

Will HITECH Heal Patient Data Hemorrhages?

M. Eric Johnson
Center for Digital Strategies
Tuck School of Business at Dartmouth
m.eric.johnson@dartmouth.edu

Nicholas Willey
Center for Digital Strategies
Tuck School of Business at Dartmouth

Abstract

Hemorrhages of confidential patient health data create privacy and security concerns. While the US HIPAA legislation on privacy and security went into effect over five years ago, healthcare information security remains a significant concern as organizations migrate to electronic health records. The recent HITECH legislation aimed at accelerating this migration contained mandates for greater security, including the addition of new requirements on breach reporting. We overview this legislation and examine a recently collected sample of inadvertently disclosed files found on internet-based file-sharing networks. We characterize the security risk of these files and also present evidence of the threat by analyzing user-issued searches. Our analysis indicates that the threat and vulnerability for the healthcare sector continued, even as HITECH became effective.

1. Introduction

Inadvertent disclosures of private customer information have occurred in nearly every industry from banking to healthcare. Such leaks directly impact customers through embarrassment, fraud, and identity theft. In the healthcare sector, data hemorrhages have multiple consequences [1]. In some cases, the losses translate to privacy violations and social stigma. In other cases, criminals exploit the information to commit fraud or medical identity theft. The fragmented nature of the US healthcare system results in data hemorrhages from many different sources including acute-care hospitals, physician groups, ambulatory healthcare providers, medical laboratories, insurance carriers, back-offices of health maintenance organizations, and outsourced service providers such as billing, collection, and transcription firms.

In this paper, we examine the recent Health Information Technology for Economic and Clinical Health (HITECH) legislation and its potential impact on the security of protected health information (PHI). HITECH was enacted as part of the 2009 American Recovery and Reinvestment Act (ARRA) to spur the

adoption of Electronic Health Records (EHR). The act earmarked \$20 Billion dollars to be used as incentives and investments in the creation of a digital health information infrastructure. The act also followed up on earlier HIPAA legislation to enhance privacy and security rules, which became effective in 2003 and 2005 respectively. It expanded the breach notification process by extending the parties covered under HIPAA, i.e. care providers and insurers, to include their business associates. It also defined new conditions and penalties for noncompliance. According to US Health and Human Services (HHS) guidance, affected individuals must be notified that a breach has occurred within 60 days after the discovery of the breach. If the HIPAA party does not have contact information for an individual, then they must post the breach on their website or make media notifications (local newspaper, television station, etc.). If a breach affects more than 500 people, state media and government notifications are required.

We examine the impact of HITECH on availability of hemorrhaged PHI on internet-based file-sharing networks. In earlier work [1], we showed that inadvertent disclosures of medical information collected in 2008 made PHI readily available on P2P file-sharing networks. We found leaks throughout the health chain including care providers, laboratories, and financial partners. In one case involving an AIDS clinic in Chicago that was leaking patient data, impacted individuals who experienced fraud and social stigma have recently filed a class action lawsuit against the clinic [2]. In another study [3], estimated that 0.4% of Canadian and 0.5% of US IP addresses exposing documents and spreadsheets on P2P networks, leaked ones containing PHI. Their results show that, with tens of millions of simultaneous P2P users in North America, the exposed PHI at any given point of time is substantial.

We begin by surveying recently reported healthcare data losses and the fraud that data fuels. Next we briefly summarize HITECH and its implications for

data security. Then we turn to an analysis of inadvertent data hemorrhages. Our research objective was to investigate the availability of sensitive PHI leaked onto file-sharing networks. Given the September 23, 2009 effective date of the new HITECH breach reporting requirements, we sampled healthcare related files both before (August) and immediately after (October) the effective date to ascertain the availability of the leaked PHI. We present an analysis of thousands of files we found. These files were published in peer-to-peer file-sharing networks like Limewire, eDonkey, and Bearshare and could be downloaded by anyone searching for them. The files found included sensitive patient correspondence, business documents, and PHI-laden spreadsheets. We found files from healthcare firms that contained private employee and patient information for thousands of individuals, including addresses, Social Security Numbers, birth dates, and treatment information. We also found private patient information including medical diagnoses and psychiatric evaluations. Besides our analysis of files, we also present evidence from user-issued searches on these networks that demonstrate the threat to leaked medical data. We conclude with a brief discussion on reducing these inadvertent data hemorrhages.

2. Healthcare Data Losses

Data losses in the healthcare have become all too common. The Open Security Foundation [4], an organization that tracks data breaches worldwide, has documented 277 incidents of medical data loss since 2004. These 277 incidents include over 12 million individual records, and an average breach size of 46,105 records [5]. The high average size may be surprising, but is biased by the fact that only large breaches are typically reported—an issue that has changed under HITECH.

