).

History: Gedas Adomavicius, Senior Editor; Vijay Khatri, Associate Editor.

policies (Jensen et al. 2012, Safran et al. 2007, Wylie and Mineau 2003). As a result, there is a growing trend for using EHR data beyond direct healthcare delivery purposes.

While secondary use of health data has significantly enhanced the quality and efficiency of medical research and healthcare management, there are growing concerns about patient privacy due to the widespread practice of health information sharing (Garfinkel et al. 2007, Safran et al. 2007). To respond to these concerns, the Health Insurance Portability and Accountability Act (HIPAA) established a set of privacy rules, which was strengthened by the HITECH Act (2009). The HIPAA *Safe Harbor* (SH) 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). A full list of PHI elements is provided in Table 1 (DHHS 2000, pp. 82818–82819). A strict implementation of the SH rule, however, may be inadequate for protecting privacy or preserving data quality. Recognizing this limitation, HIPAA also provides alternative guidelines that enable a *statistical* assessment of privacy disclosure risk to determine if the data are appropriate for release.

Along the lines of the statistical 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 that can be stored in well-defined relational databases (Aggarwal and Yu 2008). Various techniques have been developed to anonymize structured data. However, health and medical data in EHR systems and medical research platforms come from heterogeneous sources with various representations (Carter 2008, Murphy et al. 2010). They include well-structured data (e.g., identity, registration, and demographic data), set-valued data (e.g., codes such as the International Classification of Diseases (ICD) and the Logical Observation Identifiers Names and Codes (LOINC)), free text (e.g., clinical narratives, pathology reports, and discharge summaries), and images from radiology systems. These diverse data representations and formats present challenges that existing privacy-preserving approaches for structured data cannot effectively deal with. Therefore, in spite of the abundance of research in data privacy, its application in health and medical domains lags behind.

This research concerns privacy protection issues when medical documents containing free text, such as clinical narratives and discharge summaries, are shared for secondary use. In such cases, some identifying information is embedded in the text, where anonymization techniques designed for structured data are not readily applicable. The majority of privacy research in sharing and releasing medical documents has followed the SH rule and focused on the automatic detection of PHI attributes in the documents (Meystre et al. 2010, Uzuner et al. 2007, Murphy et al. 2011). The identified PHI values are then simply removed from, or replaced with a constant value in the released text documents. Studies have shown that such a simple de-identification strategy lacks the flexibility to adequately meet the diverse needs of data users (Meystre et al. 2010, Uzuner et al. 2007). There is a lack of research on how to mask the PHI fields (other than simple removal) and how to cope with non-PHI but potentially identifying information to improve privacy protection and data utility. This research addresses this problem.

It is well documented that anonymization approaches are more effective when applications on the anonymized data are more specific than general (Aggarwal and Yu 2008, Fung et al. 2010). Applications and analyses based on structured data typically include traditional statistical analysis such as regression and multivariate analysis and data mining applications such as classification and clustering. Anonymization techniques are developed for these applications accordingly (Agrawal and Srikant 2000, Aggarwal and Yu 2008, Fung et al. 2010, Li and Sarkar 2011, Melville and McQuaid 2012). For unstructured data such as medical text data, applications often include text mining tasks such as medical keyword-based search query and information extraction, as well as those that are also suitable for structured or semistructured data such as counting query and association analysis (Jensen et al. 2012, Meystre et al. 2008, Murphy et al. 2010). We develop our approach with these applications in mind.

In this study, we propose a framework to extract, cluster, de-identify, and anonymize the medical documents, focusing on the analysis and design aspects of privacy-preserving technology. 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 documents; (ii) document clustering, which clusters patient documents based on the similarity of the health and medical concept, using a recursive binary partitioning method; and (iii) de-identification and anonymization, which remove explicit individual identifiers and anonymize potentially identifying attributes using a novel cluster-level value-enumeration method. The proposed framework is implemented into a prototype system. The main contributions of this research are summarized below.

Identifying weakness in HIPAA Safe Harbor based de-identification

We provide clear evidence from real-world data that the SH rule for de-identifying data can be underprotective (i.e., de-identified data having high disclosure risk) in some cases and overprotective (i.e., resulting in poor data utility) in others. Identifying and investigating this problem has significant research and practical implications, since the current mainstream approaches in both research and practice follow the SH rule to protect privacy by detecting and removing the PHI elements from medical documents.

Document clustering using medical information and recursive partitioning

Existing statistical or computational approaches in the data privacy literature typically cluster the records based on demographic attributes, which often work well for the applications based on structured data but are not so effective for those based on text documents. We introduce a novel idea of clustering documents based on health and medical information, which significantly increases the utility of the anonymized data. The proposed clustering approach is implemented with a recursive binary partitioning algorithm for controlling the level of disclosure risk, which is new to the literature.

Value enumeration for anonymizing data at the cluster level and the dataset-level

Unlike traditional methods that use generalization or noise perturbation to anonymize the data after clustering, we propose a novel value-enumeration method for anonymizing data,

which results in a smaller information loss than traditional methods. We introduce the notion of cluster-level and dataset-level anonymity for value enumeration and develop a drill-down method to further reduce information loss in the anonymized data.

In the next section, we motivate our research using a real-life example and demonstrate the problems with the current SH approaches and the basic idea of our proposed approach using similar examples. In Section 3, we review related work in data privacy and health informatics. In Section 4, we present the analysis and design of the proposed document sharing framework. We report the results of the experimental study on a set of real-life data in Section 5. We then discuss implications and limitations of our research in Section 6. In the last section, we summarize our work and discuss future research.

2. Motivating Examples

We first provide a real example of clinical text data, which is a discharge summary originally provided by Partners Healthcare in Massachusetts (Uzuner et al. 2007). To make the data available for research, the authentic PHI values (e.g., patient and physician names, dates, and hospital names) in the original data were replaced with reasonable surrogates, which are realistic semantically and consistent with the original format, by the data provider before the data were released. The document, which is shown in Figure 1, includes a header that provides admission and discharge dates. The main body of the discharge summary includes several sections, such as history of present illness, past medical history, physical examination, and discharge medications. Presumably, there will be a record of structured data associated with this text document, which may include patient name, date of birth, gender, address, phone number, and so on. In this study, we focus on anonymizing and sharing medical text data without considering the structured data. We will discuss later in Section 6 how to apply our approach to anonymize the associated structured data as well.

The information in the document in Figure 1 can be classified into six categories: (a) patient and physician names, (b) admission and discharge dates, (c) hospital name, (d) patient age, (e) patient gender, and (f) health and medical information. The first three categories are PHI while the remaining three categories are not. When medical text data are used for research and analysis, typically a set of such documents are shared together (Safran et al. 2007, Jensen et al. 2012). To examine the problem with a set of documents, consider an example of five short clinical notes shown in Figure 2, created based on the real example in Figure 1. While the text in the notes are highly simplified because of space limitations (particularly for the health and medical information), the six categories of information mentioned above still appear in these simplified notes.

The original text is shown in the left panel of Figure 2. In compliance with the HIPAA SH rule, the patient and hospital names must be removed, and the date expression must be truncated to include only the year. On the other hand, HIPAA permits releasing non-PHI items, such as age, gender, and health, and medical information, without any change. As a result, the HIPAA-compliant de-identified notes are shown in the right panel of Figure 2.

Although the de-identified documents are HIPAA-compliant, they can be overprotective or underprotective. 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 hard to identify the 88-year-old man in record 1 using the other information (e.g., gender, admission year, and geographical area) in the notes. So, existing de-identification approaches could put a patient's privacy at risk and at the same time provide data with significant information loss.

To overcome these limitations, the proposed approach first clusters the patients' documents based on health and medical information, and then anonymizes the potentially identifying attributes using a cluster-level value-enumeration method. Figure 3 illustrates how the proposed approach works for the example shown in Figure 2. With the proposed approach, the five records are clustered into two groups based on the patient's symptoms and health data. The first group includes records 1–3, which show flu-like symptoms. The second group includes records 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 very 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. Compared to the SH approach, the value-enumeration method may provide more detailed information for some fields (e.g., dates and hospitals); it may also provide less detailed information for other fields (e.g., age) if privacy concerns are present. The level of detail can be controlled by the data owner. For example, the age values are enumerated because the 88-year-old is highly distinguishable. If individual age values do not cause privacy concerns, they can be directly presented without value enumeration.

To see why the proposed clustering method provides better data utility than traditional clustering-based methods, in Figure 4 we show the results using a popular traditional method called *k*-anonymity (Sweeney 2002), where records are clustered into two groups based on similarity in demographic and date attributes, and concerned values are replaced with generalized values. The first group includes records 1, 2, and 5, while the second group includes records 3 and 4. The data set satisfies 2-anonymity since each record cannot be distinguished from at least another record with respect to the generalized values. Suppose the anonymized data are used to support a keyword-based search query. When a user enters the terms "abdomen pain" and "fatigue," records 4 and 5 will be retrieved from the data anonymized by either the proposed method or *k*-anonymity. The age value from the *k*-anonymized data will show [5–17] and [52–88]; the hospital values will be [Southern Region Hosp] and [Northeastern Region Hosp], and the date values will be [Mar–Jul 2009] and [Apr–Aug 2009]. The proposed method clearly provides more useful age, hospital, and date information than *k*-anonymity. A similar comparison can be observed for the other group if the user enters "runny nose" and "fever."

We should mention that for convenience we have placed the records belonging to the same cluster together in Figures 3 and 4, but it is clear that the anonymized records in the released

set do not have to be grouped by cluster. If necessary, these records can be shuffled across clusters before releasing.

3. Related Work

There are two types of privacy disclosure widely recognized in the literature (Duncan and Lambert 1989): (a) *identity disclosure* (or *reidentification*), which occurs when a data intruder is able to match a record in a data set 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 (Fung et al. 2010).

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 that 87% of the population in the United States can be uniquely identified with three QID attributes—gender, date of birth, and five-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 a PHI field (e.g., gender). The third type is *sensitive attributes*, which contain private information that a person typically does not want disclosed, such as sensitive diseases, sexual orientation, and personal financial information. None of the items listed in Table 1 are a sensitive attribute. That is, HIPAA does not provide guidelines on how to protect sensitive attribute data; instead, the basic idea of the HIPAA SH rule is to protect privacy by preventing identity disclosure. This study also focuses on identity disclosure only.

