Preface

The Health Information Technology for Economic and Clinical Health (HITECH) Act mandates that every American have access to their electronic medical records by 2014. There are several important issues that need to be addressed before the electronic medical records can be beneficial to the individuals as well as effective for the healthcare providers. From an operational data processing standpoint, the EMRs should provide an interconnected system that would be accessible to the various stakeholders in the system of healthcare delivery. From a strategic data usage standpoint, it is important to be able to access, use, and make decisions on the vast amount of information that will be stored in these systems. There are many issues connected to the strategic data usage including protecting privacy of the individual data and analyzing electronic medical records. While there is a lot of research on privacy preserving data mining (e.g. Agrawal et al., 2001, Aggrawal et al., 2008, Kargupta et al., 2003, Liu et al., 2006, Samarati et al., 1998, Sarathy et al., 2011, Szarvas et al., 2007, Vaidya et al., 2006), very few have been used for protecting privacy in electronic medical records. Furthermore, most of the methods are designed for numerical data. Similarly, while there are a lot of methods for pattern recognition and data mining (e.g. Han et al., 2006, Tan et al., 2006), very few methods exist for data visualization, which is both intuitive and appealing to domain experts. In this paper we focus on privacy for text data, analyzing text data, and data visualization as they apply to electronic medical records.

PRIVACY IN ELECTRONIC MEDICAL RECORDS

Privacy issues dominate the landscape of electronic medical records. The Health Insurance Portability and Accountability Act (HIPAA) requires that personally identifiable information of patients be protected against access by non-covered entities. Such entities include medical researchers, administrative staff, and healthcare professionals who are not directly involved in the diagnosis and treatment of the patient. The “safe harbor” de-identification of medical record Information of HIPAA specifies eighteen identifiers of an individual patient, or of the relatives of the patient, and employers or household members of the patient that must removed.

Automatic removal of identifying information is relatively easy in numeric data. However, when it comes to unstructured data such as text, automatic removal of identifying information becomes non-trivial. At the same time textual data abounds in medical records in the forms of progress notes, discharge summaries, laboratory reports, x-ray reports, etc. Moreover, it has been shown by research in privacy preserving data mining that the mere removal of identifying information is no guarantee that the privacy of individual patients will be protected. In this section we first discuss issues related to de-identification of text data in electronic medical records, and then the issues related to privacy protection of medical records in general.
DE-IDENTIFYING MEDICAL TEXT

The elements identified as protected health information (PHI) in electronic medical records include eighteen specific categories of information such as location information in terms of geographic subdivisions smaller than a state; dates for admission, discharge, birth and death that are more detailed than the year, age information including the year if the age is more than 89; contact information such as telephone, fax numbers, and email addresses; Internet-based information such as URL and IP addresses; business related information such as account numbers, health plan beneficiary numbers, medical record numbers, and certificate/license numbers; personal identification such as names, social security numbers, biometric identifiers; and vehicle identifiers such as license plate numbers and vehicle serial numbers. In addition, finger prints, voice prints, photographic facial images, etc. are also considered to be protected health information.

Research on de-identification of free medical text can be divided into the following categories: those based on Natural language processing (e.g., Szarvas et al., 2007, Uzuner et al., 2007, Wellner et al., 2007); data scrubbing (e.g., Neamatullah et al., 2008, Gupta et al., 2004, Sweeney, 1996, 2001, Ruch et al., 2000); name removal (e.g., Taira et al., 2002, Thomas et al., 2002), concept match scrubbing (e.g., Berman 2003), and statistical approaches (e.g., Sibanda 2006).

Natural language processing (NLP) based systems typically use labeled test and training sets to de-identify discharge summaries. These approaches have recall and precision rates in the range of 96-99%. However, such approaches may not work well when the text may contain ungrammatical sentences. Also, NLP-based de-identification systems have been typically used for structured text such as discharge summaries and have not been shown to be effective on less structured text such as progress notes.