The majority of breaches publicly reported to the Open Security Foundation are the result of physical loss. Lost or stolen laptops, computers, disks, media, hard drives, tapes, and documents account for 156 of the 277 incidents. Many of these incidents (like a lost laptop) represent inadvertent disclosures, rather than technical hacks. Likewise, laptops that are stolen are often stolen for the laptop itself—not the data. However, the data is inadvertently disclosed.

For example, on August 25, 2009, a Blue Cross and Blue Shield laptop was stolen from an employee at the company's Chicago headquarters. The laptop contained the personal information of more than 800,000 physicians. To put that number in context, the 800,000 number is greater than the 732,000 practicing physicians in the United States in 2007 according to the AMA, meaning that nearly every physician's unencrypted personal information was available on this

single laptop [6]. This information could be used by criminals to facilitate false billing of Medicare and insurers, general identity theft, and other forms of fraud. Sensitive Information can also simply be lost by healthcare organizations. For example, in November 2009, insurance provider Health Net lost a single hard drive containing the personal information and medical records of 1.5 million members. The information on the hard drive dated as far back as 2002 and because customers could be potential victims of identity theft, Health Net has offered 2 years of identity protection to each customer [7]. The total cost to Health Net of providing this protection to 1.5 million customers at market rates of approximately \$10 per month is estimated to be \$360 million.

While physical loss may be the largest component of reported data breaches, a significant amount of data loss has occurred over the internet. In the open security foundation database there are 33 internet based breaches including email breaches, hacks, viruses, and web-based data losses. Web-based breaches are the most common type of internet breach in the database and in many cases are caused by unfamiliarity with IT systems and employee error that result in an inadvertent disclosure. For example, in April of 2008, it was reported that two improperly configured servers holding Wellpoint data, exposed the personal and medical information of nearly 130,000 individuals over the internet [8]. On one server, 1320 enrollees' data was so freely available that it had been indexed by search engines. In the case of the second server, a form of data protection was in place, but the protected files were left available for download for over a year. In this instance, the company contracted to maintain Wellpoint's data security was found to be at fault.

Even if the actual files on the computer are well protected sometimes, a virus spread over a network can still expose sensitive information. In July of 2009, the Alberta Health System was infected with the Coreflood virus [9]. The Coreflood virus works by capturing screen shots of documents as they are opened and then relaying the information to the website or computer of the virus originator. The virus in the Alberta Health System network transferred the information of an estimated 11,582 patients to its creator before finally being discovered 14 days later. The virus is believed to have infected the network via email. Of course, email is an important everyday part of medicine, but incidents like the Coreflood breach emphasize the dangers of email. In some cases, like in the October 2009 leak from Baptist Hospital East in Kentucky, employees themselves may inadvertently disclose private information by sharing data directly over email [10]. The hospital accidentally sent out a list of 350 employees' social security numbers to a large mailing

list, but had intended to simply notify nurses of the need to update their medical licenses.

Finally, employee fraud is also a significant contributing source of data loss. In April of 2009, Johns Hopkins Medicine reported that a patient registration secretary was suspected of operating as part of an identity theft scheme that had affected up as many as 47 people [11]. During employment at Johns Hopkins, the employee had accessed the personal information of over 10,000 people. This information included addresses, social security numbers, parents' names, date of birth, place of birth, and medical insurance information, enough to perpetrate fraud.

3. Background on HITECH

As part of HITECH's push towards the development of a digital health information infrastructure, the act created a new office—the National Coordinator for Health Information Technology, to oversee the transition to electronic health records, and enhance patient privacy. According to Congressional projections, the act should produce a rapid and expansive adoption of EHR technology, with 90% of doctors and 70% of hospitals employing the digital systems sometime in the next decade.

The act also offered powerful incentives to encourage doctors and hospitals to switch to EHR technology as soon as possible. Beginning in 2011, independent physicians can receive a yearly payment of as much as \$15,000. These incentive payments will decrease over time as the program is phased out over a maximum of 5 years of eligibility. According to the House Ways and Means committee, a Doctor qualifying for all incentives could be awarded as much as \$65,000 for adequately demonstrating meaningful use of EHR technology over a 5 year period [12]. Similarly hospitals can receive large incentive payments for converting to EHR technology. Hospitals demonstrating meaningful use of EHR will receive a base amount of \$2,000,000 for the initial year of EHR use. This amount will be adjusted upwards for hospital discharging more than 1150 patients per year (\$200 per patient), and scaled downwardly according to the fraction of total patients using the government run Medicare program. Like the payments to physicians, these incentives will be phased out over a 5 year period. Long term care or rehabilitation facilities, such as nursing homes and psychiatric hospitals, are not eligible for incentives under the stimulus act.

