Personal Health Information Leak Prevention in Heterogeneous Texts

Marina Sokolova[1], Khaled El Emam[1,2], Sean Rose[2], Sadrul Chowdhury[1], Emilio Neri[1], Elizabeth Jonker[1], Liam Peyton[2]

[1]Electronic Health Information Lab
Children's Hospital of Eastern Ontario
401 Smyth Rd., Ottawa, Canada, K1H 8L1

[2] University of Ottawa
800 King Edward, Ottawa, Canada, ON K1N 6N5

{*msokolova, kelemam,srose,schowdhury,eneri,ejonker*}*@ehealthinformation.ca*
lpeyton@site.uottawa.ca

Abstract

We built a system which prevents leaks of personal health information inadvertently disclosed in heterogeneous text data . The system works with free-form texts. We empirically tested the system on files gathered from peer-to-peer file exchange networks. This study presents our text analysis apparatus. We discuss adaptation of lexical sources used in medical, scientific, domain for analysis of personal health information.

Keywords

information leak prevention, personal health information

1 Introduction

When electronic means became the prime instrument for storage and exchange of personal health data, the risks of inadvertent disclosure of personal health information (i.e., details of the individual's health) had increased. Inadvertently disclosed personal health information facilitates criminals to commit medical identity theft, i.e., allows an imposter to obtain care or medications under someone else's identity [10]. Furthermore, PHI is an important source of identity theft [14], and has been used by terrorist organizations to target law enforcement personnel and intimidate witnesses [21]. PHI security breaches had happened in various domains. PHI has leaked from a Canadian provincial government agency [6] and from health care providers, through documents sent by employees and medical students [18]. There are several examples of the confirmed leaks on peer-to-peer file sharing networks: a chiropractor exposed his patient files on a peer-to-peer network, including notes on treatments and medications taken [20], a criminal obtained passwords for 117,000 medical records through a file sharing network [24]. In this work, we present a system which detects personal health information (PHI) in free-form heterogenous texts. It can be used to detect the inadvertent disclosure of PHI, thus, benefit information leak detection.

Texts which contain personal health information can be written by doctors, nurses, medical students or patients and can be obtained from various sources within the health care network. Hospitals provide patient health records (e.g. speech assessment, discharge summaries, nurse notes), patients write letters, notes, etc. These texts can be found on the web, within peer-to-peer file exchange networks, and on second-hand disk drives [12, 25]. Within those texts, we seek the information which refers to individual's health: disease (pneumonia)[1], treatment procedures (X-rays), prescribed drugs (aspirin), health care providers (the Apple Tree Medical Centre). Our system contributes to information leak prevention, a growing content-based part of *data leak prevention*.

There are several differences between our tool and the previous work on PHI leak prevention. Our system detects personally identifiable and health information. Previous work focussed on detection and de-identification of personally identifiable information (.e.g,, person names, phone numbers, age-related dates), but did not retrieve health information. Our system processes data of unknown content, context and structure. Whereas, previously the PHI leak prevention systems operated within a closed domain of hospital patient records, where the input data was guaranteed to contain PHI. As we mentioned, these systems were built to find and alter personally identifiable information, e.g., name, age, phone.

In our case, the input files come with unknown content. Sometimes the file content excludes a possibility of personal information, e.g., a young-adult vampire-romance novel Twilight, a research presentation Statistical Learning Theory, a song Quel Temps Fait Il A Paris.

[1] Hereinafter, this font signifies examples.

Workshop Adaptation of Language Resources and Technology to New Domains 2009 - Borovets, Bulgaria, pages 58–69

Sometimes, file contents may suggest holding personal information and PHI, e.g., personal correspondence, documents from lawyer or physician offices. In many other cases, files fall between these two categories. We discard files which we identify as being highly unlikely to contain PHI and concentrate on the analysis of the remaining files. In the remainder of the presentation, we define personal health information, provide examples of texts containing PHI and discuss the extent of confirmed inadvertent PHI leaks. We define pairs of possible/impossible and probable/improbable PHI containers. Our data and empirical results are presented after that. We follow with discussion of related work and motives for the adaptation of medical knowledge sources. At the end, we present plans for future work and conclusions.

2 Background

Our group works on prevention of inadvertent disclosure of personal health information in heterogenous text data. Personal health information (PHI) refers to ailments, treatments and other health-specific details of an individual. In Ontario, Canada, the Personal Health Information Protection Act [1] defines PHI as information that:

1. relates to the physical or mental health of the individual, including information that consists of the health history of the individual's family
2. relates to the providing of health care to the individual, including the identification of a person as a provider of health care to the individual,
3. is a plan of service within the meaning of the Long-Term Care Act for the individual
4. relates to payments or eligibility for health care, or eligibility for coverage for health care, in respect of the individual
5. relates to the donation by the individual of any body part or bodily substance of the individual or is derived from the testing or examination of any such body part or bodily substance
6. is the individual's health number
7. identifies an individual's substitute decision-maker.

It also states that "[identifying information] identifies an individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify an individual" [1]. Below we present samples of texts with PHI:

[THE PATIENT] ADMITTED IN TRANSFER FROM [HOSPITAL NAME] FOR MENTAL STATUS CHANGES POST FALL AT HOME AND CONTINUED HYPOTENSION AT CALVERT HOSPITAL REQUIRING DOPAMINE;

[The person]'s heart attack happened on a sidewalk in Midtown Manhattan last May. He was walking back to work along Third Avenue with two colleagues after a several-hundred-dollar sushi lunch. There was the distant rumble of heartburn, the ominous tingle of perspiration.

I, [John Doe], want to make the following statement under Oath: On 02 May 2007, in duty hours between 1030 1100 hrs I was practicing drill exercises for our next mobilization to [Place Name, Country]. In that exercise trying to put someone under arrest the person fell down on my left knee causing me a contusion.

The act protects the confidentiality of PHI and the privacy of individuals with respect to that information, while facilitating the effective provision of health care. Similar PHI protection acts have been enabled: US has the Health Insurance Portability and Accountability Act, often known as HIPAA [2], EU – Directive 95/46/EC , or, Data Protection Directive [3], although some details vary.

We divide PHI into two broad categories: personally identifiable information (PII), e.g., name, birth data, address, and health information (HI), e.g., diagnosis, prescribed drugs. Table 1 lists some PII and HI subcategories and their examples. Further, in this paper, we concentrate on the HI category.