Systems for name removal typically use a large repository (ranging from hundreds of thousands to millions) of first and last names. These systems scan each sentence of a medical text and assign for each word the probability that it is a name. The words with high probability are marked for subsequent removal. Although such systems have impressive (98-99%) recall rates, their usefulness is limited because these do not remove PHIs that are not names.

Statistical approaches (e.g. Sibanda 2006) involve identifying the semantic category of each word using lexical clues of the context and ontologies such as UMLS (Unified medical Language System: http://www.nlm.nih.gov/research/umls/). A semantic category organizer is trained to identify eight semantic categories from discharge summaries. The system shows high recall and precision values on test datasets. Another related machine learning approach uses a linear kernel to train a support vector machine using human-annotated data to classify each word as a PHI or non-PHI. This system is demonstrated to have high (almost 93%) sensitivity and almost 99% specificity.

The concept match scrubber (Berman 2003) uses UMLS to categorize each word into one of three categories: a non-PHI concept, a stop word such as article, conjunction, preposition etc, and a PHI. Words that are identified as a PHI are replaced with a blocking symbol (*). The system has been applied to pathology reports. However, the resulting redacted document suffers from readability because of the amount of text that is blocked out. A subsequent paper (Berman 2008) presents an improved concept-match scrubber that uses a list of 200,000 non-identifying doublets or word pairs to determine whether word pairs in the text should be blocked out. However, even with the improved version the readability of the resulting text is poor.
A comprehensive data scrubbing method that removes all eighteen PHIs is described in Neamatullah (2008). The system was tested against nursing notes and discharge summaries in the MIMIC II (Multi-parameter Intelligent Monitoring in Intensive Care) database (more details can be found in http://mimic.physionet.org/). The system consists of four look-up dictionaries: (1) look-up table for names of patients and hospital staff from MIMIC II (2) look-up tables for generic names of individuals, hospitals, and locations, (3) PHI indicator look-up tables containing keywords or phrases such as “Dr,” name indicators such as “mother,” location indicators such as “town,” and age indicators such as “patient is,” and (4) non-PHI lookup tables such as common words and UMLS terms.

Medical text such as discharge summaries and progress notes typically contains dates. The dates can be replaced with pre-defined templates, or completely blocked out. It is important for medical research to preserve information such as length of stay, intervals between events, or season-specific information. Hence, instead of completely blocking out the dates, actual dates can be obfuscated while the relevant information can be preserved by methods such as shifting the month and the year.

Numerical data that contain identifying information such as telephone, fax, and social security numbers often have specific formats. However, some of these patterns might overlap with other non-PHI numerical data that are relevant to treatments. Hence data scrubbing algorithms must be designed to differentiate between these two types.

Automatic removal of PHI from medical text remains as one of the research challenges. The unigram or bi-gram-based techniques described in Berman (2003, 2008) could be extended to n-grams in order to protect the utility of the redacted text. Machine learning approaches including spectral analysis could be used to develop more effective solutions than currently exists. Topic models could be used to separate words that are PHIs from those that are not.

**PRIVACY IN EMRS**

Various privacy models have been suggested in the literature, particularly in the literature on privacy preserving data mining and statistical disclosure control. Apart from methodological differences, these methods vary in terms of the amount of background information that an adversary is assumed to have. However, there is little guidance available on the realistic assessment of the amount of background of an attacker. This problem is overcome by one of the latest privacy models, differential privacy, that makes no assumption on the attackers background information (Dwork 2008, Kifer et al., 2011).

The general notion of differential privacy is that an attacker should not be able to gain knowledge from the results of a query whether a patient record has been added or deleted from a database no matter what background knowledge the attacker has. The increase in the knowledge gained by an attacker who queries the database instance before and after the insertion or deletion of a record is called the differential privacy. The goal is to restrict the knowledge gain to a parameter $\epsilon$, which can be publicized. The typical value of $\epsilon$ is 0.01, or 0.1 (Danker et al., 2012).