The act also follows up on the federal government's HIPAA legislation and outlines plans for required privacy and security controls on EHR systems. While electronic health records can streamline admittance, billing, and the administration of care, the increased accessibility creates new risks. In a system where almost all PHI is digital, patient vulnerability to

massive scams is a concern, thus driving stricter measures governing the transmission and storage of PHI. To address this issue, the initiative establishes protocols and certifications for health information technology products. Certification will be necessary for health providers to qualify for the aforementioned incentives, and the development of a certification program will be handled by the National Institute of Standards and Technology. Furthermore, Congress has established a protocol to be followed by health providers in the event of a PHI leak. This so-called breach notification process has varying response levels depending on the severity of the breach. Following the passage of the HITECH act, the Department of Health and Human Services outlined their requirements for breach notification [13]. The guidance lists the necessary steps to be followed in the event of a breach of *unsecured* public health information. The special definition given to "unsecured" by the HHS is information that is not secured with technology that renders PHI "unusable, unreadable, or indecipherable" to unauthorized individuals. In plain terms, EHR systems must use a form of encryption technology, and health practitioners must destroy unencrypted copies of health information after use. If health information is to be used for scientific purposes, only a "limited data set" of information relevant to the study is to be provided and such data must adequately obscure the identity of the patients. Furthermore, the Act extends the requirements on data security from parties covered under HIPAA, i.e. care providers and insurers, to also include their business associates. According to the guidance, affected individuals must be notified that a breach has occurred within 60 days after the discovery of the breach. If the HIPAA party does not have contact information for an individual then they must post the breach on their website or make media notifications (local newspaper, television station, etc.). If a breach is larger than 500 people, state media and government notifications are required. The Secretary of Health and Human Services will maintain a web hosted list of entities involved in a breach affecting more than 500 people.

Groups covered under HITECH must also abide by several other new requirements [14]:

- Must honor an individual's request that information be withheld from health plan providers if care is paid for in cash.
- Must be capable of providing a 3-year audit trail of patient health information disclosures upon request.
- May not communicate with patients for marketing purposes resulting in monetary gain without the permission of the patient.

Additionally, HITECH increases the severity of HIPAA fines. The federal government imposed monetary penalties for both inadvertent and willful disclosure of unsecured patient information. The penalties under HITECH increase with the severity of the violation, ranging from \$100 to \$1.5 million [15].

Much of the debate around EHR and government incentives to adopt new systems have hinged around the words “meaningful use.” Providers must be able to demonstrate meaningful use to receive incentive payments. On January 13, 2010 HHS issued a Notice of Proposed Rulemaking (NPRM) to implement provisions that provide incentive payments for the meaningful use of certified HER [16,17]. The proposed rule included the definition of meaningful use following a three-stage definition phased in over time, along with other requirements for qualifying for incentive payments. That rule was finalized in July 2010 [18]. HHS maintains a website at healthit.hhs.gov, which provides funding announcements, tools, and information on health information technology. The Office has published an implementation plan which outlines the future goals of the office and its strategy for meeting its obligations under the Recovery Act [19].

The success of the HITECH Act will be highly dependent upon the attractiveness of the incentives to healthcare providers and their willingness to transition to digital technologies. The in-trenched familiar and reliable paper systems supported by ad hoc digital systems like Excel spreadsheets will likely take some time to disappear due to resilient legacy use. For hospitals working with independent physicians, this likely means maintaining both legacy electronic and paper systems with new EHR during the transition phase, which could lessen the economic benefits that the recovery act hopes for. Furthermore, the government will need to work diligently to make sure that the sensitive data in transition into EHRs remain secured. Firms across all business sectors struggle with data security problems and it is unlikely that there is a prescribable out-of-box solution that will work for all parties handling EHR. Any platform that does become widely adopted will become a larger and larger target for parties seeking to exploit EHRs for personal and financial gains. In the end, the hope is that health professionals will respond to the offered incentives and that the HITECH Act will make health care faster, less expensive, and higher quality. The focus of this paper is security during the early transition as HITECH security rules became effective.