3.1. 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 data set 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 k is called the *minimum group size*. In fact, it is easy to see that the *maximum* reidentification risk for any individual record in a k-anonymized data set is $1/k$. So, parameter k is an important privacy risk measure. Privacy risk can also be measured by *average* reidentification risk based on *actual group sizes* (which may be greater than k), discussed in more detail in Section 5.

The k-anonymity approach generalizes different but similar QID values into a higher-level value within a group; it leaves sensitive attribute values unchanged. Because individuals in a group have the same generalized QID values, they are subject to high attribute-disclosure

risk if their sensitive values are the same or similar. That is, while *k*-anonymity is effective in protecting against identity disclosure, it does not consider attribute-disclosure risk. To address this issue, a privacy principle called *l*-diversity has been proposed (Machanavajjhala et al. 2006). The *l*-diversity principle requires that a sensitive attribute should include at least *l* well-represented values in each group of the *k*-anonymized data. Another privacy principle, called *t*-closeness (Li et al. 2007), addresses the issue by requiring that, for each group, the distance between the distributions of the sensitive attributes in the group and the overall distribution of the sensitive attributes cannot be larger than a threshold value *t*. The *l*-diversity, *t*-closeness, and the other privacy principles focusing on sensitive value protection for structured data typically assume that there is one or a few predefined sensitive attributes, which is not a realistic scenario for text data. Medical text documents typically have a large number of unstructured (not predefined) sensitive attributes, such as symptoms, conditions, test results, diagnosis, and treatments. It is difficult, if not impossible, to apply the idea of *l*-diversity or *t*-closeness for the sensitive attributes in medical document data.

Various anonymization methods have been proposed in the literature, including generalization, suppression, swapping, and noise perturbation (Cooper and Collman 2005). Generalization and suppression either generalize the original values to a higher level category or remove the values (Sweeney 2002). Swapping involves an 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, Melville and McQuaid 2012), which applies mainly to numeric data. There are also studies that address privacy disclosure problems by hiding sensitive information (Menon et al. 2005). Applying an anonymization method to data causes information loss and reduces data utility. It is thus necessary to evaluate anonymization methods based on some data utility measures together with the privacy risk measures. The data utility measures are typically related to the changes in summary statistics or data analysis results. We discuss more about data utility measures in Section 5.

3.2. De-Identification for 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 have been mainly based on the SH principle, focusing on the detection and removal of PHI items from the documents. Meystre et al. (2010) and Uzuner et al. (2007) have reviewed more than a dozen state-of-the-art techniques in the field, all of which follow the SH approach. The task of detecting PHI items from the 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 (Friedlin and McDonald 2008). 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 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 data sets. In terms of performance, manually crafted rules often miss unexpected or nonstandard variations of PHI terms, resulting in a low proportion of relevant terms being retrieved. To overcome these limitations, supervised machine learning methods are often used to categorize PHI terms (Wellner et al. 2007). Popular machine learning methods used include, for example, support vector machines (SVM) (Cortes and Vapnik 1995) and conditional random fields (CRF) (Lafferty et al. 2001). Studies have shown that in general, machine-learning-based techniques perform better than pattern-matching-based techniques (Uzuner et al. 2007, Meystre et al. 2010). Many machine-learning-based techniques also incorporate some rules and dictionaries to further improve performance. A disadvantage of a machine-learning-based approach is that it requires fairly large corpuses of annotated documents as training data.

The effectiveness of machine-learning techniques can be further improved by using ensemble learning, 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 to improve machine learning performance on classification problems (Wolpert 1992, Tan and Gilbert 2003). Prior research has shown that performances of different PHI detection techniques vary in different scenarios (Uzuner et al. 2007). Therefore, combining multiple detection algorithms through ensemble learning is likely to improve the overall performance for PHI detection. While there are a considerable number of studies that apply the ensemble approach for classification problems on structured data (Polikar 2006) and for text document classification (Sebastiani 2002), we have not found any publication that uses the ensemble approach for detecting PHI and QID fields in the literature.

Gardner and Xiong (2009) propose a framework for de-identifying medical text data. Their framework includes, in addition to the SH implementations, a *k*-anonymity-based alternative, which groups the documents based on QID attributes such as age and location, and anonymizes the QID values by generalization. We have shown in Section 2 that grouping data by QID attributes may not serve well for the purposes of sharing medical documents. Therefore, it is desirable to have documents categorized based on the health and medical information. In addition, Gardner and Xiong (2009) directly adopt the *k*-anonymity's generalization method to anonymize data, which, as we have shown, tends to be overprotective and undermine the utility of the anonymized data. The value-enumeration method we propose provides better data utility than the generalization method with the same disclosure risk.

In summary, current privacy research in sharing medical documents focuses mainly on the SH approach. 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.

4. The Proposed Approach

Regarding the statistical approach to de-identification, the U.S. Department of Health and Human Services has recommended the following principles to determine disclosure risk (OCR 2012): (i) replicability, which refers to the chance the health information will consistently occur in relation to the patient; (ii) availability of external data sources that contain the patients' health information; and (iii) distinguishability of the patient's data. In general, the greater the replicability, availability, and distinguishability of the health information, the greater the disclosure risk. For example, identity and demographic data are highly distinguishing, highly replicable, and are available in public data sources; they are thus considered high-risk information. On the other hand, laboratory results may be very distinguishing, but they are rarely replicable and seldom disclosed in externally available data sources; they are thus considered low-risk information. The three principles are well justified and we develop our approach based on these principles.

Based on these principles, we classify the information in medical documents into three classes 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 reidentification, we apply the cluster-level value-enumeration method to anonymize QID values. (c) *Health and medical details* (HMDs), such as symptoms, test results, disease, and medications. We will follow the common practice to keep HMDs unchanged because they are of low replicability and availability, but are critical for clinical analysis and healthcare research. In our framework, HMDs are the basis for clustering medical documents. The proposed framework, which we call DAST (*D*e-identification and *A*nonymization for *S*haring medical *T*exts), consists of three functional modules as shown in Figure 5.

4.1. Information Extraction

Module 1 of DAST reads the original documents and 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 elements defined by HIPAA but also additional QIDs not covered by HIPAA (such as patients' age, marital status, ethnicity), as well as HMDs such as patients' 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 extracted HMDs serve as input for Module 2 to perform document clustering. Therefore, the effectiveness of this module is critical since the performances of the other modules depend on it. Figure 6 shows the design of Module 1, which contains three components: (1.1) feature extractor, (1.2) base classifiers, and (1.3) result aggregator.

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 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 four classes: EID, QID, HMD, or irrelevant. The results of the base classifiers are then fed into component 1.3 to produce the final combined results. For example, in the combined result for record 2 in Figure 2, "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.

The result aggregator (component 1.3) uses an ensemble approach to reach the final classification decision. An ensemble classifier consolidates multiple base classifiers to obtain a better predictive performance than those obtained from the base classifiers individually. In term-classification problems, two measures are typically used for evaluating the performance of the classifiers: (a) *recall*, which is the proportion of relevant terms that are retrieved, and (b) *precision*, which is the proportion of retrieved terms that are relevant. In many applications, precision tends to be more important than recall. For classifying PHI terms, however, it is critical to understand that recall is far more important than precision since any PHI term not detected may cause identity disclosure. This aspect has been overlooked by the existing approaches for PHI term detection. Our ensemble approach takes this priority for recall into account when aggregating the results.

More specifically, the result aggregator uses a "set-union" method to combine result sets from the base classifiers to improve the performance in recall. For example, if the PATIENT set generated by base classifier 1 is {Mrs. Brown: PATIENT; Mrs. White: PATIENT} and that by base classifier 2 is {Mrs. Black: PATIENT; Mrs. White: PATIENT}, respectively, then the result aggregator will generate a PATIENT set that is the union of the two base PATIENT sets {Mrs. Brown: PATIENT; Mrs. White: PATIENT; Mrs. Black: PATIENT}. In this way, the result aggregator will find every term that is identified as PATIENT by any base classifier, resulting in high recall.

When there are conflicts between base classifiers, the result aggregator resolves the conflicts based on disclosure risk priority. If a term is recognized as an EID by any classifier, it will be classified as an EID even though it is recognized as a QID or an HMD by the other classifiers. For example, if a base classifier recognizes "white" as a PATIENT name and the other base classifiers consider it as a RACE, "white" will be classified as a PATIENT and removed from the anonymized text. This enables maximum protection for the EID attributes. Similarly, if a term is classified as a QID by one classifier but as an HMD by the other classifiers, it will be classified as a QID because a QID value will be subject to value-enumeration while an HMD value will remain unchanged in the anonymized text.

In this module, the most time-consuming tasks are the training of base classifiers. The time complexities for training SVM and CRF classifiers are quadratic in the size of the training set (Platt 1998, Sutton and McCallum 2012). We note that classifier training is an off-line process that needs to be performed only once. When these classifiers are used for de-identification of medical documents, their runtime performance is linear in size.

4.2. Document Clustering

Clustering-based anonymization approaches are common in data privacy research (Li and Sarkar 2013). These approaches typically cluster the records based on QID attributes, such as age and location, and then anonymize the QID values. As explained earlier, clustering by QID attributes may not serve well for the purposes of medical document sharing, where data users are typically more interested in patients' health and medical information such as symptoms, diagnosis, and treatments. To address this problem, we propose to cluster the documents based on the HMD.

Module 2 of the proposed DAST framework clusters patient documents based on the HMD terms extracted from Module 1. There exists 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 purposes. 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, causing significant information loss; whereas with an overly small size, too few QID values would be listed, resulting in 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 state-of-the-art document clustering technique called nonnegative matrix factorization (NMF) (Lee and Seung 1999). Given a collection of n documents with a total of m terms, let $\mathbf{B}$ be an $m \times n$ term-document matrix. The NMF factorizes $\mathbf{B}$ into two nonnegative matrices, $\mathbf{W} = [w_{jp}]$ ($j = 1, \ldots, m$; $p = 1, \ldots, c$) and $\mathbf{H} = [h_{qi}]$ ($q = 1, \ldots, c$; $i = 1, \ldots, n$), such that $\mathbf{B} \approx \mathbf{WH}$, where $c \ll \min(m, n)$ is a prespecified parameter, representing the number of clusters. The NMF problem can be formulated as

$$\begin{aligned}&\underset{\mathbf{W},\mathbf{H}}{\text{minimize}} \|\mathbf{B} - \mathbf{WH}\|_F^2 \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n}\left(\mathbf{B}_{ji} - (\mathbf{WH})_{ji}\right)^2 \text{subject to } \mathbf{W} \text{ and } \mathbf{H} \text{ are non} \\ &-\text{negetive; i.e., } w_{jp} \geq 0, h_{qi} \geq 0, \forall j, p, q, i. \end{aligned} \quad (1)$$

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 cluster or concept, with each element w_{jp} indicating the degree to which term j belongs to cluster p. Matrix $\mathbf{H}$ describes the contribution of the documents to these concepts or clusters, with each element h_{qi}

representing the degree to which document i 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 nonnegative, which is consistent with the nonnegativity nature of the original term-document matrix. In the context of medical document clustering, a cluster may represent an observable or latent medical concept such as a disease or a condition; a cluster may also contain a mixture of several similar medical concepts. Some clusters may even show similar patient demographics related to certain diseases or conditions. These medical concepts are not specified a priori; instead, they are "learned" by the NMF clustering algorithm.