Personal information	
Info categories	Examples
Given names	Serge, Jasmine
Locations	London, Osaka
Addresses	401 Smyth Rd., Empire State Bldg.
Dates	02 May 2008, 05/14/07
Health information	
Info categories	Examples
Disease names	Pneumonia, arthritis
Symptoms	calcium deficiency
Drug names	Aspirin, Fosamax
Health care providers	CHEO, Dr. Joe Doe

Table 1: *PHI categories and their examples*

Previously, second hand disk drives and peer-to-peer file exchange networks (p2p networks) were searched for the presence of texts with PHI [12, 25]. In [12], we studied the extent of inadvertent PHI exposure on second hand computer hard drives. We purchased functional disk drives from various second-hand computer equipment vendors, then examined sixty drives using

[2] http://www.hhs.gov/ocr/privacy/index.html
[3] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=-CELEX:31995L0046:EN:HTML

digital forensic tools. The focus was on drives with a capacity range of 10 GB to 40 GB, which were used by individual end users in desktop machines and servers. The recovered data was examined manually by two experts. PHI was found in 425 files gathered from 11(18%) disk drives. 5 disk drives contained PHI of the computer owner, 6 disks – PHI of other people. In [25], we suggested evaluation measures for automated detection of PHI files. We semi-manually examined 859 files gathered from two p2p networks. 8(1%) files contained PHI. Although the number of files was comparatively small, the personal health information contained in these files potentially was exposed to millions of on-lookers.

3 Methods

3.1 Files as Personal Health Information Containers

We consider that among heterogeneous files of unknown content, some files can be *possible* containers of personal information, whereas others cannot, i.e. *impossible* containers. Only the possible containers which contain personal information may become *probable* sources of PHI leaks. The files become actual PHI leaks if they additionally contain health information. The impossible containers are improbable to leak personal information, and consequently, improbable to leak PHI. In other words,

$$DATA = POS + POS' \quad (1)$$

$$PHI \subset PROB \subset POS \quad (2)$$

$$POS' \subset PROB' \subset PHI' \quad (3)$$

where $DATA$ denotes the data, POS is a set of possible containers, $PROB$ – a set of probable containers, PHI – a set of files with PHI. $'$ marks the set complement. Based on the set relations 1–3 , we apply the following rules of inference (f denotes a file):

$$f \in POS' \rightarrow f \in PROB' \rightarrow f \in PHI' \quad (4)$$

$$f \in PHI \rightarrow f \in PROB \rightarrow f \in POS \quad (5)$$

To categorize files into POS or POS', we recall that personal information is information which identifies an individual, either by itself or jointly with other information (see Section 2). According to this definition, two types of files immediately fall into the impossible container category:

1. contents not concerned with individuals;
2. contents not able to identify individuals.

To find the first type, we look for files with content unrelated to individuals, e.g., fictions, songs . To find the second type, we look for files whose content is unreadable for end users, e.g viruses. All the other files might or might not contain personal information, and are put into the possible container category. From those, only the files that are shown to contain personal information are marked for further investigation, i.e. health information detection.

Rules 4 – 5 are used in our text analysis system. The system's work cycle has three phases. On the first two phases, it identifies impossible containers and removes them from the set. On the third phase, it works with the remaining files which we consider to be possible containers. The utilized text analysis gradually deepens:

shallow analysis the file titles and contents are treated as streams of characters (phase 1);

partial content analysis for each file, a limited keyword search is performed on a small portion of text (phase 2);

deep content analysis texts are mined for syntactic and semantic patterns (phase 3).

3.2 Creating the Corpus

For our experiments, we used files gathered from two peer-to-peer file sharing networks (p2p networks). Peer-to-peer networks allow decentralized sharing of computer resources, including those that provide the infrastructure for direct, real-time communication and collaboration between peer computers. The networks are known for hosting files with PHI information.

The usage, together with the observed security weakness, marks p2p network data as a possible source of information leaks, including PHI leaks. This assumption was confirmed by data management studies [18].

To gather data, we obtained the project approval from the Research Ethics Board of Children's Hospital of Eastern Ontario. The files were gathered from April 2008 till June 2009. The *Gnutella* and *eD2K P2P* networks were selected due to their prevalence and global popularity.[4] To automatically capture samples of p2p files, we modified the publicly available ShareAZA[5] p2p client. The tool is a software package which allows one to connect to multiple P2P networks simultaneously in order to search for and download files. Modifications to this client included changes to the search function as well as increased logging capabilities. The search function was modified to automatically search for any document file (Microsoft Word, Raw Text, Rich Text, Excel, Powerpoint, PDF, Wordpad, XML, etc) and automatically retrieve it. Automatic searches were conducted by the code at fifteen minute intervals. A semi-manual analysis of the first data sample (859 files) showed the presence of PHI on the two p2p networks [25]. In total, we have gathered 2852 files. The data was sent for processing "as is", without preliminary pre-processing: we preserve all the initial spelling, capitalization, grammar, etc.

[4] `http://www.kolabora.com/news/2004/01/09/popular_p2p_tools_and_programs.htm`, retrieved Aug 12, 2009

[5] `http://shareaza.sourceforge.net/?id=source`

Detection of publishable and educational text			
Categories	Examples	Categories	Examples
Books	ebook, ISBN, publisher	Periodic	magazine, article, volume
Retail	manual, readme, copyright	Genre	biography, fiction, sci-fi, whodunit
Book type	dictionary, novel, cookbook	Publishable	abstract, acknowledgement, introduction
Education	theses, assignment, course-work	Special (NA)	Bible, dummies, Microsoft, software
Detection of non-personal text			
Categories	Examples	Categories	Examples
Music	album, ballad, song	Fictionals	Harry Porter, Scarlet O'Hara
Advertisement	Mrs Tiggy Winkles, Tim Hortons	Politics	Al Gore, Winston Churchill

Table 2: *Categories and terms for partial content analysis*

3.3 Empirical Text Processing

Shallow analysis On this step, we aimed to remove files which were the most unlikely candidates to leak PHI. We assumed that any published text was not leaking PHI: fictions described non-existing heros, magazines and newspapers obtained person's consent on information disclosure, songs were not providing enough details, etc. Corrupted files and non-text files (images, music) were other candidates for a fast removal. On this step, we applied string matching and character-based N-gram modelling methods.

First, we processed only the file titles. The titles were compared with the `Amazon.com` database. 723(25.35%) files were removed after their titles were found in the database, e.g. New York New York, Abba_-The_Winner_Takes_It_All, abominable snowman were discarded. However, if there was no exact title matching, the file was passed for further processing.