4. Research Method - Data Leaks

To examine the challenges of protecting ad hoc electronic health information, we examine inadvertent disclosures during the HITECH transition period of

July-October 2009. We characterize the extent of medical information leakage by analyzing healthcare related disclosures and search activity in peer-to-peer file-sharing networks. To collect a sample of relevant, leaked files, we developed a digital footprint of promising healthcare industry-related terms. Using National Institute of Health (NIH) data on research funding [20] to rank medical research institutions in the United States, we selected the top 25 institutions (Figure 1). The ranking is based simply on funding from the listing of NIH grants, contracts, and agreements with non profits, research institutions, medical schools, independent hospitals, higher education, and individuals.

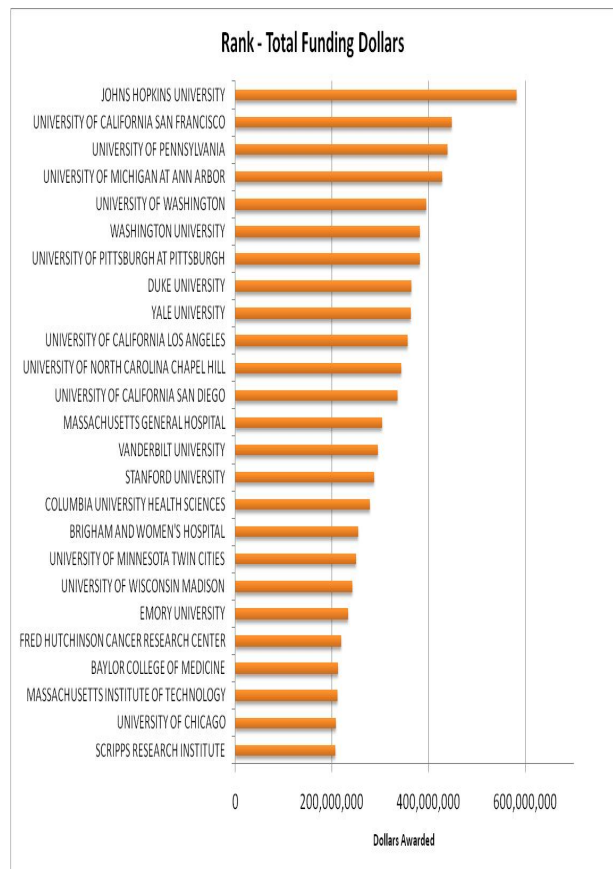


Figure 1. Institution ranking.

Rather than focusing on those institutions, we used them as inspiration to develop a list of searchable terms for healthcare. Thus, for each institution, we developed a set (average of 52) of terms and phrases related to that institution and medical specialties. For example, for John Hopkins this would include terms like JHM Institute for Basic Biomedical Science, Hopkins Institute for Cell Engineering, Hopkins Institute for Computational Medicine Hopkins Institute of Genetic

Medicine, Welch Center Biostatistics Center, Johns Hopkins Bayview Medical Center, Middle Atlantic Spectrometry Lab, Wilmer Eye Institute, Blaustein Pain Center, Brady Urological Institute, Sidney Kimmel Cancer Center, etc. Notice that these included the names of many clinics, research centers, and specialties (over 1250 in total across the top 25 institutions). Those terms along with many generic ones (medical, hospital, health...) were used to search for files in the major P2P networks. With the help of Tiversa Inc., we used these terms to search the four most popular networks (each of which supports the most popular clients) including Gnutella (e.g., Limewire, Frostwire, BearShare), FastTrack (e.g., KaZaA, Grokster), Aries (Aries Galaxy), and e-donkey (e.g., eMule, EDonkey2K). We conducted our first search over a 2-week period in July 13-27, 2009. Files containing any one or combination of these terms in our digital footprint were captured along with related user-issued searches in those networks. We focused on text files, files from the Microsoft Office Suite (Word, Powerpoint, Excel, and Access) and Adobe (pdf). Our goal was to gather a sample of files to characterize the ongoing data hemorrhage. Since users randomly join P2P networks to find and share media (and then depart), the network is constantly changing. The dynamic nature of P2P networks with many different networks, clients, and users makes it difficult to estimate the exact population size in aggregate or at any particular moment. Estimates have placed the P2P population at around 10 million simultaneous users [21]. New clients are constantly being introduced while others have been shut down through legal action. Using Tiversa's systems, we participated in those networks over a two-week period and collected a very large sample of this activity – a total of 7,911 files.

Given our search approach, we often captured files that were not remotely relevant to our search (in this case, having nothing to do with healthcare) or were duplicates. After an initial analysis to remove irrelevant and duplicate files, we arrived at 2,966 files for further (automated and manual) analysis.