It is computationally prohibitive to find a global optimal solution for the NMF model (1) (Lee and Seung 2001). Lee and Seung (2001) and Xu et al. (2003) have developed efficient NMF algorithms for finding near optimal solutions, which run either faster than or comparable to the other popular algorithms. They have shown that their algorithms are nondecreasing and converging in objective function (1). Studies have shown that the NMF-based methods outperform, in terms of clustering quality, many other well-known 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, however, cannot be used directly for our clustering purpose because it requires the number of clusters, c, as an input and lacks a mechanism to control the size of a cluster. In our clustering task, the number of clusters cannot be prespecified because each cluster must satisfy the requirement on the minimum number of records in a cluster. To address this issue, we propose a recursive NMF procedure, as a component of Module 2. Essentially, this procedure performs a recursive binary partitioning of the data, using NMF for each binary split, until each group cannot be further partitioned because of the minimum group size requirement. A sketch of the recursive NMF algorithm is given in Figure 7. It is easy to show that the recursive NMF algorithm results in clusters that have low within-cluster variation and high between-cluster variation because each NMF in the recursive process has this property.

The computational complexity of the traditional NMF algorithm is of $O(cnt)$, 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 $O(cnt)$.

To better understand how the proposed clustering method works, we provide a clustering example in Figure 8, which is based on a real data set used in the experiment (the medication data described in Section 5). The tree-structured diagram illustrates the recursive partitioning process and each node in the tree represents a cluster (partitioned subset) of documents. We show in each cluster the terms that describe the principal diagnosis of the case; these terms appear frequently in multiple documents within the same cluster. It is clear that for each cluster there is more or less an underlying medical concept. The concepts at higher levels are more general and they become more specific at lower levels. For example, the first split divides the entire data set into two clusters, one largely representing nonchronic conditions

while the other representing mostly chronic diseases. For the four example final clusters (leaf nodes), the first cluster includes bodily injury cases; the second contains mostly childbirth cases; the third appears to be a cluster of gynecologic cancer cases; and the fourth include those with cardiovascular diseases.

4.3. De-Identification and Anonymization with Value Enumeration and Drill-Down

Data privacy studies in the health domain typically focus on the risk of identity disclosure, which is consistent with the privacy principles underlying HIPAA. Similarly, this study also focuses on identity disclosure. That is, we consider limiting the risk of reidentifying an individual from the released data. A medical document typically contains rich information about HMDs such as symptoms, test results, and diagnosis. It is virtually impossible to protect against attribute disclosure; i.e., to develop a mechanism that masks HMDs. This perhaps explains why HIPAA focuses exclusively on the reidentification issue.

When illustrating the value-enumeration method in Figure 3, we have assumed that the entire data set contains only those five clinical notes. The reidentification risks mentioned above is also based on this assumption. When the data set includes more than those five clinical notes, the actual reidentification risk can be lower because there may be additional clinical notes in the data set that have the same attribute values as those enumerated in Figure 3. To further examine the reidentification risk in a general setting, we first define some terms.

Definition 1—A *grain value* of a QID attribute refers to a value of the QID attribute to be used for value enumeration.

We also use the term “grain” to describe the level of detail for a QID attribute in value enumeration. Information contained in a grain value is typically more detailed than that allowed by the HIPAA SH rule or that generated by traditional *k*-anonymity approaches. For example, in Figure 3 the grain is “month-year” for the Visit Date attribute and hospital name for the Hospital attribute (and it is the age itself for the Age attribute). They are more detailed than the corresponding attribute values in the right panel of Figure 2 (HIPAA Safe Harbor) and in Figure 4 (*k*-anonymity). Like generalization hierarchy in *k*-anonymity, the grain is determined by the data owner.

Definition 2—Given d QID attributes denoted by A_j (j = 1, …, d), let V_j (j = 1, …, d) be the set of distinct grain values of the jth QID attribute. Let $v_{a_j}(a_j=1,\ldots,|V_j|)$ be a distinct grain value in V_j. A *profile* refers to a possible combination of d grain values from the d QID attributes, expressed as $\{A_1=v_{a_1},\ldots,A_d=v_{a_d}\}$. In other words, a profile is an element of the Cartesian product $V_1\times\cdots\times V_d=\left\{(v_{a_1},\ldots,v_{a_d})|v_{a_j}\in V_j,j=1,\ldots,d\right\}$.

For instance, one profile (possible combination of the QID attribute values) for record 1 in Figure 3 is {Visit Date = “Mar-2009,” Hospital = “Mass General Hosp,” Age = 5}. Note that the number of distinct values of Visit Date in Figure 3 is four, while this number is five originally in Figure 2.

Definition 3—A cluster of records with value-enumerated QID attributes are said to satisfy *cluster-level k-safety* if the cluster has at least k records and all records within the cluster have the same set of profiles.

In Figure 3, for example, records 1–3 satisfy cluster-level 3-safety, and records 4 and 5 satisfy cluster-level 2-safety. With appropriately specified grains, the document clustering algorithm described in the previous section can always result in a set of records that satisfy the cluster-level k-safety requirements. It is easy to see that, like k-anonymity, the maximum reidentification risk for any individual record in a data set satisfying k-safety is $1/k$.

As discussed earlier, the cluster-level k-safety may be too conservative in estimating reidentification risk. To illustrate, consider Figure 9, which shows a different scenario of value-enumeration for the example data in Figure 2. In this scenario, the values of the Visit Date and Age attributes are enumerated in the same way as in Figure 3, but the values of the Hospital attribute are listed with a single hospital name for each record. This data set provides more detailed information about patients than that in Figure 3. Since the hospital name is provided for a patient, if a doctor from another organization receiving this anonymized data is interested in getting more detailed information about the patient, the doctor can contact the hospital and request an approval (e.g., patient consent) to access the original record. The value-enumerated records in Figure 9, however, do not satisfy the cluster-level k-safety requirements. If the entire data set contains only these five records, then they all have unique hospital names and thus have high reidentification risks. Suppose, however, these five records are included in a data set of a few thousand records, all of which came from the five hospitals shown in the example. Then, it is likely that some other records in the data set have the same profile as these records, which would lower the reidentification risks for these records. Based on this observation, it is necessary to define reidentification risk at the dataset-level.

Definition 4—A set of records with value-enumerated QID attributes are said to satisfy *dataset-level k-safety* if for each possible profile associated with a record there exist at least $k - 1$ records in the data set that contain the same profile.

Obviously, cluster-level k-safety is more conservative in considering reidentification risk than dataset-level k-safety. We point out that this difference in reidentification risk between cluster-level and dataset-level arises in our approach because our clustering method groups the records based on the HMD attributes, not on the QID attributes. In traditional k-anonymity approaches (e.g., Sweeney 2002, Gardner and Xiong 2009), grouping of records into clusters are based on the QID attributes. As a result, the QID values in different clusters are mutually exclusive in general. In this situation, the difference in reidentification risks between cluster-level and dataset-level is relatively insignificant.

Suppose the records in Figure 9 all satisfy the dataset-level k-safety. Then, it is natural to further examine if it is possible to provide more detailed values for the other QID attributes while satisfying dataset-level k-safety. Figure 10 shows a scenario, where the value of the Visit Date attribute for each record is listed with a single month-year value. Suppose records 4 and 5 satisfy the dataset-level k-safety while records 1–3 do not. Then records 1–3 can be

released as they appear in Figure 9, while records 4 and 5 can be released as they appear in Figure 10. It is not difficult to see that this process of finding more specific value-enumeration representation can be continued as long as the dataset-level *k*-safety is satisfied. We call the process of replacing a list of enumerated values with a single grain value for a QID attribute A_j a *drill down* on A_j. The drill-down process starts from a set of cluster-level *k*-anonymized records and continues until no attribute value can be displayed in a more detailed manner without violating the dataset-level *k*-safety requirements.

Given a set of value-enumerated data with G clusters that satisfy cluster-level *k*-anonymity, let $m_j^g=(j=1,\ldots,d;g=1,\ldots,G)$ be the number of enumerated values of the jth QID attribute for each record in group g. Then, the number of profiles for each record in group g is $\prod_{j=1}^{d} m_j^g$. We call this quantity the *profile size* of group g. That is, the profile size of a group refers to the number of possible combinations of the enumerated values of all QID attributes in the group. For example, the profile size for each record in the first group in Figure 3 is 2 × 2 × 3 = 12, and it is 6 and 3 in Figures 9 and 10, respectively. The profile size can be considered as a measure of information loss due to value enumeration. A small profile size suggests a small information loss, thus it is desirable. Therefore, it is natural to set the objective of the drill-down process to minimizing the total profile size $\sum_{g=1}^{G}\prod_{j=1}^{d} m_j^g$. On the other hand, the dataset-level *k*-safety requirements must be satisfied.

Let M be the total number of possible profiles in the entire data set whose QID attribute values are value enumerated. Let $\{v_{a_1},\ldots,v_{a_d}\}_i(i=1,\ldots,M)$ be a profile and $|\{v_{a_1},\ldots,v_{a_d}\}_i|$ be the number of times this profile appears in the data set. Then, given a set of data with G clusters that satisfy cluster-level *k*-safety, the drill-down problem can be formulated as

$$\text{minimize} \sum_{g=1}^{G}\prod_{j=1}^{d} m_j^g \quad (2a)$$

$$\text{subject to } |\{v_{a_1},\ldots,v_{a_d}\}_i| \geq k, \ \ \forall i. \quad (2b)$$

As explained earlier, each drill-down on an attribute causes the objective value (2a) to decrease. However, finding a global optimal solution for this optimization problem is computationally expensive because the process involves computation of all possible combinations of the attributes to drill down. Our strategy for an efficient solution is to select one QID attribute at a time for drill down. The criterion is to select the attribute having the smallest number of distinct grain values in the data set. This allows the drill-down process to be performed more gradually and the dataset-level *k*-safety constraint (2b) to be satisfied more easily at each step. The detailed drill-down value-enumeration algorithm is given in Figure 11.