On all of the following steps, we worked with whole file data, i.e., body and title. Immediately after the `Amazon.com` search, the files were passed through a text extractor. 37(1.30%) non-text files (images, music, viruses) were removed. Then we applied a modified version of a publicly available language identifier TextCat [6]. For each file, the tool built character-based N-grams. The N-grams were compared with language models for 69 languages. The best-fitting model provided the language text tag. On that step, 724(25.39%) non-English texts were discarded (e.g., 13 de octubre 20008, canserle ilgili bilgiler).

Partial content analysis The goal of this phase is to identify and remove publishable, educational and non-personal texts which cannot be identified by the title string matching. #11 The Dragonfiend Pact_1, Copy of VintagePatternBook are book files with "camouflaged" titles that passed through the string matching method of the shallow filter; chriscolumbus is a student assignment which could not be detected earlier. To find such files, we define categories of terms characteristic to publishable and educational texts. One category represents local North American (NA) preferences in the files; see the upper part of Table 2 for examples.

We also look for texts with non-personal content. Music texts, discussion of popular fictional characters, current political events, and advertisements would be unlikely candidates for leaking explicit, detailed PHI. The lower part of Table 2 lists detection categories and examples of terms.

In practice, we applied the key-word search to the titles and the first 200 words of the body text. Many detected texts were educational (assignments, reports, theses), some represented small literary forms (essays, poetry, self-published books). Manuals, tech reports, articles were also detected on this step. As a result, 605(21.21%) publishable and non-personal texts were filtered out.

Another task was to find and remove multiple copies of the same file. For each pair of remaining files, we compared their sizes, titles, and first and last sentences. If all parameters were the same, we tagged two files as duplicates and kept only one for further processing.[7] 41(1.44%) files were removed on this step. The remaining p2p files are deemed susceptible to information leaks, i.e., possible containers, and passed to the deeper analysis stage; Figure 1 shows the proportional distribution of the processed files.

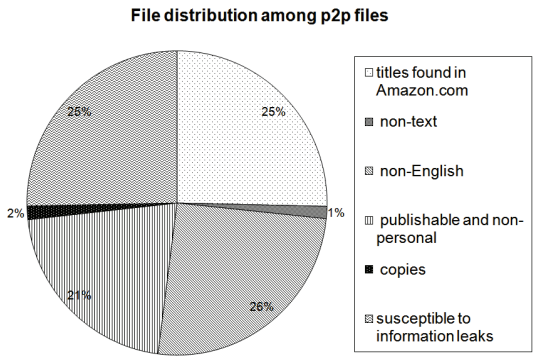

Fig. 1: *File content distribution found by shallow and partial content analysis*

[6] http://odur.let.rug.nl/ vannoord/TextCat/

[7] Here and everywhere, when appropriate, we used hashes and compared hash values.

Deep content analysis 722(25.31%) files remained after the shallow and partial content analysis phases. Potentially, those files may have personal content, thus, hold personal health information (i.e., possible containers). In processing these files, we want to identify a set of probable containers first and then work with this set only. For this, complete contents are analyzed with a combination of syntactic, and semantic methods. This phase uses external resources: dictionaries and knowledge sources.

The heterogeneity of the data makes it unrealistic to expect such file commonalities as text structure, grammar style, content word vocabulary, etc. We, instead, rely on the definition of PHI which can be expressed through a reasonably limited number of semantic categories, e.g., person and geographic names, disease names and symptoms. Sets of syntactic rules are used to identify references to individuals, locations and age-identifiable events (birth, death). We parse sentences to find preposition phrases, noun phrases and verb phrases. Soft Regular Expressions(RE) are used to extract numeric-based categories, such as phone number, street number and unit, dates, and email.

In data management and privacy protection, geographic information is shown to be the single most important category responsible for person identification [16, 3]. We implement geographic information extraction for the following categories:

country : all the UN-recognized countries and their capitals on all the continents (France, Paris; Liberia, Monrovia), and self-proclaimed entities (Eritrea, Abkhazia) ;

place : in US: state name, state capital, the largest city (Illinois, Springfield, Chicago); in Canada: province, province capital, largest cities, tourist attractions (Alberta, Edmonton, Calgary, Banff), the same – for territories; in Europe, Latin America, Asia, Africa, Australia: Alpha, Beta and Gamma world cities [8].

code : US' ZIP code (Massachusetts 02163, NY 10027), Canada's postal code (K1H 8L1);

street : for US and Canada – type (Avenue, Ch., Street, Beach), number (401 Smyth Rd.);

landmarks : Empire State Building, CN Tower , Niagara Falls, etc.

From other named entities, we concentrate on recognition of person names (John Smith), organizations (Nepean High School) and health care providers (Ottawa General Hospital). We considered organizations to be geographic pointers, as they can tie an individual to a certain location. The Contact Us link on the organization web site or a name part (Boston University) are strong indicators of a person's geographic affiliation. Hospital, doctor, and registered nurse information in the file makes it a strong candidate for a PHI leak. We combined a key word search and syntactic patterns to find these named entities. We did not apply look ups of health care providers, as no confirmed high quality sources are available

To reduce computationally expensive person name look-up, we first searched for patterns of family relations (My daughter, an uncle of) and self-identification (my name, sincerely). Other patterns are event-related (was born, died in). Depending on the patterns, either preceding or following capitalized words are stored in the name list. Further, when the tool checks for a person name, it will first check with the file dictionary. The pattern search is augmented with an RE-based search. The latter is combined with the person name look up. We use three proprietary dictionaries: female and male first names and last names. Our dictionaries contain formal and informal name forms (William, Bill, Billy) and non-Anglo-Saxon names (Meehai, Leila).

To be marked as probable containers, files should contain a geographic identifier (e.g., street address, place name, organization) and two other identifiers, e.g., first name and last name, first name and another geographic identifier, last name and a phone number. The 345(12.10%) probable containers were then passed into the final phase of health information extraction where 12 PHI files were found (Figure 2). We discuss in detail PHI extraction in Section 3.4.