The approaches of differential privacy can be categorized into interactive and non-interactive. The basic difference between the two approaches is that for interactive differential privacy noise (typically Laplace noise) is added to the result of each query, whereas in the non-interactive approach the noise is added to the data a-priori, i.e., prior to the release of the data. However, it has been shown (Sarathy et al., 2011) that an attacker may be able estimate the values of sensitive numeric attributes with a high level of accuracy from the perturbed data with some background information. Violation of privacy can also occur with multiple queries.
As noted by Dankar et al., (2012), there are many challenges in applying current privacy preserving methods to healthcare data. For example healthcare data consists of both categorical and numerical data. Hence, methods that are primarily designed for numeric data are not application in healthcare settings. Secondly, many events in healthcare data are correlated. Examples include laboratory results and diagnoses, diagnoses and treatments, treatments and outcomes, etc. Hence data distortions to attributes individually may result in data that are meaningless for healthcare research when the attributes are considered collectively.

There is a need to explain to the healthcare and patient communities as to how privacy preserving techniques and especially the parameter settings can affect the privacy of the individuals while preserving the usefulness of the data. Also, researchers on data privacy must collaborate with those in healthcare to make sure that such techniques result in meaningful outcomes. There are still many open issues that need to be addressed before user communities adopt any of the privacy protection techniques proposed in the literature.

ANALYZING ELECTRONIC MEDICAL RECORDS

Electronic medical records contain a rich source of clinical information in the form of text data. However, the analysis of the text data requires methods automated methods for extracting clinical information. In the following section we review some of the current work and potential research opportunities. The related work can be categorized into (1) extracting clinical information from text, (2) mapping relevant text into standardized coding systems such as UMLS (Unified Medical Language Systems), (3) analyzing the text for knowledge discovery that could be useful to clinicians, (4) presenting information to clinicians in an effective manner, and (5) help end users understand clinical information in electronic medical records.

Extracting Clinical Information from Text

Several methods have been suggested for mapping clinical information to standardized coding systems such as the UMLS (e.g., Zou et al, 2003, Friedman et al., 2004). The approach taken by Zou et al., (2003) uses string matching, statistical and linguistic processing, and part of speech tagging and identification of noun phrases to map words in clinical text to UMLS. However, such methods can result in erroneous coding resulting from modifiers such as concept negation (e.g., no significant growth of tumor), reference to past diseases (e.g., previously had chicken pox), family history (e.g., family history of diabetes), or mention of events that have not actually occurred (e.g., exposed to a disease). Unless adequately handled, text such as these might be included as clinical information directly related to the patient.

Here two systems are widely reviewed that are used for automatically extracting clinical information from text: MedLEE (Friedman et al., 1994, 1995), and HANDS (Keenan et al., 2008). MedLEE has been used to extract clinical information from free text such as discharge summaries, whereas HANDS is used to extract clinical data from progress notes in nursing reports.

MedLEE consists of four major modules: pre-processor, parser including an error recovery module, phrase regularizer, and encoder. The preprocessor scans an input text and divides into smaller text units such as words and multi-word phrases. Next, using lexical look up it classifies each word and multi-word into types such as abbreviations, quantitative, temporal, parts of speech, and a special type called finding, representing clinical terms. Some word sense disambiguation is also performed at this step.
The preprocessor uses UMLS codes, a lexicon, list of abbreviations, and rules for word sense disambiguation. The parser takes the output of the preprocessor and creates a structure using a grammar based in syntactic and semantic rules. It also contains an error recovery module that helps skip words or text segments that cannot be parsed for some reason. The output of the parser goes to a phrase regularizer that deals with noncontiguous phrases where words are separated instead of being contiguous, that the parser cannot handle. For example, the phrases “increased triglyceride” and “triglyceride has increased” would result in the same output structure despite their syntactic differences. The encoder maps the regularized form into UMLS codes. The original version of MedLEE used a semi-automated method for the encoder, which was later changed to a fully automated encoder as explained in the next section.