As one might suspect, the majority of medical-related files available on P2P networks fall into the educational category (publically available health-education materials) along with reports and journal articles. This is consistent with earlier results and the fact that many P2P users are students. We also found a range of files related to organizational operation such as billings, insurance claims, marketing documents, and legal forms. Likewise we found many HR related documents including job listings, cover letters, and résumés. Of course, we did find a number of files that posed risks to organizations and patients, including many with PHI. For example, we found nursing notes (see Figure 2), medical histories, patient diagnoses,

psychiatric evaluations, letters to patients, and spreadsheets with patient data. Each file was rated on a simple 3-point scale (see Table 1), with 0 representing low risk, 1-medium risk, 2-high risk (ones with PHI including such identifying information as name, address date of birth, social security numbers, insurance numbers, and health related information). Of course a scale with five- or seven-points would permit greater discrimination. In practice, however, we found we could not further distinguish between the files (without further file information). Thus, for our study, a more finely grained scale would increase the scale's variance through the induction of random noise rather than a systematic variance attributable to the underlying risk phenomenon [22]. Ratings were conducted by a single researcher with a second-party review of the ratings. As such, inter-rating reliability was not an issue.

Table 1. File Risk Rating Scale

Level	Definition
High 2	<ul style="list-style-type: none"> Substantial PHI such as patient evaluation and diagnosis and significant (multiple) identifying information such as address, date of birth, social security numbers, insurance numbers, and health related information. Marked as confidential.
Medium 1	<ul style="list-style-type: none"> PHI such as patient evaluation and diagnosis and with limited identifying information (name and address). Letters and forms with limited PHI.
Low 0	<ul style="list-style-type: none"> Information that is commonly shared with others in course of business but not with the general public (and is therefore quasi-public) –e.g., resumes, cover letters, forms, sales presentations. Public information. Education material.

After categorizing the files, we found that about 15% of the relevant files posed some risk (a 1 or 2 on our scale) and 8% had significant PHI (a 2 on our scale). The files we found illustrate that PHI is often electronically stored outside of enterprise-class EHR in ad hoc file formats like documents and spreadsheets that are vulnerable to inadvertent disclosure.

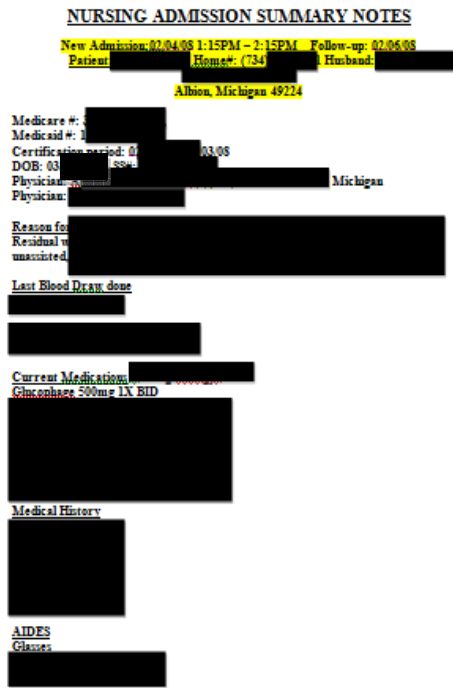


Figure 2. Redacted example of document with PHI.

5. Threat

We have shown that sensitive medical information is available on P2P networks. In addition, we have found that there are users who seek out these documents. Using the same set of terms, we were able to identify thousands of suspicious incoming searches.

We collected about 417,000 searches with nearly 125,000 related to medicine and healthcare (Figure 3). Those user-issued searches were also categorized by discernable threat level on a simple four-point scale. Of these searches, 102,000 searches were not considered to have a discernable malicious intent. Searches like “medical”(17,993 searches), “dna”(12,983), “eating disorders”(2380), “heart care” (2395), “health secrets”(378), “medicine”(8329) may be malicious, but are more often intended to recover general information rather than sensitive documents. While these terms could relate to medical data, it seems more likely that the intent of the searcher was benign. For example, “medicine” could be a search for the film “Medicine Man”, or the TV show “Dr. Quinn Medicine Woman”, or the Bon Jovi song “bad medicine”, the punk rock band “Medicine”, or songs by Kim Leoni, Three Six Mafia, and Guster that are all titled “medicine.”