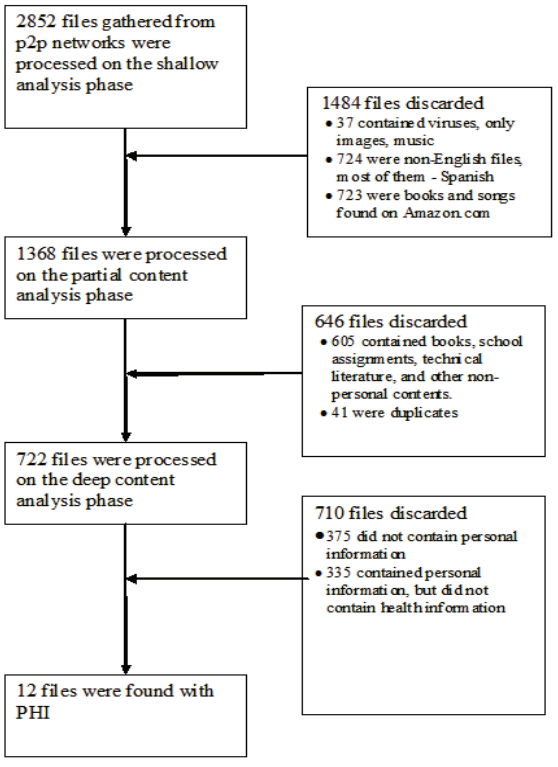

Fig. 2: *Gradual reduction in the number of analyzed files*

[8] http://en.wikipedia.org/wiki/Global_city#GaWC_Inventory_of_ World_Cities

3.4 Health Information Extraction

3.4.1 Ontology structure

Our ontology building works as follows:

i to identify a small number of semantic categories which correspond to the main categories of Health Information;

ii work with each category separately, identifying the information that should be analyzed;

iii apply Information Extraction methods to find the information indicators in the existing sources.

At the initial step, we form three genetic semantic categories – disease, drugs, and symptoms – as was discussed in Section 2; see Table 1 for examples. For diseases and drugs, we concentrate on extraction of their names. The category contents were derived from Webster's New World Medical Dictionary [15], the International Classification of Diseases (ICD9 codes)[9], the Medical Dictionary for Regulatory Activities (MedDRA)[10] and Canadian Drug Product Database (Active and Inactive) [11].

ICD9 codes [2] are used by health care professionals to tag and classify morbidity data from inpatient and outpatient records, physician offices, as well as most of the National Center for Health Statistics (NCHS)[12] and the Canada Institute for Health Information[13] surveys. The codes are divided into two sections: one containing diseases and injuries (ICD9CM Disease and Injury), and another containing surgical, diagnostic, and therapeutic procedures (ICD9CM Procedures). ICD9 provides the hierarchy of diseases where terms on every level relate to the individual's health. The following sample presents the complete hierarchical snapshot for cholera:

1 INFECTIOUS AND PARASITIC DISEASES (001-139)
INTESTINAL INFECTIOUS DISEASES (001-009)
Excludes: helminthiases (120.0-129)

001 Cholera

001.0 Due to Vibrio cholerae

001.1 Due to Vibrio cholerae el tor

001.9 Cholera, unspecified

This succinctness allows reduction in the source processing and simplifies information extraction steps.

The Canadian Drug Product Database (CDPD) contains product specific information on drugs approved for use in Canada. It includes human pharmaceutical and biological drugs, veterinary drugs, and disinfectant products. Additionally, a database of previously available drugs is maintained. However, an average, non-expert individual may treat drugs as consumer goods and refer to them in different ways. That is why we expect that drug names can vary from a generic, non-proprietary, name as Ibuprofen to a more specific brand name as Advil.

To accommodate extraction of various drug names, we sought information provided by Merck & Co., an international pharmaceutical company[14]. We obtained a list of generic drug names and the trade names associated with them [15].

Patient symptoms such as chest pain or headache, as well as mentioned procedures such as heart surgery, also consist of health information, as they may allow one to infer a specific medical, behavioural or psychological condition or ailment of another individual. To identify patient symptoms, we use the MedDRA dictionary which covers a wide range of terminology including symptoms and signs (i.e. visible symptoms).

However, the listed above resources leave some gaps in PHI detection. The most noticeable absentees are acronyms (ICU) and providers (therapist, surgeon), but also some condition names (blood pressure, tube fed). To fill the gaps, we manually searched the Webster's medical dictionary. Figure 3 shows the structure of our ontology.

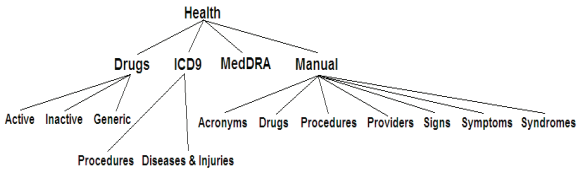

Fig. 3: *The structure of the knowledge source*

3.4.2 Terms and term units

We aimed to populate the ontology with PHI-related single terms (diabetes) and term units (Felty's syndrome). We first minimized the above mentioned resources by removing un-related categories (e.g., animal diseases, animal drugs). Then the remaining resource texts were normalized : converted to lowercase, punctuation marks and numbers were removed, and stop words (of, when) were eliminated.

We consider that a term unit is a sequence of two or more consecutive units words, that has characteristics of a syntactic and semantic unit, i.e. collocation. To identify collocations, we used a subset of 700,000 articles from the MEDLINE corpus [16], a repository of medical documents. The normalized text was processed by TEXT::NSP[17], a collocation extraction and N-gram building tool. We looked for N-grams of

[9] http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm
[10] http://www.meddramsso.com/MSSOWeb/index.htm
[11] http://www.hc-sc.gc.ca/dhp-mps/prodpharma/databasdon/index-eng.php
[12] http://www.cdc.gov/nchs/
[13] http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=home_e
[14] http://www.merck.com/
[15] http://www.merck.com/mmpe/appendixes/ap2/ap2a.html
[16] http://medline.cos.com/
[17] http://search.cpan.org/ tpederse/Text-NSP-1.09/lib/Text/NSP.pm

length 2 and 3. From these counts a log-likelihood statistical significance test was performed to determine if a given textual unit qualifies either as an N-gram or a collocation. We used the tool default settings. The sets of trigrams and bigrams were then merged into a single set of collocations which is used later in the process.

Each of the IC9CM, CDPD, Merck, MedDRA, and manually created datasets were then used to find and extract PHI entities by applying the following procedure: try to match collocations, and if a match is found mark it as a PHI indicator; if a word is not matched as part of a collocation, mark it as a single word PHI indicator. However, this inclusiveness may reduce the detection power of the terms. For example, cat and magic would be extracted from the drug base entry:

```
CAT IV - SUNBURN PROTECTANTS, LEG MAGIC
```

Thus, the list of HI indicators had to be filtered and pruned to eliminate false HI indicators like these. For filtering of these words, we used data obtained from the British National Corpus (BNC)[18] The BNC is used to filter out terms that occur over 1800 times. This threshold was chosen, as hospital occurs slightly less than 1800 times. Additionally, we eliminated content words which appear among the top frequent 5000 words in the Brown corpus [19]. No collocations were filtered out by this process, which may in part cause false PHI indicators to be included in the final list. This process resulted in the ontology with 62004 entities, including 25528 unigrams, 27641 two term collocations, and 8835 three term collocations. Figure 4 sketches the ontology building process.