The most time-consuming component of the drill-down algorithm is the computation of the profile count $|\{v_{a_1}, \ldots v_{a_d}\}_i|$. If this is a concern, the Apriori-algorithm (Agrawal and Srikant 1994) can be used to dynamically eliminate any profile having a subset of the profile (i.e., combination of the grain values of fewer than d QID attributes) whose frequency count is smaller than k.

4.4. Data Utility Analysis

We have emphasized earlier that our approach is developed for medical text data sharing applications such as association analysis and medical keyword-based search query that involve both QID and HMD data. To analyze data utility with anonymized data, let X be a set of QID attributes and Y be a set of HMD terms in a document set. For any $x \in X$ and $y \in Y$ (where x may contain one or more QID attribute values and y may contain one or more HMD values), the probability of associating x with y is

$$P(X=x, Y=y)=P(X=x|Y=y)P(Y=y) \text{ or } P(x,y)=P(x|y)P(y), \quad (3)$$

where $P(y)$ can be interpreted as the "support" item-set y in the association rule mining context and as the "recall" for y in the medical keyword-based query context (and recall that our approach is designed with priority to improve this measure). With the original data, the probability in (3) can be estimated as

$$\hat{P}(x,y)=\hat{P}(x|y)\hat{P}(y)=\left(\frac{n_{x|y}}{n_y}\right)\left(\frac{n_y}{n}\right), \quad (4)$$

where n is the total number of records in the "support" context and the number of relevant records in the "recall" context, n_y is the number of records (out of the n records) containing y, and $n_{x|y}$ is the number of records containing x given $Y = y$. For anonymized data, because HMDs are not changed, n_y and n remain unchanged; but $n_{x|y}$ needs to be estimated from the domain of x given y, based on the anonymized data. It is easy to see from the examples in Section 2 that this domain contains only a limited number of discrete values with the value enumeration method, but it is typically a region that covers much more discrete values with the SH or k-anonymity method. In estimating $n_{x|y}$, the smaller the domain size is, the closer the estimate $\hat{n}_{x|y}$ is to $n_{x|y}$. For example, if x refers to QID values that have been drilled down, then $\hat{n}_{x|y}=n_{x|y'}$, since the domain in the anonymized data contains the same values as in the original data.

Let $\hat{P}(x,y)$, $\hat{P}_{\text{SH}}(x,y)$, and $\hat{P}_{\text{VE}}(x,y)$ be the estimated $P(x, y)$ calculated using the original data, the SH-compliant data, and value-enumerated data, respectively. Based on the above discussion, we have the following property.

Proposition 1—If the grain values for value enumeration are more detailed than that allowed by the Safe Harbor rule, then $\hat{P}_{\text{VE}}(x,y)$ is closer to $\hat{P}(x,y)$ than $\hat{P}_{\text{SH}}(x,y)$ that is

$$|\hat{P}_{\mathrm{VE}}(x,y) - \hat{P}(x,y)| < |\hat{P}_{\mathrm{SH}}(x,y) - \hat{P}(x,y)|. \quad (5)$$

Proof—Let $D(x|y)$ be the domain of x given y based on the anonymized data. Without any knowledge about the distribution of x and y, we assume a uniform distribution, which has a probability density of $1/|D(x|y)|$. Then with anonymized data, $\hat{n}_{x|y}=n_{x|y}/|D(x|y)|$. When the grain values are more detailed than that allowed by the SH, $D(x|y)$ associated with value enumeration contains only the enumerated grain values or their possible combinations if x involves multiple attributes, while with SH, it must cover all possible grain values suppressed by SH or all possible combinations of these values. Thus, $\hat{n}_{x|y}$ calculated from value-enumerated data is closer to $n_{x|y}$ than that calculated from SH-compliant data. Since n_y and n in (4) remain unchanged, we have (5).

Likewise, we can show a similar result to Proposition 1 when the SH is replaced by *k*-anonymity with its generalization method. Therefore, the proposed approach enables a more accurate estimation of the association or relationship between QID and HMD information than SH or *k*-anonymity.

5. Experimental Evaluation

We have developed a prototype system based on the proposed DAST framework. The system is implemented in Java and text data is organized using XML. We conduct a set of experiments to compare DAST with the HIPAA SH implementation and the *k*-anonymity method in terms of reidentification risk and data utility. As we reviewed in Section 3, the SH based approach represents the mainstream approach for medical document de-identification and *k*-anonymity is well accepted in the field as well.

We use the real-world data sets provided by the Informatics for Integrated Biology and the Bedside (i2b2) project for this study. The i2b2 project has collected multiple sets of clinical documents from different healthcare organizations and made them available for research. We use three data sets for the experimental evaluation. The first set, which has 889 records, was provided for research on the medication aspect of patient care and thus is called Medication. The second set, called Obesity, includes 1,237 records that are related to obesity disease. The third set contains 871 records, collected from a project initiated by a Veterans Affairs (VA) healthcare organization. So, the data set is named VA. Each record above is a medical document, with the format similar to that in Figure 1. Detailed information about the three data sets can be found on the i2b2 website (www.i2b2.org/NLP/DataSets). The elements of information contained in these text data sets include patient name, admission and discharge date, age, gender, hospital, symptoms, test results, diagnosis, disease, medications, and so on. To protect privacy, all PHI values had already been replaced with reasonable surrogate values by the data providers before the data were released. The surrogate values are realistic semantically and consistent with the original format. For example, “Mr. Anderson” may be replaced by “Mr. Taylor” and “6/21/2003” may be replaced by “9/16/1997.”

5.1. Information Extraction and Classification

We implemented three base classifiers in the system. The first is an SVM-based classifier that relies mostly on the local features to classify a term as part of a PHI or non-PHI category. 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 and record structure information to recognize PHI. On top of these three base classifiers, we implemented a result aggregator using the union and risk priority methods discussed in Section 4.1 to combine the outputs of the three base classifiers and produce the final results. Information extraction performances are evaluated based on three commonly used measures: (i) *recall*, which is the proportion of relevant items that are retrieved, (ii) *precision*, which is the proportion of retrieved items that are relevant, and (iii) *F*-measure, which is the harmonic mean of recall and precision.

Out of the three data sets, Medication is the only one that has all PHI marked in the text for testing extraction performance. In fact, the Medication data came with two sets of records: a training set of 669 records and a testing set of 220 records. We focused on the extraction of three categories in this experiment: Patient Name (EID), Admission Date (QID), and Age (QID). The results of the information extraction on the Medication data are shown in Table 2. Among the three base classifiers, the CRF-based classifier performed the best, with the highest recall, precision, and *F*-measure. The SVM-based classifier was a close second, while the rule-based classifier was the worst. These results are consistent with those found from prior studies (Meystre et al. 2010, Uzuner et al. 2007). The final results generated by the aggregator showed a recall of 90.94%, which is about 4.3 percentage points higher than that of the CRF classifier, and a precision of 89.85%, which is about 1.5 percentage points lower than that of the CRF classifier. We believe this trade-off between recall and precision is favorable because, as explained earlier, recall should be considered more important than precision for PHI detection. The aggregator also has the highest *F*-measure, indicating that it is the best when recall and precision are considered together.

5.2. Privacy Risks

All three data sets 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 analyzing privacy risks for two QID attributes in this experiment: Age and Admission Date. The grains for the two attributes are age in year and admission month-year.

We evaluate privacy disclosure risk using two criteria. The first focuses on the worst-case scenario, measuring the number and proportion of records that can be uniquely reidentified based on the QID attribute values. The second measure is related to the average reidentification risk.

Based on the HIPAA SH 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, however, there are still many unique Age and Age-AdminYear combination values. Table 3 shows the number (and percentage, within the parentheses) of records with these

unique values for each data set, resulting from applying the SH rule, *k*-anonymity, and DAST, respectively. With the data processed under the SH rule, if an adversary knew the age and admission year (or even only age) of some patients in the data sets, it is not difficult for the adversary to reidentify the patients and subsequently reveal all of the information of the patients. With the data anonymized using *k*-anonymity or DAST (with $k > 1$), there is no record that can be uniquely identified by the value of Age or Age-AdminYear combination because of the *k*-anonymity and *k*-safety requirements. Therefore, the proposed value-enumeration approach is safer than the SH rule with respect to protecting against unique identification. The DAST-anonymized data can provide more accurate time period information (month-year) than the SH based data (year only). However, to avoid uniqueness on Age value, DAST needs to enumerate a unique Age value together with at least another Age value, which causes some information loss on the Age values for the related records, as compared to the SH release.

To evaluate average reidentification risk, we first define individual reidentification risk. For a record i, let n_i be the number of the other records in the data set that also contain the QID profile of record i. Clearly, a larger n_i value implies a smaller reidentification risk for record i. Therefore, the reidentification risk for record i is defined by $1/(1 + n_i)$, represented as a percentage. Subsequently, for a data set of N records, the average reidentification risk over all of the records in the data set is

$$\text{Average reidentification risk} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+n_i}. \quad (6)$$

When using *k*-anonymity and DAST, the average reidentification risk varies with different k values. We performed the experiments using two values: $k = 3$ and $k = 6$, which are commonly used in the *k*-anonymity related studies (Sweeney 2002, Machanavajjhala et al. 2006). For the SH based data, the average reidentification risk can be calculated straightforwardly.

The results of average reidentification risk are shown in Table 4. It is observed that DAST outperforms SH in all scenarios, particularly for the cases involving the Age-AdminYear combination. When considering Age only, because most age values in the data sets have frequency counts much larger than the threshold k values, DAST performed drill down for most of the records, displaying age value individually as in SH. Consequently, the drill-down results appear only slightly better than the SH results. However, the average reidentification risks with drill down are always smaller than those with SH because drill down cannot be performed on a few rare age values, which are associated with records of very old patients. These high-risk patients were protected by drill down but not by SH. This explains why in data privacy research and practice, disclosure risk is more often measured using the maximum risk (such as the unique reidentification risk) instead of average risk. In the case of the Age-AdminYear combination, the average reidentification risks with SH are very high because many records in the original data sets have either a unique Age-AdminYear combination value or share the value with very few other records. Drill-down operations cannot be applied to these records because of the *k*-safety requirements. As a result, the

average risks with drill down are much lower than those with SH. It is not surprising that drill-down results in higher average risks than no drill-down because drill-down reveals more detailed QID information on average. However, the maximum reidentification risks, represented by $1/k$ and controllable by the user, are the same with or without drill down. It is also observed that average reidentification risk decrease as k increases, which is expected.