**HANDS**

Another automated system used by nurses to enter and track clinical information, outcomes, and interventions is HANDS (Hands-on Automated Nursing Data System, e.g., Keenan et al., 2008). HANDS was developed to provide summarized information on patient care across time and space to all clinicians involved. One of the challenges in the development of HANDS was the lack of a single terminology across the healthcare profession to represent information about diagnoses, interventions, and outcomes. The HANDS system uses the structured terminologies of NADNA (NANDA 2003), NIC (Nursing Interventions Classification, Bulechek, 2008) and NOC (Nursing Outcomes Classification, Moorehead et al., 2004). These terminological systems are designed through extensive literature survey and input from nurses. Additionally, these systems are designed so that they can be maintained and evolved over a period of time. The current version of HANDS is a Web-based that allows clinicians to enter and update a patient’s plan. The histories of the plans of care are stored in a central server with appropriate access control. Clinicians are able to see a summary of the issues during the health care process as well as progress over time. The HANDS interface shows the linkages between the diagnoses with interventions, and interventions with outcomes. Keenan et al., (2008) demonstrate the feasibility of using HANDS in four hospitals on over forty thousand episodes. The data is stored in a relational database after de-identification.

In spite of the development of systems such as MedLEE and HANDS, several challenges exist. For example, there are differences between the UMLS metathesaurus used by MedLEE and the NANDA, NIC, and NOC terminologies used by HANDS. Thus, there is a need to develop a map between the terminologies so that these systems can be integrated to capture a comprehensive picture of electronic medical records.

Electronic medical records contain large amounts of data of different types including numeric, textual, and multimedia. Proper analysis of such data can provide important insights that can help both medical research and practice. There are several challenges that need to be overcome in order to fully utilize the potential of EMRs: (1) detecting temporal trends and patterns in individual patient records, (2) clustering EMRs, (3) integrating various types of EMR data, (4) visualizing electronic medical records. A lot of literature exists in finding patterns and clustering of healthcare data (e.g., Chaovalit 2010). In the following section we focus on data visualization.

**Data Visualization**

There are various visualization techniques that are applicable to electronic medical records. Some examples of such techniques include LifeLines (Plaisant et al., 1998) used for categorization of informa-
tion into facets such as medications and lab tests, heatmaps for visualizing very large datasets as well as performing image algebra for further analysis, VizAlerts (Livnat et al., 2005a, Foresti et al., 2006) and VizAware (Livnat et al., 2005b) for detecting visual correlation, GeoTime (Kapler et al., 2005) for temporal and geospatial display, and Google Earth based systems (Sundvall et al., 2007) for analyzing temporal events.

One of the earliest examples of visualization of electronic medical records is LifeLines (Plaisant et al., 1998) that provides a linear time view of events. In this system documents such as memos, lab reports, and medication documentation are called facets. The facets are displayed in a horizontal line along with associated events. Users can open and close the facets as they navigate on the event line. It is also possible to drill down from higher to lower level of details by interactive navigation tools. Bade et al., (2004) further extended the visual timeline model by adding alerts generated from quantitative data, uncertainty and quality of data, time periods that have missing values, and non-uniform timelines so that more space is provided for more interesting time periods.

Using polar coordinates to represent correlated time series data is another innovative idea that was first proposed as a data visualization scheme by Carlis et al., (1998) and subsequently by Linvat et al., (2005a, 2005b). The idea is to use the analog clock as a model to represent time and plot the facets along concentric time circles. The correlations among the facets can be observed from the overlapping regions along the time circles. The method proposed by Linvat et al., (2005a, 2005b) uses radial time axes and plots the facets around the circle. The inner region of the circle is used for maps or network charts. The facets are connected to the events using lines.

Several other systems have been proposed to plot time series data, represent events on the time dimension and linking them to provide a graphical representation. GeoTime (Kapler et al., 2005) plots the events on a time axis that is vertical to the x-y plane. An earlier system proposed by Kirby et al., (1996) uses an anatomical map of the body to show the various organs and allows users to drill down to detailed levels to enter and view various parameters stored in EMRs. Andrienko et al., (2005) suggest representing the referring events as spatio-temporal dimensions and its attributes using retinal features such as shape, size, texture, color, and orientation.