A threat rating of 1 was given to searches that were more specific and likely to return sensitive data, even though the searcher may have not had a malicious

intent. Searches falling under this rating include: “medical school” (6,930 searches), “aids research” (389), “medical center” (371), “autism research”, “cancer center” (309 searches), and “autism research” (15). A threat rating of 2 was given to searches that were highly suspicious and very likely to be directed at recovering sensitive documents. The searches “university research park” (82 searches), “hiv center” (45), “cytology core” (7), “broad institute”(2), and “Columbia center for aids research” (4), are search terms that users are unlikely to enter into a P2P network for finding health information (keep in mind, P2P is primarily for sharing media). Searches that were clearly intended to recover confidential research documents and other sensitive files were given the highest threat rating of 3. The searches, “Public health passwords”(79 searches), “hiv diagnosis” (72), “breen lab” (42), roche labs (188), pharmacogenetics proprietary (17), “confidential embryonic stem cell 2009”(1), spctrm internal (1), and “pittsburgh cancer institute files” (2), are all intended to gain access to either patient data or confidential research. Several research labs at major institutions were targeted by name including, the Breen lab (42) at North Carolina state university, the Moore lab (19), Keating lab (5), and Sauer Lab (2) at MIT/Scripps, Cooper lab (17) at Washington University, Roberts lab (5) and Chu Chen Lab (3) at Fred Hutchinson Cancer Research Center, and the Kemp Lab (3) at Georgia Tech. This type of sophisticated searching indicates that P2P users are looking to do more than commit health care fraud, and are also actively seeking out confidential research findings and information.

Unsurprisingly, most of the collected search terms were intended to return materials commonly shared over P2P networks, pornography (52,153 searches), audio files (37,041), and videos (49,990). Searches like “eminem medicine ball” (22,623 Searches), “teens” (24,938), and “the matrix” (21,093), overlapped with our focus on searches related to medical data but were clearly not related to healthcare.

In addition to medical searches, we also uncovered suspicious searches related to the federal government, educational institutions, and corporations. Searches like “Columbia 2009 my documents” (1 search), “mit passwords” (39), and “confidential university” (291), are a clear indication that there are users seeking to gain access to documents from university networks. Furthermore, searches such as “Washington blue prints 228397”(1 search), “Washington classified” (39), “Langley building docs” (4), and “lincoln lab” (421), are unsettling from a national security standpoint. Since this data was collected inadvertently (because they overlap with terms and phrases were watch looking for), the actual threat is likely to be much larger than our data indicate. Overall, 30% of the user-

issued searches we captured (based on our terms) appeared to have a medical-related intent. Of those, nearly 18% appeared to have some level of threat with about 1% representing highly focused searches (a three on our scale).

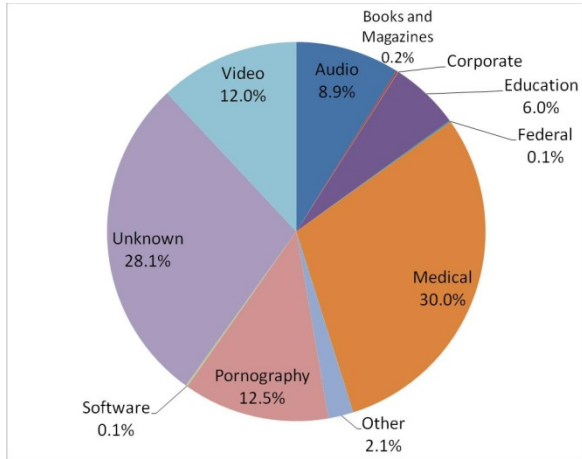


Figure 3. Categorization of user-issued searches.

In conclusion, the collected search data expose the prevalence of malicious medical-related searches on P2P networks. We have found that users have developed targeted search terms for flushing out patient data files and research findings. When this information is combined with the numerous documents collected as part of our experiments we conclude that users are not only searching for sensitive files, but also finding them.

6. Follow-up on Business-Related Spreadsheets

On August 24, 2009 US HHS published final guidance related to HITECH breach notification for unsecured protected health information [23]. Those rules applied to breaches of protected health information occurring on or after September 23, 2009. Under HITECH, within 60 days after the discovery of the breach, organizations are required to provide notification to the affected individuals. If the individuals' contact information is not available, then they must post the breach on their website or make media notifications (local newspaper, television station, etc.). Moreover, breaches greater than 500 people also require state media and government notifications. While these requirements went into effect in September, HHS provided some relief from full enforcement until February 22, 2010 to adapt to the new rules. Signaling an end to the enforcement delay, on February 23, 2010 HHS's Office of Civil Rights posted a list of over 60 organizations that reported breaches of unsecured PHI affecting 500 or more individuals on OCR's website [24].