The average reidentification risks associated with DAST are very close to those with k-anonymity, particularly for DAST without drill down. This is understandable because with the same k, the group sizes produced by k-anonymity are very close to those by DAST. With drill down, DAST incurs somewhat higher average reidentification risks than k-anonymity. However, drill-down provides much better data utility than k-anonymity, as demonstrated in the next section.

5.3. Data Utility

Medical text data can be used for knowledge discovery in, for example, disease comorbidity, patient stratification, drug interactions, and clinical outcomes (Jensen et al. 2012, Safran et al. 2007). Two basic techniques underlying these applications are search query and association rule mining (Jensen et al. 2012). In evaluating data utility, therefore, we focus on the effectiveness of the DAST-anonymized data when these two techniques are employed.

As mentioned earlier, health and medical research involving epidemic and infectious disease usually requires season information, which can be found from the admission months of the patients. However, the HIPAA SH rule prohibits the explicit release of admission month, while k-anonymity and value enumeration provide generalized or enumerated but not exact information about admission months. Therefore, it is necessary to evaluate information loss when a query about admission month is performed on the anonymized data using the SH, k-anonymity, and our proposed approaches.

We ran a set of queries on each data set to count the number of patients admitted in each month across all years. Because a record anonymized with value enumeration contains multiple month values in a cluster, we randomly choose one of them for that record based on the month value distribution in the cluster. For example, suppose a cluster contains three records with two month values: “Jan” in two records and “Feb” in one record. Then for each record in this cluster, “Jan” will have a 2/3 chance of being selected and assigned to the record and “Feb” will have a 1/3 chance. There are no month values for the SH based data. So, the month values in the data were generated randomly using the uniform distribution. Similarly, the month values for k-anonymity were generated using the uniform distribution based on the available month values within each group.

For each data set, a counting query for each month was run on the original data, and the data anonymized by the SH, k-anonymity, and value enumeration, respectively. We then compared the counts from an anonymized data set with those from the original data based on the average error measure below

$$\text{Error rate in month count} = \frac{1}{12}\sum_{t=1}^{12}\frac{|\tilde{C}_t - C_t|}{C_t}, \quad (7)$$

where C_t and $\tilde{C}_t$ are the count of month t from the original data set and the anonymized data set, respectfully. A small error value implies that the count based on the anonymized data is close to that on the original data, which is clearly desirable.

Again, we used $k = 3$ and $k = 6$ for DAST. The experimental procedure was repeated 20 times for each query on each data set considering the effect of the random selection of the month values. The average error rates over the 20 runs are reported in Table 5. It is clear that the average error rates on the data generated by DAST are generally smaller than those by SH or k-anonymity. The differences are statistically significantly at $\alpha = 0.01$ for all pairwise comparisons involving DAST except one case (for the Obesity data where $k = 3$ between k-anonymity and DAST without drill down). The DAST causes smaller error rates than SH and k-anonymity because month value distribution with DAST are closer to the original distribution than those with SH and k-anonymity. The results also show that, for the same k value, the error rate with drill down is smaller than that without drill down in each case, suggesting that the data quality associated with dataset-level k-safety is better than that with cluster-level k-safety. In addition, it can be observed that as k increases, the error rate associated with DAST increases. This is expected because a larger k value implies a greater profile size, resulting in more enumerated month values for a record. On the other hand, a larger k value implies a higher privacy protection level.

In terms of association rule mining, we evaluate the performance in data utility in the context of large itemset mining. This is a realistic context because many medical document sharing applications can be viewed as finding large itemsets and associations in the documents. We focus on mining associations between medical information and two QID attributes: Admission Date and Hospital. The grains for the two PHI attributes are admission month-year and hospital name, respectively, which are important season and location information that is not available if the HIPAA SH rule is strictly followed.

To facilitate comparisons, we converted the document data sets into medical-term-association tables. Each row of the table corresponds to a document in the original data set. Each column represents a medical term extracted from Module 1 of DAST and there are a few hundred columns for each data set. These medial terms are related to the medical concepts and topics in the data set. Additional columns were added to the table to represent the admission month and hospital. Similarly, we converted the data sets anonymized by SH, k-anonymity, and DAST into respective medical-term-association tables. The method to determine month values for the counting queries described earlier was also used in creating the tables.

For each data set, an association rule mining algorithm was run on the original table, and the tables based on the data anonymized by SH, k-anonymity, and DAST, respectively. We then compared the large itemsets discovered from an anonymized table with those from the

original table based on an error measure defined below (Aggarwal and Yu 2008, Rizvi and Haritsa 2002)

$$\text{Error rate in support count} = \frac{1}{|L|}\sum_{\ell \in L}\frac{|\tilde{F}_\ell - F_\ell|}{F_\ell}, \quad (8)$$

where L represents the set of all large itemsets with a support count larger than the minimum support threshold (which was set to 10 in this experiment); F_ℓ is the frequency count of the ℓth large itemset from the original data set; and $\tilde{F}_\ell$ is the corresponding count from the anonymized data set. Equation (8) measures the proportion of the large itemsets in the original data set that are correctly found in the anonymized data set. Clearly, a small error rate is desirable.

Considering the effect of the random selection of the QID attribute values, the experimental procedure was repeated 10 times on each data set. The average error rates over the 10 runs are reported in Table 6. It is clear that DAST outperforms the SH and *k*-anonymity approaches substantially in terms of large itemset mining. The differences are statistically significantly at $\alpha = 0.001$ for all pairwise comparisons involving DAST. Therefore, our proposed approach preserves data utility significantly better than the SH and *k*-anonymity approaches. Similar to the counting query case, for the same *k* value, the error rate with drill down is smaller than that without drill down in each case. Note that because hospital names in the original data were replaced by the data providers with surrogate names, which no longer have location meaning, generalization used by *k*-anonymity cannot be implemented for counting itemsets containing hospital names. So, the error rate for *k*-anonymity in Table 6 is calculated based on the large itemsets without hospital names.

In addition to counting query and association rules mining, we performed a third experiment, where anonymized medical documents would be used for keyword-search-based information retrieval. Our experiment focused on searches related to medical terms (HMD) and hospital names (QID). We wrote a program that allows HMD terms to be used for search queries on the document collection and retrieves a set of hospital names associated with the HMD terms. For example, for a query that contains two keywords, “diabetes” and “glucose,” the program will return a set of hospital names, indicating that these two words appear in a record that contains one or more of these hospitals.

We first composed our keyword list using all of the medical terms extracted by our HMD extractor. We then removed the terms that appear in more than 25% of the total records (these high frequency terms are not very useful to characterize medical concepts in a document). The result is a list of 387 terms that are highly relevant to the medical concepts conveyed in the text data. These terms are rare enough to represent the differences between individual records. We used all single terms and all possible two-term combinations as query terms in our experiment. This resulted in a total of $387 + \binom{387}{2} = 75,078$ queries, which

were used for retrieving hospital names from the original data sets, as well as anonymized data sets. Queries that returned empty results were disregarded.

The performance of the search results, which we call *search query score*, is measured based on the well-known Jaccard similarity coefficient

$$\text{Search query score} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{|\tilde{H}_q \cap H_q|}{|\tilde{H}_q \cup H_q|}, \quad (9)$$

where Q is the number of queries, and H_q and $\tilde{H}_q$ are the set of hospital names retrieved by the qth query from the original data set and anonymized data set, respectively. The value range of this measure is [0, 1]. A larger value indicates that the results from the anonymized data set are similar to those from the original set, which suggests a better performance. Safe Harbor does not allow releasing hospital names, so the search query score for Safe Harbor on any data set is zero. To make comparisons somewhat meaningful, we also randomly generated the hospital values for Safe Harbor based on the hospital value distribution in the original data set (by assuming that the distribution information was provided to the data user). This random selection process was repeated 10 times and the average score value over the 10 runs is reported. Because surrogate hospital names in the data do not have location meaning, generalization used by k-anonymity cannot be implemented. So, we did not include k-anonymity in this experiment.

The search query score results are summarized in Table 7, where under Safe Harbor a score of zero indicates no hospital is found, while the results inside the parentheses were computed by assuming the hospital value distribution is known (which is unlike month values where random guessing is always possible). Clearly, DAST outperforms Safe Harbor substantially in this task. The results are statistically significantly different at $\alpha = 0.001$ for all pairwise comparisons. Therefore, our proposed approach achieves better data utility significantly than the SH approach. Overall, the score values with DAST appear to still be low, which is expected because enumerated values often include both the original and irrelevant hospital names for each record. Safe Harbor essentially cannot provide hospital names for any medical-term based search query. Again, the drill-down operations improve the quality of search results. Also, as k increases, the score associated with DAST decreases.

The search query score in Equation (9) is a measure of accuracy for an individual search query. For medical research, the shared data are more likely to be used for finding patterns and relationships that are characterized at the data set level (rather than individual information retrieval). To evaluate the effectiveness of the proposed approach in this aspect, we use the following statistic to measure data utility in preserving the relationship between the search query term and the hospital:

$$\text{Error rate in hospital count for given keywords} = \frac{1}{|H|}\sum_{h\in H}\frac{|\tilde{n}_{h|w} - n_{h|w}|}{n_{h|w}}, \quad (10)$$

where H represents the set of all hospitals retrieved; $n_{h|w}$ is the number of times hospital h being retrieved by search keywords w from the original data set; and $\tilde{n}_{h|w}$ is the corresponding number from the anonymized data set. Equation (10) measures the proportion of the hospitals retrieved by given search keywords from the original data set that are correctly retrieved from the anonymized data set.

Again, we applied the distribution-based method described earlier to determine the retrieved hospital. So, the experimental procedure was repeated 10 times on each data set. The average error rates over the 10 runs are reported in Table 8. It is clear that DAST outperforms SH substantially. The differences are statistically significantly at $\alpha = 0.001$ for all pairwise comparisons. Therefore, our proposed approach preserves the relationship between the search query term and the hospital significantly better than SH.