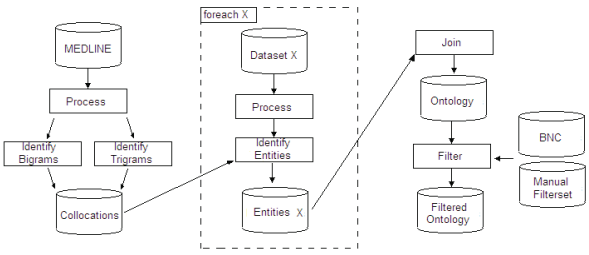

Fig. 4: *Health Information ontology building. X denotes an external source*

3.4.3 Text classification

After the ontology has been created, it can be used to classify a text as PHI or not PHI. Based on the number of trigrams T, bigrams B and unigrams U in the text, identified as HI indicators, along with the length of the document n, we classified the text as either PHI or not PHI. We then could weight N-gram contributions proportionally to the number of words N :

$$\frac{3T + 2B + U}{n} > H \rightarrow PHI \tag{6}$$

[18] http://www.natcorp.ox.ac.uk/
[19] http://www.edict.com.hk/textanalyser/

Texts were marked as PHI' if their N-grams did not satisfy Eq. 6 .

If we wanted to use the ontology structure, we would augment the formula by including the term category contributions. Each category is assigned a normalized weight based upon its quality. We rate the quality of a resource based upon the percentage of entities within it that are not filtered out. So a category C_i of original size (i.e., a number of initial terms) S_i and filtered size (i.e., a number of remaining terms) F_i has quality Q_i defined as:

$$Q_i = \frac{F_i}{S_i} \tag{7}$$

We then compute the normalizing factor M and the weight W_i, given to an entity from a given category:

$$M = \frac{1}{n}\sum_{1}^{k} Q_i \tag{8}$$

$$W_i = \frac{Q_i}{M} \tag{9}$$

The weighted formula for text classification becomes:

$$\sum_{1}^{k} W_i \frac{3T_i + 2B_i + U_i}{n} > H \rightarrow PHI \tag{10}$$

Empirical testing of several thousand documents allowed us to determine a suitable threshold value of 0.04. Yet, the threshold H was chosen from a set of empirical experiments such that it optimized the precision, while keeping the recall at 100%. Thus, the chosen threshold value may be overly fit to the test data. In the future, when more sample data becomes available, experiments should be performed to try and to determine what the optimal threshold value should be. Figure 5 depicts the PHI text classification process.

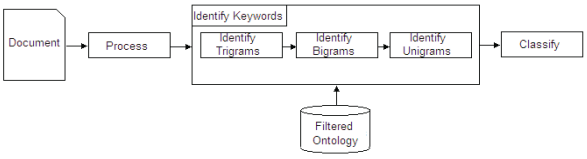

Fig. 5: *HI text classification*

4 Empirical Results

Evaluation on screened positive examples Evaluating the correct identification of PHI leaks presents a certain methodological difficulty. The number of PHI files is negligible even if compared with the number of probable containers. On the other hand, all of the PHI files exhibit specific characteristics: they contain personally identifiable and health information. Hence, we can apply measures that evaluate a tool's performance only on examples which satisfy pre-determined criteria [23]. The other examples

are ignored. The approach – evaluation on screened positive examples – has been shown effective and appropriate for PHI leak detection [25]. Table 3 presents the confusion matrix:

Actual	Predicted HI =1	Predicted HI =0
PHI=1	n^+_{PHI}	?
PHI=0	$n^+_{\overline{PHI}}$	?
	n^+	n^-

Table 3: *Confusion matrix for classification of screen positive examples*

We compute *True Detection Probability*($T\hat{D}P$) and *False Referral Probability*($F\hat{R}P$):

$$T\hat{D}P = \frac{n^+_{PHI}}{n^+ + n^-} \tag{11}$$

$$F\hat{R}P = \frac{n^+_{\overline{PHI}}}{n^+ + n^-}. \tag{12}$$

TDP shows the proportion of files an algorithm marked as having the PII and HI indicators `and` containing PHI. FRP shows the proportion of files the algorithm marked as having the PII and HI indicators but `not` containing PHI.

To put the measures in perspective, we use the ideal classification (Table 4) where all the predicted HI files are indeed the PHI files and visa-verse:

Actual	Predicted HI =1	Predicted HI =0
PHI=1	n^+	?
PHI=0	0	?
	n^+	n^-

Table 4: *Confusion matrix for the ideal classification of screen positive examples*

Then, $T\hat{D}P_i = \frac{n^+}{n^+ + n^-}$, $F\hat{R}P_i = 0$. A method works better if $\frac{T\hat{D}P}{T\hat{D}P_i}$ is close to 1 and $F\hat{R}P$ – to 0.

The ontology application We tested our tool on several sets of p2p files. Here we report typical results, in terms of accuracy.

(a) 72 files were randomly obtained from a peer-to-peer file sharing network. The set contained nine files with health care information (parents' notes, letters, documents from a lawyer office). We used our HI ontology and manually examined all the labels output by the system. Table 5 shows the results.

Here, $T\hat{D}P_i = 11.11\%$. We obtained $T\hat{D}P = 11.11\%$, thus, $\frac{T\hat{D}P}{T\hat{D}P_i} = 1$. We obtained $F\hat{R}P = 1.39\%$: HI indicators were extracted from a summary of a teen fiction which was not a PHI file.

	HI_o =1	HI_o =0
PHI=1	8	?
PHI=0	1	?
	n^+	n^-

Table 5: *HI ontology: classification of the* 72 *files*

For *per-term* extraction accuracy, we obtained Recall = 100%, i.e., all health care indicators were correctly extracted. On the relevant, true PHI, eight documents, we obtained Precision =100%, i.e. all extracted indicators were health care indicators indeed.

(b) To test the proposed system a set of 76 texts were used. This set was composed of 4 PHI texts and 72 non-PHI texts. Table 6 lists the results.