On the eve of the effective date (September 23), we reinitiated a two-week sampling of files on P2P networks, using the search terms from the August collection. However, this time we focused exclusively on excel spreadsheets that might contain significant PHI (files with the xls extension). Over that period we collected 3,766 spreadsheets, of which 788 were unique, relevant files. Of those relevant files, 45% contained information that held some risk for organizations or individuals and 2.5% represented significant risk. For example, we found spreadsheets showing medical settlements that included the individuals' names, addresses, DOBs, SSNs, phone numbers, employers, insurance information, and the amount of the financial settlements. Others included medical forms and reports with PHI or healthcare employee information. Five of the files we found appeared to qualify as major breaches under the new HITECH rules—that is they contained significant PHI for more than 500 individuals. For example, we found one detailed monthly case logs on several hundred mental health patients over a two-year period. Another contained insurance information for over 7,000 individuals, including personally identifying information, their physicians, and dates of service (see example in Figure 4). Yet another more extensive spreadsheet included similar information on over



Figure 4. Redacted spreadsheet with PHI.

16,000 patients, but also included employer information and diagnosis codes for each patient. Together the five files contained sensitive PHI on over 28,000 individuals.

7. Discussion

Certainly, we did not expect the effective date of HITECH would result in the elimination of inadvertent disclosures of PHI. In fact, given the nature of P2P networks, leaked files live-on within the network long after the original leak source is closed. We are encouraged to report that we did not find any research-related PHI (PHI disclosed as part of a research activity). Documents and studies related to research that we did see had replaced patient ID with tokens, anonymizing the data. However, our results show that significant PHI leaks continue from other areas within the health supply chain, placing organizations and patients at risk. Moreover, it is important to note the results of this effort show that the root problem is not simply P2P networks. The files we found in P2P

networks are the same ones that would be disclosed with a lost laptop, CD, or flash memory. Peer-to-peer networks simply provide a window into the types of data that can easily be inadvertently disclosed. Thus the solutions to the data hemorrhage must go beyond blocking P2P networks.

There are indeed many measures that the healthcare sector should consider, beyond blocking P2P networks and preventing files from migrating to machines that participate in P2P. While none themselves represent the single answer, efforts such as P2P monitoring, disk-level encryption, tokenization, and data truncation all help. More importantly, moving sensitive material out of ad hoc databases such as spreadsheets and documents and into enterprise-class software will likely reduce the types of inadvertent disclosure we observed.

As compared with earlier studies we conducted in the banking sector (Johnson 2008), we note that the extended enterprises of healthcare providers often include many technically unsophisticated partners who are more likely to leak information. Thus tracking and

stopping medical data hemorrhages is more complex and possibly harder to control given the fragmented nature of the US healthcare system. However, efforts to move PHI out of ad-hoc files (like spreadsheets, simple databases, and word processing) and into better-managed EHR should reduce the inadvertent disclosures we document in this paper.

8. Acknowledgements

This research was partially supported by the National Science Foundation, Grant Award Number CNS-0910842, under the auspices of the Institute for Security, Technology, and Society (ISTS). Experiments described in this paper were conducted in collaboration with Tiversa who has developed a patented technology that, in real-time, monitors global P2P file-sharing networks

