Encryption and the Loss of Patient Data

Encryption and the Loss of Patient Data

Amalia R. Miller* and Catherine E. Tucker‡

March 14, 2011

Abstract

Fast-paced IT advances have made it increasingly possible and useful for firms to collect data on their customers on an unprecedented scale. One downside of this is that firms can experience negative publicity and financial damage if their data are breached. This is particularly the case in the medical sector, where we find empirical evidence that increased digitization of patient data is associated with more data breaches. The encryption of customer data is often presented as a potential solution, because encryption acts as a disincentive for potential malicious hackers, and can minimize the risk of breached data being put to malicious use. However, encryption both requires careful data management policies to be successful and does not ward off the insider threat. Indeed, we find no empirical evidence of a decrease in publicized instances of data loss associated with the use of encryption. Instead, there are actually increases in the cases of publicized data loss due to internal fraud or loss of computer equipment.

*Economics Department, University of Virginia, Charlottesville, VA

†MIT Sloan School of Management, MIT, Cambridge, MA.

‡We thank HIMSS for providing the data used in this study and seminar participants at the Workshop on the Economics of Information Security 2010 for helpful comments. All errors are our own.

1 Introduction

Fast-paced IT advances have made it increasingly possible and useful for medical providers to collect patient data on an unprecedented scale, to improve both the diagnosis and treatment of medical conditions and the billing of insurers. However, collecting so much data is not risk-free. For example, Troy Beaumont Hospital in Detroit experienced a severe data breach when a laptop was stolen in August 2006, from the rear of a vehicle belonging to a nurse. This laptop documented the names, addresses, social security numbers, patient care details and insurance information for 28,400 patients. Large-scale losses like this are not unusual and they can have serious consequences for firms both in and outside the health sector. Cavusoglu et al. (2004) analyzed 225 security breaches and found that security breaches of firm data were associated on average with a loss of 2.1 percent of the firm's market value, or around $1.65 billion of market capitalization, within two days of the announcement. Further, 31 percent of surveyed consumers claim that they will end their relationship with a company if they are affected by a breach (Ponemon, 2008). There are also serious consequences for consumers of such instances of data loss including fraud and identity theft, leading governmental policy to take an increasingly activist stance to try to prevent consumer data losses.

Security experts and policy makers often encourage firms to adopt and use encryption software to minimize the risks of losing customer data. Encryption is a way to encode computer files so that only someone with access to a secret 'key' can read them. Theoretically, encrypting data should deter malicious hackers, because it makes the data difficult to read. Encryption should also minimize the risks of data being used maliciously if the data fall into the wrong hands. This paper presents some of the first empirical evidence about the extent to which firm adoption of encryption software limits how likely firms are to experience publicized instances of customer data loss. We focus on the health sector because that sector uniquely provides data on whether hospitals have adopted encryption software over time, as well as data about firm characteristics. This is also a sector where evidence has been

mounting for the need to secure patient data better. For example, a recent report found that health organizations may have to spend $834.3 million in total costs to address violations of the Health Insurance Portability and Accountability Act (HIPAA) in 2009 (Nicastro, 2010). Further, we find empirical evidence in this paper that increased digitization of patient data is associate with data breaches.

Surprisingly, we find empirical evidence that the use of encryption software does not reduce overall instances of publicized data loss. Instead, its installation is associated with an increase in the likelihood of publicized data loss due to fraud or loss of computer equipment. Speculatively, this may occur because firms are less careful at controlling access internally to encrypted data, and also because employees are less careful with computer equipment when they believe that data are encrypted. This reflects earlier findings that human error, rather than malicious external hackers, is often responsible for data loss: Ponemon (2009) finds that 88% of data breaches in 2008 could be traced back to insider negligence. The Troy Beaumont Hospital case in the first paragraph shows how human carelessness can undermine encryption: The nurse had kept the username and password for the encryption algorithm along with the stolen computer, rendering the encryption worthless.