	HI_o =1	HI_o =0
PHI=1	4	?
PHI=0	2	?
	n^+	n^-

Table 6: *HI ontology: classification of the* 76 *files*

We obtained $T\hat{D}P = 5.26\% = T\hat{D}P_i$. Again, $\frac{T\hat{D}P}{T\hat{D}P_i} = 1$. We obtained $F\hat{R}P = 2.78\%$. Of the false PHI files, one was a resume of a healthcare worker, and one was an unfilled health insurance form. In both cases the falsely classified texts contained both PII and HI, yet there was no link between the two. In the future, a deeper analysis phase – perhaps, co-reference resolution – of potential PHI texts could be done, potentially increasing the precision of the method as a whole.

Medical Subjects Heading application Medical Subjects Heading (MeSH), a controlled vocabulary thesaurus, is produced by the National Library of Medicine.[20]. Its hierarchical and categorized structure is often used in analysis of medical texts [7]. In our case, however, MeSH would require a considerable adjustment before it can be used for PHI leak detection. For example, the top hierarchical terms are too general: Anatomy, Endocrine system (level 1) and Bladder, Work Schedule Tolerance (level 3) do not contribute much to the knowledge of a personal health situation. The bottom level terms might be informative only to experts but not to the general population, e.g Motor Cortex (level 8), Trypanosoma cruzi(level 11). In both cases, the use of these terms could considerably increase the number of false positives. On the other hand, the un-predictability of the input data makes it desirable for our tool to use terms belonging to all the categories, perhaps mapping it to only one of the assigned fields.

To support our claim, we used topical descriptors (DC= 1) represented by the Main Heading field, without selecting specific hierarchical level. We filtered out the stop words and frequent content words. This left 21470 terms (Corneal Ulcer, Fibromyalgia). We applied

[20] `http://www.nlm.nih.gov/mesh/`

them to classify a sample of 50 probable files. All of the PHI files were detected, but the number of false positive files was very high. 12 files were identified as containing PHI, whereas the correct number was 4 files (Table 7). Publisher, technology are examples of MeSH terms contributing to the file misclassification.

	HI_m =1	HI_m =0
PHI=1	4	?
PHI=0	8	?
	n^+	n^-

Table 7: *MeSH: classification of the* 50 *files*

In this case, $\hat{TDP} = 8.00\% = \hat{TDP}_i$, $\frac{\hat{TDP}}{\hat{TDP}_i} = 1$. On the other hand, $\hat{FRP} = 16.00\%$ shows over-inclusiveness of the test. A small number of examples, however, does not allow for conclusive remarks. We plan more experiments when new files will be gathered.

5 Related Work

Information Leak Prevention From a data management perspective, our problem belongs to information leak prevention, a part of *data leak prevention*. Intentional and un-intentional leaks of data have become a major issue for businesses, end users, software and network providers, etc. Many companies (*Symantec* [21], *Websense*[22], etc.) concentrate their efforts on building tools able, ideally, to prevent or, at least, minimize such leaks. These tools are based on organizational policies and identify, monitor, and protect data at rest, in motion, and in use [23]. Information leak prevention is concerned with content analysis of data. Information leak prevention tools are deployed in banks, financial companies, government organizations [4]. While processing free-form text data, some of the tools apply NLP methods to enforce safer data management [4]. Many of those tools work on specific text structure and type. Our tool, on the other hand, contributes to the solution of a specific task (i.e., prevention of PHI leaks) without constraining this solution to predefined text structure or types.

The applied research community participates in tool development for information leak prevention [13, 19]. *Microsoft Research* developed a *defensive* tool that looks for personally identifiable information in one's own documents. The tool processes digital documents, including metadata, and removes the owner's name, username,security ID, computer NetBIOS name, names of online, email, webmail servers, etc. No NLP or TDM techniques are involved: all documents are treated as flat byte streams. First, the tool collects potentially sensitive information from the computer, then searches documents for its presence. The tool is semi-automated. User intervention is required to reduce false positives, i.e., non-sensitive information wrongly labelled as PII [5]. In related efforts, academic groups mostly work on hospital record de-identification; more details follow in the next paragraph. Few teams are actively involved in health information leak prevention outside of the de-identification of hospital records. In [8], the authors propose a method which detects the inference of sensitive information in documents. The method relies on search engines to find the most frequent association of topic-based terms. In [26], the author describes a method which warns web site owners if posted personal information enables identity theft (e.g., date of birth, address, name).

We, instead, focus on leak prevention techniques able to detect information within heterogeneous texts.

De-identification of personally identifiable information So far, PHI detection attracted only limited attention from Text Data Mining and NLP communities. Mainly, the work has been restricted to preparation of hospital records for future use by other researchers, i.e., the secondary use of health data. In Europe and North America, the law requires removal of personally identifiable information, *data de-identification*, before permitting documents for secondary use [11]. The privacy protection requirements made de-identification popular among NLP, Information Extraction, and Machine Learning applications implemented on hospital records and other PHI texts.

Typically, a de-identification method is designed for one type of documents, e.g. discharge summaries [27, 28]. De-identification consists of the detection of patients' personally identifiable information and its subsequent transformation. De-identification tasks are restricted to detection and transformation of PII, and avoid the analysis of terminology concerning health conditions of patients. The reported systems' primary methods are look-ups of person and geographic name dictionaries. Their performance, thus, depends on comparability of the dictionaries and the input data. Table 8, adapted from [22], compares the performance of a publicly available de-identifier Stat De-id when it uses customized dictionaries (the left part) and without them (the right part). The results were obtained on the de-identification of nurse notes.

Customized dictionaries			Non-customized dictionaries		
Fscore	*Pr*	*R*	*Fscore*	*Pr*	*R*
84.4	74.9	96.7	77.4	72.3	83.4

Table 8: *Classification (%) of person names, health care provider names, addresses, age-related dates.*

Unfortunately, many publications do not report on disambiguation results, e.g. person and geographic names (Washington - surname, city, or state? Sofia - name or city? Marina - name, street or area?), or

[21] http://www.symantec.com/index.jsp
[22] http://www.websense.com/content/home.aspx
[23] http://media.techtarget.com/searchFinancialSecurity/downloads/Understanding_Selecting_DLP_Solution.pdf

person and trade marks (Tim Hortons - a coffee chain or a person? Jo Malone – a private label or a person?). Withholding of disambiguation accuracy makes it difficult to correctly assess the tool's performance.

We, on other hand, focus on heterogeneous texts whose content and context is not determined before our system processes them.