9. References

- [1] Johnson, M. Eric (2009) "Data Hemorrhages in the Health-Care Sector," *Lecture Notes in Computer Science*, R. Dingledine and P. Golle (Eds.): FC 2009, LNCS 5628, 71–89, ICFA/Springer-Verlag Berlin Heidelberg.
- [2] "John Doe et al. vs. Open Door Clinic of Greater Elgin, an Illinois Corporation" 2010.
- [3] El Emam, Khaled, Emilio Neri, Elizabeth Jonker, Marina Sokolova, Liam Peyton, Angelica Neisa, Teresa Scasa (2010), "The Inadvertent Disclosure of Personal Health Information through Peer-to-peer File Sharing Programs." *Journal of the American Medical Informatics Association*, 17: 148-158.
- [4] <http://www.opensecurityfoundation.org/>
<http://datalossdb.org/>
- [5] Through September 2009. <http://datalossdb.org/>
- [6] "Laptop Theft Nets Data on 800,000 Physicians" Information Week, Thomas Claburn, October 15, 2009, Source: <http://www.informationweek.com/news/healthcare/security-privacy/showArticle.jhtml?articleID=220601030> Retrieved: December 7, 2009
- [7] "1.5 Million Medical Records at Risk in Data Breach" The Hartford Courant, Matthew Sturdevant November 19, 2009. Source: http://www.courant.com/health/hc-healthbreach1119.artnov19_0.1798384.story Retrieved: November 28, 2009
- [8] "Wellpoint Customer Information Exposed", Associated Press, Tom Murphy, April 8, 2008, Source: <http://attrition.org/dataloss/2008/04/wellpoint01.html> Retrieved :December 8, 2009
- [9] "Security of health information like 'Fort Knox': doctor" CBC News, July 9, 2009 Source: <http://www.cbc.ca/canada/edmonton/story/2009/07/09/edmonton-virus-ahs.html> Retrieved: December 9, 2009
- [10] "Email leaks 350 Baptist East employee Social Security numbers", WHAS11, Rachel Nix. October 26, 2009. Source: <http://www.whas11.com/news/consumer/Email-leaks-350-Baptist-East-employee-Social-Security-numbers-66250142.html> Retrieved: December 10, 2009
- [11] Letter: "Re: Possible Compromise of Personal Information", Johns Hopkins Medicine, Donald L Bradfield Senior Counsel, April 3, 2009. Source: http://datalossdb.org/primary_sources/0000/1552/ITU-168293.pdf Retrieved: December 11, 2009
- [12] Title IV Health Information Technology for Economic and Clinical Health. Majority Staff of the Committees on Energy and Commerce, Ways and Means, and Science and Technology. January 16, 2009 Source: <http://waysandmeans.house.gov/media/pdf/110/hit2.pdf> Retrieved: July 1, 2009
- [13] Guidance Specifying the Technologies and Methodologies That Render Protected Health Information Unusable, Unreadable, or Indecipherable to Unauthorized Individuals for Purposes of the Breach Notification Requirements under Section 13402 of Title XIII (Health Information Technology for Economic and Clinical Health Act) of the American Recovery and Reinvestment Act of 2009; Request for Information. Office of the Secretary, Department of Health and Human Services. April 17, 2009 Source: http://www.hhs.gov/ocr/privacy/hipaa/understanding/covered_entities/hitechrfi.pdf Retrieved: February 1, 2010
- [14] The Health Information Technology for Economic and Clinical Health Act (HITECH Act): implications for the adoption of health information technology, HIPAA, and privacy and security issues. Linn Foster Freedman, Nixon Peabody. February 23, 2009. Source: http://www.nixonpeabody.com/publications_detail3.asp?ID=2621 Retrieved: July 1, 2009.
- [15] Federal Register (2009) / Vol. 74, No. 209 / Friday, October 30, 2009 / Rules and Regulations, 56123- 56131. <http://www.hhs.gov/ocr/privacy/hipaa/administrative/enforcementrule/enfifr.pdf> Retrieved: February 1, 2010.
- [16] Federal Register (2010) "Medicare and Medicaid Programs; Electronic Health Record Incentive Program; Proposed Rule," Vol. 75, No. 8 / Wednesday, January 13, 2010 / Proposed Rules, 1844-2011. <http://edocket.access.gpo.gov/2010/pdf/E9-31217.pdf> Retrieved: February 27, 2010.
- [17] Federal Register (2010) "Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology; Interim Final Rule," Vol. 75, No. 8 / Wednesday, January 13, 2010 / Rules and Regulations, 2014- 2047. <http://edocket.access.gpo.gov/2010/pdf/E9-31216.pdf> Retrieved: February 27, 2010.
- [18] Federal Register (2010) "Medicare and Medicaid Programs; Electronic Health Record Incentive Program; Final Rule," Vol. 75, No. 144 / Wednesday, July 28, 2010 / Rules and Regulations 44314- 44588. <http://edocket.access.gpo.gov/2010/pdf/2010-17207.pdf>
- [19] Health Information Technology American Recovery and Reinvestment Act (Recovery Act) Implementation Plan. Office of the National Coordinator for Health Information Technology. Source: http://www.hhs.gov/recovery/reports/plans/one_hit.pdf Retrieved: July 5, 2009.

[20] Source:

<http://report.nih.gov/award/trends/AggregateData.cfm?Year=2008>

[21] Mennecke, T. (2006), "Slyck News – P2P Population Continues Climb" June 14, <http://www.slyck.com/news.php?story=1220>.

[22] DeVellis, R. F. (2003), Scale Development: Theory and Applications, Second Edition, Sage Publications, London.

[23] Federal Register (2009) "Breach Notification for Unsecured Protected Health Information; Interim Final Rule," Vol. 74, No. 162, August 24, 2009/ Rules and Regulations, 42740-42770. <http://edocket.access.gpo.gov/2009/pdf/E9-20169.pdf>
Retrieved: February 2, 2010

[24] Office for Civil Rights HHS (2010), "Breaches Affecting 500 or More Individuals," <http://www.hhs.gov/ocr/office/index.html>.
Retrieved: April 27, 2010