One issue with positing a causal relationship between the adoption of encryption software and a firm experiencing publicized data loss is that there may be unobserved heterogeneity (such as the unobserved desirability of the data collected) that may lead both to higher instances of data loss and greater adoption of encryption software. To address this, we estimate jointly the likelihood of a data loss and the adoption of encryption software, treating the adoption of encryption software as an endogenous binary variable. As a source of exogenous variation that drives the adoption of encryption software but not the loss of data, we use whether or not the state's breach notification law makes an exception for encrypted data.

Many states have enacted general regulations that require all firms in all sectors to notify customers if their data is breached. However, some of these states give a blanket exception or

'safe harbor' if the breached data were encrypted. A state-wide encryption exception should give some incremental incentive to hospitals in that state to adopt encryption software, compared to hospitals in states that do not have an encryption exception. We use state-level fixed effects to control for baseline differences in states' propensities to use data, and we control for the effect on data breaches of the passing of any data breach notification law. Therefore, our identifying assumption is that there was no unobserved change in the average hospital's propensity to lose data that occurred at the same time as the passing of a data breach notification law that had an encryption exception, compared to states with a data breach law with no encryption exception.

When we control for the endogeneity of the adoption of encryption software in this manner, adopting encryption software is still positively associated with a greater likelihood of data loss. One concern is that the enactment of encryption exceptions may lead to an underreporting of cases of data breaches if hospitals use encryption because they are then not obliged to report them. However, there is a positive correlation between encryption exceptions and the likelihood of a data breach being publicized in the data. Another concern is that the enactment of a data breach law may make it easier for volunteers and journalists to find out about a data breach, as the law may require the hospital to report the breach publicly on a website. However, we show that our result holds when we exclude data breaches that were discovered because they were publicly reported in this manner. To further support our identification arguments, we perform a falsification check. We show that there is no relative boost in encryption adoption for states that give safe harbor to encrypted data but who explicitly exclude hospitals from their data breach laws. This check reassures that there is not something unobserved about the kind of states that put in exceptions to their data breach notification laws which may also be associated with encryption adoption and data loss.

1.1 Does the loss of encrypted data matter?

Why does it matter if the adoption of encryption software is associated with an increase in data loss, if encryption makes the lost data useless anyway? If only unreadable data are lost, it is not clear whether an increased likelihood of data loss poses a security risk to firms. However, losing encrypted data may still harm firms in three ways. First, the loss of encrypted data may not be harmless. When data are encrypted, users generally access the data either via a separate key on a USB drive or password. Getgen (2009) shows that, as happened in the Troy Beaumont hospital example, keys can easily be lost or compromised. Their study shows that 8% of organizations (including those who have not had a security breach) have experienced problems with a lost encryption key over the last two years. Second, our finding that the adoption of encryption software is associated with an increase in instances of fraud emphasizes that encryption software is not effective at preventing insiders from accessing readable data and using it in a harmful way. Third, there are many instances where firms encrypt some data, but leave other data un-encrypted, and also instances where employees de-encrypt data and download it to laptops or other unsecured portable devices.

The findings of the paper matter because government policies and industry best practices often appear to present encryption as an all-encompassing solution to data security problems. Representatives of the security industry such as Warmenhoven (2006) have argued that data exceptions in data breach notification laws need to distinguish between 'companies that lose data useful to thieves and those that lose data rendered useless by encryption' since 'a thief in possession of encrypted data has stolen little more than an empty container.' Critics of exceptions, such as Schuman (2009), have argued however that such blanket exceptions are 'ludicrous' given the possibility that the encryption key could be intercepted or cracked. In general, encryption only works as well as the organization's ability to use a strong en-

cryption algorithm and protect the password or key to that algorithm. Our results suggest that rather than blanket exceptions, a broader set of policies is warranted, that encompasses training and awareness programs, manual procedures and controls, and strong identity and access-management deployments. In particular, the fact that encryption software adoption is associated with an increase in fraud may suggest that firms deploying encryption software do not also deploy effective data access controls. Andrews (2010) points out that one of the 'biggest internal vulnerabilities' is 'misuse of privilege' by hospital personnel. This vulnerability appears to not just be limited to the healthcare sector. For example, the mortgage firm Countrywide emphasized their use of encryption and access controls in their website privacy and security policies. However, these encryption techniques were not enough to prevent a Countrywide employee from 2006-2008 from downloading records on up to two million customers/prospects to sell to mortgage brokers who wanted them for sales leads.[1]