Health care information extraction There are few publications dedicated to health care information analysis in free-form, unstructured texts. Presented work often focuses on specific, rather narrow information categories. In [17], the authors compare four commercial tools which extract medication name, route, dose (number-based), strength (number-based) , and frequency (number-based) from discharge summaries and family practice notes. In [29], the authors focus on detection of obesity-related diagnostic information. They used NLP methods to extract 16 obesity diagnoses from dictated physician documentation. We, instead, opt for detection of all the HI categories.

6 Discussion

In recent years, Text Data Mining and Natural Language Processing communities have concentrated their efforts on the analysis of medical, biomedical and bioinformatics texts. With educational and research medical publications rapidly increasing (for some types, the increase fitting an exponential curve [9]), machine-readable lexical and knowledge sources were built to promote mining of medical texts: MedLine[24], GENIA corpus[25], MeSH[26], to name a few.

The services of medical and allied professionals are offered through Health Care[27], the industry which provides the prevention, treatment, and management of illness and the preservation of mental and physical well-being. Although medicine and health care are closely related, the domains produce remarkably different text data. A bulk of texts containing medical information comes from articles in medical journals, magazines, professional blogs, research publications. These are formally written, well-edited, *knowledge-rich* texts. Texts containing PHI are mostly internal reports, letters and various forms of personal communication. Often, they offer a description of the individual's health and lifestyle and related information such as treatments they are receiving and drugs they are taking. The texts are written without adherence to requirements of formal editing. Sometimes they are unedited, containing grammatical and lexical irregularities. For example, hepatitis can be shortened as hep, and future actions can be described as Assess/plan. Recent publications show the increased demand for automated Health Care text processing; for example, see the Journal of the American Medical Informatics Association [28]. However, there are no readily available lexical resources that cope well with Health Care text characteristics. The existing medical resources may require adaptation. Their "as is" application may give insufficient results in terms of effectiveness (missed relevant information, whereas non-relevant information captured) and efficiency (a long processing time and extra computational resources).

For example, the Medical Entities Dictionary (MED)[29] is an ontology containing approximately 60000 concepts, 208000 synonyms, and 84000 hierarchies. This powerful lexical and knowledge resource is designed with medical research in mind, as opposed to detection of personal health information which may require a more concise knowledge base. We borrow Figure 6 from the MED web site. [30] It shows the term Plasma Glucose Test with its relationship to other terms in the MED database. Solid lines connect it to parents in the *isa* hierarchy, broken lines are nonhierarchic semantic links.

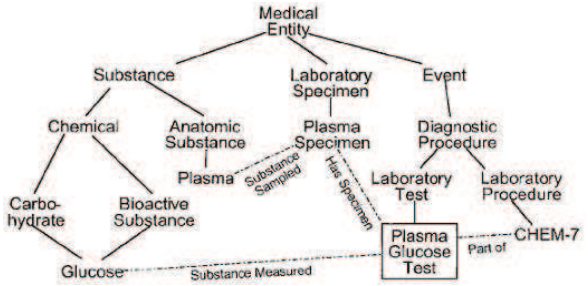

Fig. 6: *The term Plasma Glucose Test and relations to other terms in the MED database (adapted from the MED web site)*

Consider a text which contains the term Plasma Glucose Test. When referring to an individual, the term indicates examination for diabetes. There are two types of the test. Random Plasma Glucose Test refers to a simple blood sugar test. No fasting or glucose administration is required. Results are processed within 24 to 48 hours or faster. Fasting Plasma Glucose Test is more demanding: the patient should avoid food or drink, except water, for at least 12 hours prior to the procedure [15]. Both test types and the generic term Plasma Glucose Test relate to the physical health of a person, thus, are HI (see Section 2, point 1). CHEM-7, metabolic panel testing[31], is a more general term which indicates 7 possible tests (glucose, serum sodium, serum potassium, etc.). The term is HI, although it is less revealing than Plasma Glucose Test.

Other terms, e.g., Bioactive Substance, Plasma, Event, may or may not reveal HI, depending on the context. Searching texts for all the sixteen terms in-

[24] http://www.nlm.nih.gov/databases/databases_medline.html
[25] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi
[26] http://www.nlm.nih.gov/mesh/
[27] http://medical-dictionary.thefreedictionary.com/health+care
[28] http://www.jamia.org/
[29] http://med.dmi.columbia.edu/
[30] http://med.dmi.columbia.edu/struc.htm
[31] http://www.nlm.nih.gov/medlineplus/ency/article/003462.htm

creases the computing time by a factor of eight (the near linear processing was confirmed in laboratory testing). The extra time might not be a problem for topic classification and text mining which process published and otherwise legitimately dispensed documents. We, on the other hand, want to limit information exposure to a reasonable minimum. The excessive terms can also increase the probability of texts being falsely tagged PHI.

We opted to build a HI lexical source from the knowledge sources used by medical and allied professionals in health care organizations. The two primary sources are the International Classification of Diseases (ICD9 codes)[32], and Canadian Drug Product Database[33].

7 Future Work and Conclusions

The separation of possible and impossible containers is a key factor of our tool performance. To insure high accuracy, we want to implement a more diligent testing of the file titles. In future, before being fed into the `Amazon.com` search, the titles will be pre-screened to find possible legal and health-related documents (Jane Doe letter of assessment, My affidavit). We plan to use stemming during this pre-screening. The sought after key words belong to two groups: authorized and unauthorized evidence (e.g., affidavit, permission, statement) and health records (e.g., discharge, hospital, referral). We also plan to extend geographic information analysis: (i) add more categories, e.g. population hubs such as major airports (Heathrow, Pearson), international resorts (Varna, Saltsburg), etc.; (ii) rank the found names according to their contribution for person identification; for example, New York, pop. $> 8,300,000$, can be ranked lower than Ottawa, pop. $812,000$. In future, we want to reinforce person names with statistical evidence of their use, e.g., a reverse rank on the list of popular North American names. These techniques should allow file ranking with respect to a potential risk of information leaks. In this study, we focus on contents seen by end users and do not collect the hidden file metadata. We may want to investigate the metadata impact on the scale of information leaks.

By all means, we also plan to continue testing the tool. A restricted number of PHI files can make our tool prone to data over-fitting. Hence, we continue to gather new data samples. Our future work may include analysis of the MeSH hierarchical levels, in order to reduce the number of false positive examples. We also want to use BNC to find frequent collocations, as filtering these out from the PHI indicator dictionary can, too, reduce the false positives. Other directions of future work are related to detection of rare events (e.g., a rare coma complication Lock-in). If found in text, such event can correct identify a person. However, automated extraction of rarely supplied information is difficult and may require expanding our system with a new element.