6. Discussion

Experimental results from the real-world data in Section 5.2 provide clear evidence that the HIPAA SH approach for de-identifying data can be underprotective (i.e., de-identified data having high disclosure risk). On the other hand, the results in Section 5.3 show that the SH approach can also be overprotective (i.e., resulting in poor data utility). These findings have significant research and practical implications, since the mainstream approaches in both existing research and current practice follow the SH rule to de-identify data for medical document sharing.

To overcome these weaknesses in the existing approaches, we propose clustering text documents based on health and medical information, using a recursive binary partitioning algorithm. The clustered data are then anonymized using a novel value-enumeration method. We introduce the notion of cluster-level and dataset-level anonymity for value enumeration and develop a drill-down method to further reduce information loss in the anonymized data. The results of the experimental study demonstrate that the anonymized data generated by the proposed approach have lower disclosure risk and better data utility than those generated by the SH based existing approach.

We have so far focused on anonymizing and sharing medical text data without considering the related structured data. It is fairly straightforward to apply our approach to anonymize the structured data associated with the text. When the text documents are clustered, the structured record will go together with its associated text document. The structured data can also be classified into three categories: EID, QID, and HMD. So, the same treatments for the three categories as for the text data can be applied; that is, EIDs will be removed, QID values will be enumerated, and HMDs will remain unchanged.

When using the proposed technique in practice, the user needs to determine the minimum group size. In general, the smaller the group size, the closer the anonymized record to the original record, and the higher the reidentification risk. In the clustering-based anonymization approach, it is common to use a group size of between 3 and 10 records (Sweeney 2002, Machanavajjhala et al. 2006, Li et al. 2007). Our experiments have demonstrated that a group size within this range is effective in limiting reidentification risk

and preserving data utility. Another important decision for the user is to choose appropriate grain values for QID attributes. In principle, information contained in a grain value should be more (or no less) detailed than that allowed by the HIPAA SH rule or that generated by traditional *k*-anonymity approaches. For admission/visit date, we suggest using month-year or season-year as the grain value to provide season information. For date of birth, we believe year of birth (age) should be detailed enough. For location attributes, we suggest using a five-digit zip code and/or hospital name as the grain value. In general, the gain value depends more on the information need and data utility consideration rather than on the disclosure risk concern because the latter can be addressed by adjusting the minimum group size.

We have used three real-world data sets for experimental evaluation. As we mentioned earlier, to protect privacy, all PHI values in the data sets had been replaced with reasonable surrogate values by the data providers before the data were released. However, it is possible that a true value was replaced with an inappropriate surrogate. Because of privacy concerns, it is also likely that the data provider acted overly conservatively in processing the original data. In these cases, the results on the processed data may be more or less different from those on the original data (although it can be expected that the effect would be similar for all techniques under evaluation). Ideally, the original data should be used whenever possible.

Another limitation of this study is that the proposed framework is designed for sharing a set of text documents (instead of a single record) for the purposes of medical research, public health reporting, and health-care management study. This is the same setting as that assumed in data privacy literature on the structured data (Aggarwal and Yu 2008, Sweeney 2002). When the purpose of sharing medical documents is to study an individual case, the proposed approach may or may not be applicable, depending on whether a set of medical documents is available together with the individual document.

7. Conclusions and Extensions

In this study, we investigate the issues related to privacy protection in sharing medical documents. We propose a novel privacy protection framework that integrates and improves techniques developed in the data privacy and health informatics fields to reduce privacy disclosure risks in sharing medical documents while preserving the utility of the data. We demonstrate the validity and effectiveness of the proposed framework by developing and evaluating a prototype system based on the proposed framework. Our experiments show that the proposed approach significantly outperforms the HIPAA SH approach. Therefore, our proposed approach may be a promising alternative to the HIPAA SH rule for sharing patient medical documents with appropriate privacy protection.

As medical data sharing is being increasingly considered in practice, there is a rising concern that patient privacy is being compromised. The proposed framework will reduce the risks of reidentification of individuals from anonymized data, while improving the utility of the data. This should alleviate patients' concerns about loss of privacy and confidentiality and increase their willingness to participate in research that uses patient data. The proposed approach will also reduce organizations' concerns about potential privacy violations and

enable organizations to safely share and publish high-quality data for legitimate research and analysis.

Future research can be pursued along several lines. In the proposed value-enumeration approach, the values of a QID attribute are shown either with a list of all distinct values within the cluster or with a specific grain value after drill down. Alternatively, we can consider providing a subset of all distinct values for a QID attribute, which could be implemented using a divide-and-conquer approach. Another possible extension of this work is to apply the idea of privacy by design or value sensitive design (Cooper and Collman 2005) to the value-enumeration scheme and to allow individual patients to determine the grain details that can be released. In addition, it is worthwhile to explore more effective ways to further improve the performance of machine learning and document clustering techniques in our framework.

This work studies the problem where each individual has only one record. It is also interesting and challenging to examine the privacy problem where an individual may have multiple records in the data set. This is an understudied problem in data privacy research even for structured data. When an individual has multiple records in a data set, it is likely that the grain values of some or all QID attributes will be the same for these records. We believe the proposed *k*-safety principle may be extended to consider the number of distinct grain values to deal with the problem of one individual with multiple records. This topic deserves an in-depth investigation in future research.

Acknowledgments

The authors are grateful to the senior editor, associate editor, and three anonymous reviewers for their insightful comments and suggestions that have improved the paper considerably. The authors thank Dr. Anna Rumshisky for sharing the source code for a program to extract PHI terms from clinical notes.

Funding: X. Li's research was supported in part by the National Library of Medicine of the National Institutes of Health under [Grant R01LM010942]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

References

Aggarwal, CC., Yu, PS. Privacy-Preserving Data Mining: Models and Algorithms. Springer; New York: 2008.

Agrawal, R., Srikant, R. Proc 20th Internat Conf Very Large Databases. Morgan Kaufmann; San Francisco: 1994. Fast algorithms for mining association rules in large databases; p. 487-499.

Agrawal, R., Srikant, R. Proc 2000 ACM SIGMOD Internat Conf Management Data. ACM; New York: 2000. Privacy-preserving data mining; p. 439-450.

Carpineto C, Osinski S, Romano G, Weiss D. A survey of Web clustering engines. ACM Comput Surveys. 2009; 41(3) Article 17.

Carter, JH. What is the electronic health record?. In: Carter, JH., editor. Electronic Health Records: A Guide for Clinicians and Administrators. 2nd. ACP Press; Philadelphia: 2008. p. 3-20.

Cooper, T., Collman, J. Managing information security and privacy in healthcare data mining: State of the art. In: Chen, H.Fuller, SS.Friedman, C., Hersh, W., editors. Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Springer; New York: 2005. p. 95-137.

Cortes C, Vapnik V. Support-vector networks. Machine Learn. 1995; 20(3):273–297.

Dalenius T, Reiss SP. Data swapping: A technique for disclosure control. J Statist Planning Inference. 1982; 6(1):73–85.

Department of Health and Human Services (DHHS). Standards for privacy of individually identifiable health information. Federal Register. 2000; 65(250):82462–82829. [PubMed: 11503738]

Duncan GT, Lambert D. The risk of disclosure for microdata. J Bus Econom Statist. 1989; 7(2):201–217.

Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Amer Medical Informatics Assoc. 2008; 15(5):601–610.

Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: A survey of recent developments. ACM Comput Surveys. 2010; 42(4) Article 14.

Gardner J, Xiong L. An integrated framework for de-identifying unstructured medical data. Data Knowledge Engrg. 2009; 68(12):1441–1451.

Garfinkel R, Gopal R, Thompson S. Releasing individually identifiable microdata with privacy protection against stochastic threat: An application to health information. Inform Systems Res. 2007; 18(1):23–41.

Health Information Technology for Economic and Clinical Health Act (HITECH Act). Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA) (Pub. L. 111-5). 2009. https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf

Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. Nature Rev Genetics. 2012; 13(6):395–405. [PubMed: 22549152]

Kuang, D., Ding, C., Park, H. Symmetric nonnegative matrix factorization for graph clustering; Proc. 12th SIAM Internat. Conf. Data Mining; SIAM, Philadelphia. 2012. p. 106-117.

Lafferty, J., McCallum, A., Pereira, F. Proc 18th Internat Conf Machine Learn. Morgan Kaufmann; San Francisco: 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data; p. 282-289.

Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401(6755):788–791. [PubMed: 10548103]

Lee, DD., Seung, HS. Algorithms for non-negative matrix factorization. In: Dietterich, TG., Tresp, V., editors. Advances in Neural Information Processing Systems. Vol. 13. MIT Press; Cambridge, MA: 2001. p. 556-562.

Li, N., Li, T., Venkatasubramanian, S. Proc 23rd IEEE Internat Conf Data Engrg. IEEE Computer Society; Washington, DC: 2007. t-Closeness: Privacy beyond k-anonymity and l-diversity; p. 106-115.

X-B, Li, Sarkar, S. Protecting privacy against record linkage disclosure: A bounded swapping approach for numeric data. Inform Systems Res. 2011; 22(4):774–789.

X-B, Li, Sarkar, S. Class-restricted clustering and microperturbation for data privacy. Management Sci. 2013; 59(4):796–812.

Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M. Proc 22nd IEEE Internat Conf Data Engrg. IEEE Computer Society; Washington, DC: 2006. l-Diversity: Privacy beyond k-anonymity; p. 24-35.

Melville N, McQuaid M. Generating shareable statistical databases for business value: Multiple imputation with multimodal perturbation. Inform Systems Res. 2012; 23(2):559–574.

Menon S, Sarkar S, Mukherjee S. Maximizing accuracy of shared databases when concealing sensitive patterns. Inform Systems Res. 2005; 16(3):256–270.

Meystre, SM., Savova, GK., Kipper-Schuler, KC., Hurdle, JF. Extracting information from textual documents in the electronic health record: A review of recent research. In: Geissbuhler, A., Kulikowski, C., editors. IMIA Yearbook of Medical Informatics. Vol. 2008. Schattauer Publishers; Stuttgart, Germany: 2008. p. 128-144.

Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: A review of recent research. BMC Medical Res Methodology. 2010; 10 Article 70.

Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy in i2b2. J Amer Medical Informatics Assoc. 2011; 18(Suppl 1):i103–i108.