From a government policy standpoint, our findings matter because 'safe harbors' for encryption are at the heart of the recently proposed federal 'Data Breach Notification Act' (Senate Bill 139). The overall efficacy of data breach notification has been under question since Romanosky et al. (2008) found only weak effects from breach notification laws on the number of identity theft cases in that state. We emphasize that if federal or state laws give safe harbor to all encrypted data, this may lead firms to focus on encryption and this may be to the detriment of focusing on controlling internal access to data and employee caution when managing personal data. In other words, by promoting a technological solution in isolation, and not also promoting human-based firm processes which complement encryption's effectiveness, giving a safe harbor to encrypted data may not have the intended effect.

We also find that large hospitals are more likely to lose data. This is understandable, since they theoretically have more data to lose, but this finding does suggest that organiza-

[1] 'Security oversight may have enabled Countrywide breach' By Nancy Gohring, IDG News Service, 08-04-2008

tional or financial capacity is not sufficient to counter the underlying risk of data loss. Our finding emphasizes the need for public policies regarding data-security issues to cover all organizations, since size is not sufficient to ensure that data are safeguarded appropriately.

The empirical findings of this paper also suggest that digitization of patient records may increase the likelihood of data breaches. This supports the fact that federal policy encouraging the digitization of patient data, such as the 2009 HITECH Act, also addresses issues of data breaches and patient protection. The paper also finds that this is is primarily a function of the extent to which the hospital uses electronic systems that make it easy to consolidate data about a single patient. Therefore, health data security policy may want to focus on ensuring that these kind of organizational master keys have appropriate protections and safeguards built into the hospital's system. In particular, our results suggest that prior to adopting EMRs hospitals must address both the insider threat and ensure that encryption policies are both comprehensive and universally applied in reality.

The findings of this paper contribute to a small empirical literature that has focused on the consequences of customer data loss for firms. Generally, on the firm side, research has focused on the stock market impact of the announcement of a security breach, finding large effects in empirical event studies (Cavusoglu et al., 2004; Acquisti et al., 2006; Telang and Wattel, 2007; Gaudin, 2007; Goel and Shawky, 2009). This research also builds on a theoretical literature that has emphasized the role of coordination failure in explaining data breaches. Since early research such as MacKie-Mason and Varian (1996), most research has presented encryption as a positive measure that firms can take against the security risks inherent in electronic data. Anderson and Moore (2006) summarizes the complex relationship between information security, moral hazard and coordination failure. Roberds and Schreft (2009) find that a lack of coordination across firms leads to too much data collection and too little security. Gal-Or and Ghose (2005) find that firms have sub-optimal incentives to share information about security failures with each other. The importance of employee compliance

for data security and the difficulty of giving correct incentives has also been emphasized by work such as Bulgurcu et al. (2010). We add to this literature by focusing on empirical evidence in the healthcare sector and presenting new evidence about the robustness of a commonly used security software tool.

2 Data

The paper uses four sources of data for the empirical analysis. We describe each in turn: (1) Data on Security Breaches (2) Data on Hospitals (3) Data on Hospital IT systems (4) Data on State Regulation.