We have introduced a system which helps to prevent leaks of personal health information. Our system is able to work within the complex environment of previously unseen data types. It prevents the leakage of PHI texts in heterogeneous input, i.e. files in which context, content, and type vary unlimitedly. The uncertain content of the files contrasts our data set with homogenous data sets of known content; for example, company employee files and hospital records, where each file is pre-disposed to contain personal information, or movie script archives, where scripts are a work of fiction. To accommodate the uncertainty, we have introduced a taxonomy of files related to the possibility and probability of the files leaking personal health information.

On empirical evidence, we have shown that the medical sources are well-suited to analyze formally written, well-edited texts, often with an abundance of scientific terms and relations. For our task, however, the sources contain excessive information, making text analysis too slow, inefficient and prone to false positive identification. A series of experiments was performed on files exchanged in peer-to-peer file sharing networks with encouraging results.

Acknowledgements This work has been funded by the Natural Sciences and Engineering Research Council of Canada and the Ontario Centre of Excellence. We thank Terry Copeck for his assistance with text extraction. We thank anonymous reviewers for their helpful comments.

[32] `http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm`

[33] `http://www.hc-sc.gc.ca/dhp-mps/prodpharma/databasdon/index-eng.php`

References

[1] Personal Health Information Protection Act. Legislation of Ontario, 2004. `http://www.e-laws.-gov.on.ca/html/statutes/english/elaws_statutes_04p03_e.htm`, accessed Sept. 7, 2008.

[2] The International Classification of Diseases, 9th revision, clinical modification, 2007. National Centre for Health Statistics, Centres for Disease Control and Prevention, US Government.

[3] Review of systems for extracting and anonymizing geographic information. Technical report, Electronic Health Information Lab, CHEO, submitted to GeoConnections, 2008.

[4] Understanding and selecting a data loss prevention solution. Technical report, Securosis, L.L.C, the SANS Institute, 2009.

[5] T. Aura, T. Kuhn, and M. Roe. Scanning electronic documents for personally identifiable information. In *Proceedings of the 2006 ACM Workshop on Privacy in the Electronic Society (WPES 06)*, pages 41–50, 2006.

[6] M. Baird. Personal files were accessible for more than three weeks. The Western Star, 2008.

http://www.thewesternstar.com/index.cfm?sid-=104156&sc=23, retrieved Feb 5, 2009.

[7] L. Bouma and M. de Rijke. Specificity helps text classification. In *Proceedings of European Conference on Information Retrieval (ECIR 2006)*, pages 539–542. Springer, 2006.

[8] R. Chow, P. Golle, and J. Staddon. Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 893–901, 2008.

[9] J. DeShazo, D. LaVallie, and F. Wolfe. Publication trends in the medical informatics literature: 20 years of 'medical informatics' in mesh. *BMC Medical Informatics and Decision Making*, 9(7):e13, 2009.

[10] P. Dixon. Medical identity theft: The information crime that can kill you. The World Privacy Forum, 2006. http://www.worldprivacyforum.org/medicaliden-titytheft.html, retrieved June 7, 2009.

[11] B. Elger and A. Caplan. Consent and anonymization in research involving biobanks. *European Molecular Biology Organization reports*, 7(7):661–666, 2006.

[12] K. E. Emam, E. Neri, and E. Jonker. An evaluation of personal health information remnants in second hand personal computer disk drives. *Journal of Medical Internet Research*, 9(3):e24, 2007.

[13] A. Evfimievski, R. Fagin, and D. Woodruff. Epistemic privacy. In *Proceedings of the 27th ACM Symposium on Principles of Database Systems (PODS 2008)*, pages 171–180, 2008.

[14] J. Gayer. Policing privacy: Law enforcement's response to identity theft. CALPIRG Education Fund, 2003. http://www.calpirg.org/home/reports/reportarchives, retrieved June 7, 2009.

[15] F. Hecht and W. Shiel, editors. *Webster's New World Medical Dictionary*. Wiley Publishing, second edition, 2003.

[16] T. Herzog, F. Scheuren, and W. Winkler. *Data Quality and Record Linkage Techniques*. Springer, 2007.

[17] V. Jagannathan, C. Mullett, J. Arbogast, K. Halbritter, D. Yellapragada, S. Regulapati, and P. Bandaru. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *International Journal of Medical Informatics*, 78(4):284 – 291, 2008.

[18] E. Johnson. Data hemorrhages in the health-care sector. In *Financial Cryptography and Data Security*, 2009.

[19] R. Jones, R. Kumar, B. Pang, and A. Tomkins. Vanity fair: privacy in querylog bundles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM08)*, pages 853–862, 2008.

[20] J. Long. *No Tech Hacking: A Guide to Social Engineering, Dumpster Diving, and Shoulder Surfing*. Syngress Press, 2008.

[21] C. McGuigan and M. Browne. Hospital leak linked to witness in lvf case. Belfast Telegraph, 2007. http://www.belfasttelegraph.co.uk/sunday-life/news/hospital-leak-linked-to-witness-in-lvf-case-13904797.html , retrieved June 7, 2009.

[22] I. Neamatullah, M. Douglass, L. Lehman, A. Reisner, M. Villarroel, W. Long, P. Szolovits, G. Moody, R. Mark, and G. Clifford. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(32):e17, 2008.

[23] M. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2004.

[24] C. Preimesberger. Cyber-criminals use p2p tools for identity theft, security analyst warns. eWeek.com, 2006.

[25] M. Sokolova and K. El Emam. Evaluation of learning from screened positive examples. In *Proceedings of the 3rd workshop on Evaluation Methods for Machine Learning (EMML-ICML 2008)*, 2008.

[26] L. Sweeney. Protecting job seekers from identity theft. *IEEE Internet Computing*, 10(2):74–78, 2006.

[27] O. Uzuner, Y. Luo, and P. Szolovits. Evaluating the state-of-the-art in automatic de-indentification. *Journal of the American Medical Informatics Association*, 14:550–563, 2007.

[28] O. Uzuner, T. Sibanda, Y. Luo, and P. Szolovits. A de-identifier for medical discharge summaries. *Journal of Artificial Intelligence in Medicine*, 42:13–35, 2008.

[29] H. Ware, C. Mullett, and V. Jagannathan. Natural Language Processing (NLP) Framework to Assess Clinical Conditions. *Journal of the American Medical Informatics Association : JAMIA*, 16:585–589, 2009.