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Amer Medical Informatics Assoc. 2010; 17(2):124–130.

Office for Civil Rights (OCR). Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Department of Health and Human Services; Washington, DC: 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#protected

Platt, J. Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B.Burges, C., Smola, AJ., editors. Advances in Kernel Methods—Support Vector Learning. MIT Press; Cambridge, MA: 1998. p. 185-209.

Polikar R. Ensemble based systems in decision making. IEEE Circuits Systems Magazine. 2006; 6(3): 21–45.

Rizvi, SJ., Haritsa, JR. Proc 28th Very Large Data Base Conf. Morgan Kaufmann; San Francisco: 2002. Maintaining data privacy in association rule mining; p. 682-693.

Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Detmer DE. Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper. J Amer Medical Informatics Assoc. 2007; 14(1):1–9.

Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. J Amer Medical Informatics Assoc. 2010; 17(5):507–513.

Sebastiani F. Machine learning in automated text categorization. ACM Comput Surveys. 2002; 34(1): 1–47.

Sutton C, McCallum A. An introduction to conditional random fields. Foundations Trends Machine Learn. 2012; 4(4):267–373.

Sweeney L. *k*-Anonymity: A model for protecting privacy. Internat J Uncertainty, Fuzziness Knowledge-based Systems. 2002; 10(5):557–570.

Tan A, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics. 2003; 2(3 Suppl):S75–S83. [PubMed: 15130820]

Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Amer Medical Informatics Assoc. 2007; 14(5):550–563.

Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, Yeh A, Hitzeman J, Hirschman L. Rapidly retargetable approaches to de-identification in medical records. J Amer Medical Informatics Assoc. 2007; 14(5):564–573.

Wolpert DH. Stacked generalization. Neural Networks. 1992; 5(2):241–259.

Wylie JE, Mineau GP. Biomedical databases: Protecting privacy and promoting research. Trends Biotechnology. 2003; 21(3):113–116.

Xu, W., Liu, X., Gong, Y. Proc 26th Annual Internat ACM SIGIR Conf Res Development Inform Retrieval. ACM; New York: 2003. Document clustering based on non-negative matrix factorization; p. 267-273.

6/7/1999 12:00:00 A.M.
Discharge Summary
Admission Date: 06/07/1999
Discharge Date: 06/13/1999

HISTORY OF PRESENT ILLNESS: Mr. Cornea is a 60 year old male who noted the onset of dark urine during early January. He underwent CT and ERCP at the Lisonatemi Faylandsburgnic Community Hospital with a stent placement and resolution of jaundice. He underwent an ECHO and endoscopy at Ingree and Weamanshy Medical Center on April 28. He was found to have a large, bulging, extrinsic mass in the lesser curvature of his stomach. Fine needle aspiration showed atypical cells, positively reactive mesothelial cells. Abdominal CT on April 14 showed a 12 × 8 × 8 cm mass in the region of the left liver, and appeared to be from the lesser curvature of the stomach or left liver. He denied any nausea, vomiting, anorexia, or weight loss. He states that his color in urine or in stool is now normal. PAST MEDICAL HISTORY: He has hypertension and nephrolithiasis. PAST SURGICAL HISTORY: Status post left kidney stones × 2, and he has had a parathyroid surgery. PHYSICAL EXAMINATION: CHEST: Clear. HEART: Regular. ABDOMINAL INCISION: Clean, dry and intact. No drainage. VITAL SIGNS: He is afebrile and otherwise vital signs are stable. He is having good p.o. intake on present diet and he is moving his bowels. DISPOSITION: He is going to a rehabilitation facility until he is able to live independently. DISCHARGE MEDICATIONS: Same as pre-op, with the addition of Roxicet elixir. Dictated By: THAMETO DOYLE, M.D. OS43 Attending: PRO R. KOTEFOOKSSHUFF, M.D. 06/13/99 KE9 [report_end]

Figure 1.
An Example of a Medical Discharge Summary

Original text	De-identified text (HIPAA safe harbor)
1. Admitted on 4/25/2009, Tufts Med Ctr. The 88 year old man is complaining fever, sore throat, headache, runny nose.	1. Admitted on [2009], [HOSPITAL]. The 88 year old man is complaining fever, sore throat, headache, runny nose.
2. Mrs. Brown is a 52 year old female. Visited on 3/13/2009. Having joint pain, sore throat, fever. . . . Mass General Hosp.	2. [NAME] is a 52 year old female. Visited on [2009]. Having joint pain, sore throat, fever. . . . [HOSPITAL].
3. Admitted on 4-5-2009, patient is a 5 year old female. Having runny nose, headache, vomiting. . . . Emory Univ. Hosp.	3. Admitted on [2009], patient is a 5 year old female. Having runny nose, headache, vomiting. . . . [HOSPITAL].
4. Mark is a 17 year old male. Having pain on right side of abdomen, fatigue, dark urine. . . . Admitted 8/20/2009, UT Southwestern Med Ctr.	4. [NAME] is a 17 year old male. Having pain on right side of abdomen, fatigue, dark urine. . . . Admitted [2009], [HOSPITAL].
5. Visited on 7/19/2009. Female, 64 year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. . . . Johns Hopkins Hosp.	5. Visited on [2009]. Female, 64 year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. . . . [HOSPITAL].

Figure 2.
An Example of Clinical Text—Original and De-Identified

1. Admitted on {*Mar*-2009, *Apr*-2009}, {*Emory Univ. Hosp, Mass General Hosp, 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. . . . {*Emory Univ. Hosp, Mass General Hosp, Tufts Med Ctr*}.
3. Admitted on {*Mar*-2009, *Apr*-2009}, patient is a {5, 52, 88} year old female. Having runny nose, headache, vomiting. . . . {*Emory Univ. Hosp, Mass General Hosp, Tufts Med Ctr*}.

4. [NAME] is a {17, 64} year old male. Having pain on right side of abdomen, fatigue, dark urine. . . . Admitted {Jul-2009, Aug-2009}, {*Johns Hopkins Hosp, UT Southwestern Med Ctr*}.
5. Visited on {Jul-2009, Aug-2009}. Female, {17, 64} year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. . . . {*Johns Hopkins Hosp, UT Southwestern Med Ctr*}.

Figure 3.
An Example of Clinical Text Anonymized with the Proposed Approach (Scenario 1)

1. Admitted on [*Mar-Jul* 2009], [*Northeastern Region Hosp*]. The [52–88] year old man is complaining fever, sore throat, headache, runny nose.
2. [NAME] is a [52–88] year old female. Visited on [*Mar-Jul* 2009]. Having joint pain, sore throat, fever. . . . [*Northeastern Region Hosp*].
5. Visited on [*Mar-Jul* 2009]. Female, [52–88] year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. . . . [*Northeastern Region Hosp*].

3. Admitted on [*Apr-Aug* 2009], patient is a [5–17] year old female. Having runny nose, headache, vomiting. . . . [*Southern Region Hosp*].
4. [NAME] is a [5–17] year old male. Having pain on right side of abdomen, fatigue, dark urine. . . . Admitted [*Apr-Aug* 2009], [*Southern Region Hosp*].