2.1 Data on Security Breaches

The analysis uses data from 2005-2008 on publicized security breaches within the US. These data were collected by the 'Open Security Foundation.' OSF collects information on security breaches by monitoring news feeds about security breaches and by submitting Freedom of Information Act Requests for breach information that is collected by state governments.[2]

Table 1 summarizes the relative rates of different breach types for the medical sector that we study in this paper relative to non-medical businesses that also experienced data breaches in the OSF data. These categories reflect the way the OSF volunteer categorized the breach into the most appropriate of the different groupings. Where there was some overlap, they chose the category that seemed most appropriate. We verified these categorizations by cross-referencing the news story to the record in multiple cases and found no inaccuracies or inconsistencies. Data breach due to the loss, misplacement or improper disposal of equipment is relatively more common in the medical sector. This is unsurprising given that a third of health care professionals store patient data on laptops, smartphones and USB memory sticks (Dolan, 2010). A similar share of data breaches in non-medical and medical sectors is due to

[2]More information about the Open Security Foundation can be found at http:www.opensecurityfoundation.org.

Table 1: Relative rates of Different Breach Types: Medical and Non-Medical

Data Breach Type	Non-Medical	Medical	Difference	T-test
Equipment Loss	0.26	0.36	-0.10	-3.25
Theft	0.40	0.42	-0.02	-0.56
Fraud	0.14	0.20	-0.06	-2.67
Hack	0.20	0.02	0.18	7.75
Number of Instances	324	1196		

Source: Open Security Foundation

the theft of computer equipment (in the majority of cases a laptop). Data fraud represents a higher share of breaches in the medical sector, perhaps reflecting the increasing incidence of medical identity theft. Finally, data breaches due to 'hacking' are relatively rare in the medical sector, perhaps because of the relative lack of use of company websites and intranet sites to store data, which represented one of the largest sources of hacked data for the non-medical sector.

An obvious disadvantage of the breach incident data is that it is maintained and collected by volunteers rather than having being collected by a government body. In the US in the period that we study, there was no official central repository of information about data loss.[3] However, the distribution of different types of data breaches in the OSF database resembles statistics in the official government repository for the UK. The data from the UK are collected by the Information Commission as a consequence of the UK Data Protection Act. These data also emphasize the extent to which data are lost due to internal negligence or misconduct.[4]

Another consideration is completeness. We have information only on data breaches that were significant or newsworthy enough to have been picked up on by OSF volunteers. There

[3]This changed at the end of 2009, when under 13402(e)(4) of the HITECH Act, HIPAA was revised to require reporting of data breaches that affected more than 500 patients.

[4]The Deputy Information Commissioner, said 'Unacceptable amounts of data are being stolen, lost in transit or mislaid by staff. Far too much personal data is still being unnecessarily downloaded from secure servers on to unencrypted laptops, USB sticks, and other portable media.' 'Press Release', Information Commissioner's Office 11 Nov 2009.

may have been other instances of data loss that we have no way of finding out about. Therefore, all our estimates should be taken as reflecting correlations with a newsworthy data breach rather than any data breach. From a public relations and consequently a marketing and financial perspective, these are the data breaches that firms care about, so the conditional nature of the dependent variable is in line with the purpose of the study.

2.2 Data on Hospital IT systems

A major advantage to studying the hospital sector is that, almost uniquely, there are detailed data available about the IT systems that each hospital has adopted. We use these technology data from the past 4 years of releases of information from the Healthcare Information and Management Systems Society (HIMSS) Analytics^{TM} Database (HADB).