Given the wide range of possible visualization schemes that can be used to represent and manipulate data in EMRs, there are several practical questions that arise about the potential for adoption of these methods (Sundval et al., 2007). For example, if visualization methods are developed for EMRs, will clinicians adopt such methods? What changes will take place in how clinicians currently interact with EMR data? How will these impact the patients in terms of healthcare service quality?

Some positive feedback from clinicians is reported in Sundval et al., (2007) using a system based on Google Earth (http://earth.google.com). In this visualization scheme a digital “map” of the human body is placed in the background and the details of the anatomical parts are shown through region based loading and display of details. As users select and zoom into different parts of the body a different visualization piece is displayed. Similar to “placemarks” in Google Earth, notes are placed in relation to body parts. Both new and old notes are displayed as stacks with newer notes on top of older notes. Web-based data entry forms can be made available to users along with Google Earth based displays. Several navigation and display options can be used to enhance the usability of the system. Examples include: logarithmic time scaling for emphasizing recent notes over older notes, using linear and polar time lines, displaying different images of a region on a timeline simultaneously, using parallel timelines for different events such as documentation, clinical events, and presenting critical information such as allergy warnings as screen overlays. One important design criterion is aggregating the data, which can be done by facets,
Various facets can be used for data aggregation including archetypes such as observation, evaluation, or ontologies (Garde et al., 2007), openEHR folder structure for linking entries related to health problems and care episodes, terminology systems such as ICD diagnosis codes or SNOMED CT categories, or based on healthcare provider or clinician roles. Another innovation inspired by Google Earth is to use time as the fourth dimension to view a certain health condition such as diabetes or the growth of a tumor over a period of time. Animation over the time dimension can provide clinicians an opportunity to observe temporal patterns. Usability studies can be conducted on data use by collecting information such as time spent and number of user views. This can provide further insight into both the importance of the data as well as possible alternative data modeling options.

Data visualization in electronic medical records is still at its infancy. Despite the availability of a large number of visualization methods only a few have been tried and tested for electronic medical records. Here we discuss a few additional data visualization methods that can be potentially useful in finding patterns in electronic medical records. We divide these into cell-based visualization and dimension-based visualization.

Heat maps are one of the cell-based visualization methods. Heat maps are two-dimensional displays of data values that use color codes to depict different ranges of values. Heat maps have been used in various applications such as tracking web page usage in terms of the number of visits, representing the expression levels of genes (DNA microarrays) in molecular biology, and depicting financial data and calculations. Heat maps are useful to represent very large amount of data in a relatively limited amount of space. It is also possible to perform simple image-based algebraic calculations to derive new variables. Various adaptations of heat maps exist including height maps where a height field is used instead of color. Height maps have been used in representing document themes (Wise et al., 1995). Other cell-based visualization schemes include table lens, survey plots, and iconographic displays.

Dimension-based visualization methods use orthogonal dimensions as axes to provide a system of coordinates for plotting the data as points. Dimension-based systems include dimensional stacking, parallel coordinates, single and multi-line graphs. Dimensionality reduction techniques such as multi-dimensional scaling (MDS) and principal components analysis (PCA) (Jolliffe 2002) are often used to project high dimensional data into a low-dimensional subspace.

Data visualization is a somewhat unexplored area for analyzing and discovering patterns in electronic medical records. Visualization techniques can help overcome several limitations of traditional pattern recognition techniques. The ability to observe the data in a comprehensible manner allows human experts to detect patterns by manual inspection. In this respect, it automatically incorporates domain knowledge in the pattern discovery process. Visualization is not a replacement of data mining and knowledge discovery but rather a complementary tool to enhance the discovery process.

Aryya Gangopadhyay
University of Maryland Baltimore County, USA

REFERENCES