Figure 4.
An Example of Clinical Text Anonymized with *k*-Anonymity

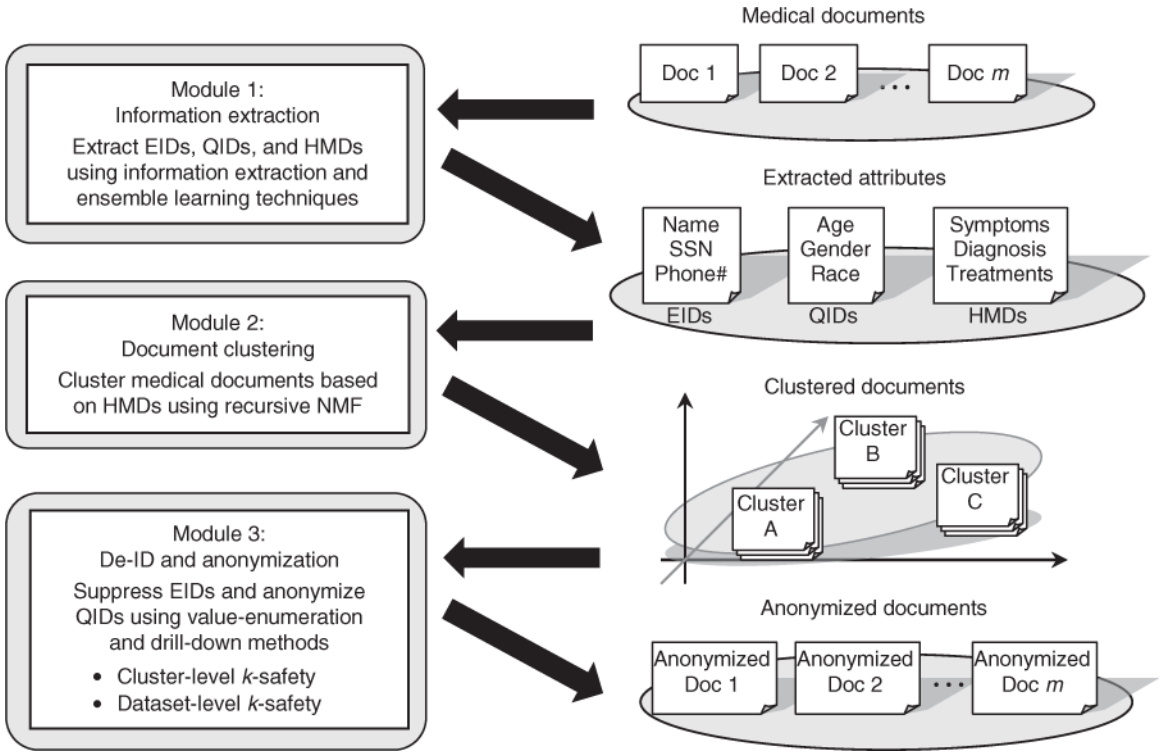

Figure 5.
Proposed DAST Framework Architecture

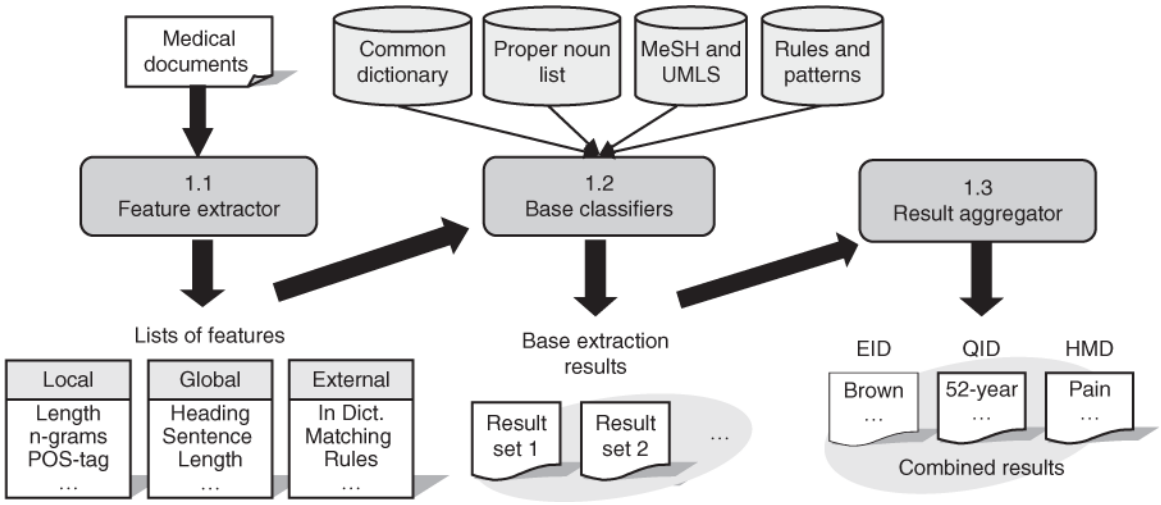

Figure 6.
Design of the Information Extraction Module

1. Given a term-document matrix **B**, partition **B** into two clusters using the NMF by Xu et al. (2003).
2. Partition the subset of the data in each cluster into two subclusters using the NMF again.
3. Repeat Step 2 for each of the two subclusters. Stop the process if the subcluster cannot satisfy the minimum group size requirement.

Figure 7.
Algorithm for Recursive Binary Partitioning with NMF

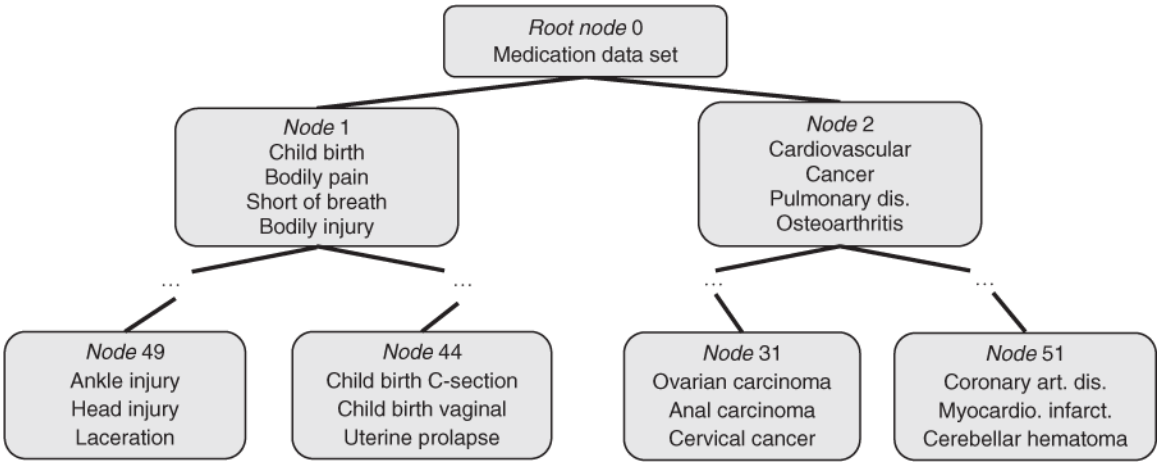

Figure 8.
Example of Document Clustering

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. . . . {*Emory Univ. Hosp*}.

4. [NAME] is a {17, 64} year old male. Having pain on right side of abdomen, fatigue, dark urine. . . . Admitted {*Jul*-2009, *Aug*-2009}, {*UT Southwestern Med Ctr*}.
5. Visited on {*Jul*-2009, *Aug*-2009}. Female, {17, 64} year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. . . . {*Johns Hopkins Hosp*}.

Figure 9.
An Example of Clinical Text Anonymized with Value Enumeration (Scenario 2)

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. . . . {*Emory Univ. Hosp*}.
4. [NAME] is a {17, 64} year old male. Having pain on right side of abdomen, fatigue, dark urine. . . . Admitted {*Aug*-2009}, {*UT Southwestern Med Ctr*}.
5. Visited on {*Jul*-2009}. Female, {17, 64} year old. Feeling abdomen pain, sore muscles, fatigue, jaundice. . . . {*Johns Hopkins Hosp*}.

Figure 10.
An Example of Clinical Text Anonymized with Value Enumeration (Scenario 3)

1. Given a set of data with G clusters that satisfy cluster-level k-safety, for each possible profile $\{v_{a_1}, \ldots, v_{a_d}\}_i$, compute $|\{v_{a_1}, \ldots, v_{a_d}\}_i|$.
2. Find the QID attribute that has the smallest number of distinct grain values in the data set and denote it as attribute j^*.
3. For each record, replace the list of attribute j^* enumerated values with the true grain value. Recompute $|\{v_{a_1}, \ldots, v_{a_d}\}_i|$ under the new value-enumeration scenario. If $|\{v_{a_1}, \ldots, v_{a_d}\}_i| < k$, roll back the attribute j^* value of corresponding records to the list of enumerated grain values.
4. Repeat Steps 2 and 3 for the remaining QID attributes.

Figure 11.
Drill-Down Value-Enumeration Algorithm

Table 1

Protected Health Information Defined by HIPAA

Category	Description
1.	Names
2.	*Locations*: 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 corresponding area contains more than 20,000 people.
3.	*Dates*: (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.
4.	Telephone numbers
5.	Fax numbers
6.	E-mail addresses
7.	Social security numbers
8.	Medical record numbers
9.	Health plan beneficiary numbers
10.	Account numbers
11.	Certificate/license numbers
12.	Vehicle identifiers and serial numbers, including license plate numbers
13.	Device identifiers and serial numbers
14.	Web Universal Resource Locators (URLs)
15.	Internet Protocol (IP) address numbers
16.	Biometric identifiers, including finger and voice prints
17.	Full face photographic images and any comparable images
18.	Any other unique identifying number, characteristic, or code

Source. Adapted from DHHS (2000).

Table 2

Results of Information Extraction Performance

	SVM-based (%)	CRF-based (%)	Rule-based (%)	Aggregator (%)
Recall	82.72	86.64	37.78	90.94
Precision	88.93	91.36	79.36	89.85
F-measure	85.71	88.94	51.19	90.39

Table 3

Results of the Unique Reidentification Risk

Data set	Age only			Age-AdminYear combination		
	Safe harbor	*k*-anonymity	DAST	Safe harbor	*k*-anonymity	DAST
Medication (889 records)	22 (2.5%)	0	0	309 (34.8%)	0	0
Obesity (1,237 records)	12 (1.0%)	0	0	476 (38.5%)	0	0
VA (871 records)	17 (2.0%)	0	0	296 (33.9%)	0	0

Table 4

Results of the Average Reidentification Risk

	Age only (%)							Age-AdminYear combination (%)						
		k = 3			*k* = 6				*k* = 3			*k* = 6		
Data set	**SH**	***k*-anonymity**	**DAST (no drill)**	**DAST (drill)**	***k*-anonymity**	**DAST (no drill)**	**DAST (drill)**	**SH**	***k*-anonymity**	**DAST (no drill)**	**DAST (drill)**	***k*-anonymity**	**DAST (no drill)**	**DAST (drill)**
Medication	8.66	7.23	6.27	8.34	4.19	4.16	8.30	37.12	16.98	14.29	17.97	8.02	7.54	10.68
Obesity	7.84	6.57	5.83	7.78	4.01	3.98	7.71	34.92	15.37	13.02	16.83	7.63	5.79	8.52
VA	8.27	7.49	6.09	8.09	4.12	4.04	8.02	36.58	16.07	13.96	16.23	8.30	7.02	9.86

Table 5

Error Rate for the Counting Query

		***k* = 3**			***k* = 6**		
Data set	**Safe harbor**	***k*-anonymity**	**DAST (no drill)**	**DAST (drill)**	***k*-anonymity**	**DAST (no drill)**	**DAST (drill)**
Medication	0.2171	0.1837	0.1536	0.0683	0.2176	0.1745	0.1173
Obesity	0.2638	0.2131	0.2147	0.1085	0.2507	0.2394	0.1636
VA	0.2269	0.1879	0.1724	0.0954	0.2165	0.1926	0.1249

Table 6

Error Rate for Large Itemsets

		k = 3			*k* = 6		
Data set	**Safe harbor**	***k*-anonymity**	**DAST (no drill)**	**DAST (drill)**	***k*-anonymity**	**DAST (no drill)**	**DAST (drill)**
Medication	0.4569	0.2343	0.1358	0.0713	0.3109	0.1569	0.1096
Obesity	0.4762	0.2775	0.2079	0.1198	0.3441	0.2482	0.1521
VA	0.4381	0.2560	0.1369	0.0804	0.3348	0.1753	0.1042

Note. The results are statistically significantly different at $\alpha = 0.001$ for all pairwise comparisons.

Inf Syst Res. Author manuscript; available in PMC 2018 March 19.

Table 7

Results of the Search Query Score

		DAST (k = 3)		DAST (k = 6)	
	Safe harbor	**No drill down**	**With drill down**	**No drill down**	**With drill down**
Medication	0 (0.0671)	0.1723	0.2114	0.1186	0.1437
Obesity	0 (0.0712)	0.1964	0.2418	0.1256	0.1694
VA	0 (0.0547)	0.1620	0.2031	0.1053	0.1396

Note. The results are statistically significantly different at α = 0.001 for all pairwise comparisons.

Table 8

Error Rate for the Hospital Count with Given Keywords

		DAST (k = 3)		DAST (k = 6)	
	Safe harbor	**No drill down**	**With drill down**	**No drill down**	**With drill down**
Medication	0.6452	0.1173	0.0632	0.1374	0.0694
Obesity	0.7109	0.0904	0.0516	0.1151	0.0613
VA	0.5872	0.1162	0.0589	0.1387	0.0628