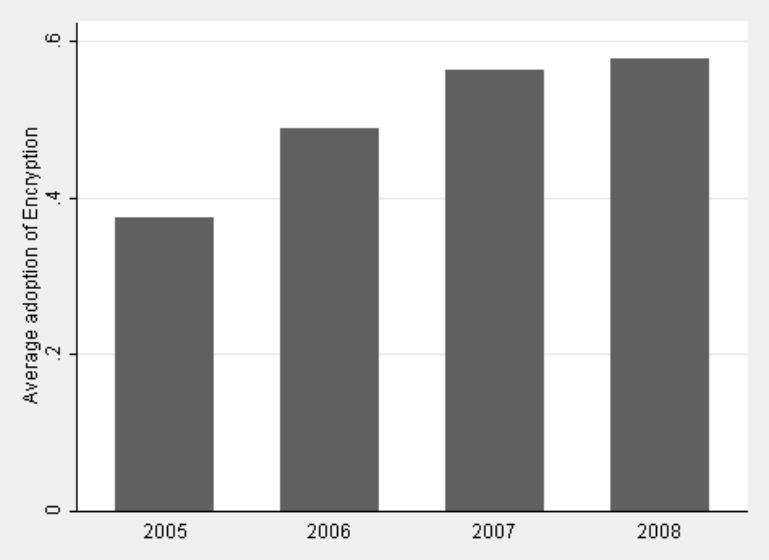
Encryption involves taking data and converting it into cipherdata using an encryption algorithm and an encryption key. Encrypted data is meaningless on its own without being deciphered with a key. The processes of encryption and decryption are customarily achieved using encryption software. The question in the survey only asked (under the heading of 'Security Systems') whether that hospital was currently using encryption. It did not ask how extensively encryption was used. It also did not distinguish between the use of encryption for data stored on disks or for communications.[5]

As shown by Figure 1, there was a substantial increase in adoption of encryption software over the period we study. The level of adoption in 2008, at 57%, is higher than in the non-medical sector. A recent survey (Getgen, 2009) of 655 IT professionals, found that on average 43% of businesses use database encryption.

We also collected data on adoption of six different IT and software systems: Firewall software, financial data warehousing, physician documentation software, clinical data repositories, clinical data warehouses, and enterprise master person index software. The last four

[5]Another limitation is the data are based on an annual survey so we do not know what month a hospital adopted encryption. To ensure that this does not lead to measurement error we show the robustness of our main results in Table A-3, to dropping years where the hospital first reported using encryption software.

Figure 1: Growth in use of encryption software



IT systems are crucial inputs of an electronic medical record (EMR) for a patient. EMRs have been the focus of public policy under the 2009 HITECH Act, which committed $19 billion to their promotion. Further, Section 3 of the Administrative Simplification Compliance Act (ASCA), Pub.L.107-105, requires that all all initial claims for reimbursement from Medicare be submitted electronically, with limited exceptions (42 CFR 424.32) However, the computerization of patient data has been the focus of both privacy and security concerns (Miller and Tucker, 2009). We largely use these IT systems as controls for the technological sophistication and inherent data risks of the organization in our regressions, but some of the raw correlations that we find are suggestive about which parts of EMRs are most vulnerable to security risks.

2.3 Data on State Regulation

In our empirical analysis, we both study the main effect of security breach notification laws on instances of data loss, and use blanket exceptions for encrypted data as a source of exogenous variation that can explain the adoption of encryption software. We collected data by studying the text of each law as listed in Alexander (2009) and cross-referencing this

with the National Conference of State Legislatures listing of laws. We used the text of these laws to distinguish whether or not that state has a blanket exception or 'safe harbor' for encryption.[6] We defined a law with a blanket encryption exception as being a law that allowed firms to not have to notify customers individually of breaches if the data involved in the breach were encrypted, regardless of whether the encryption key was compromised.

Table 2 summarizes the data we collected on laws. At the end of 2008, there were no state laws in Alaska, Alabama, Kentucky, Mississippi, Missouri, New Mexico, South Carolina or South Dakota. As shown in column (3) of Table 2, some states excluded organizations already covered by HIPAA (The Health Insurance Portability and Accountability Act of 1996 Privacy Rule). We use this variation when we perform a falsification check for the influence of encryption exceptions on adoption. There were also other data breach laws that excluded hospitals. Georgia's data breach law applied only to information brokers while Oklahoma's law only applied to state government organizations, so we count both states as having no law applying to the hospitals in our data.

These state laws set costs of data loss for hospitals above and beyond those imposed by the federal government in this period. HIPAA laid down various guidelines designed to protect the privacy of protected health information (see Miller and Tucker (2009) for a description) but it did not actually require notification in the event of security breaches. The final HIPAA Security Rule did not require encryption, but instead listed encryption as an addressable implementation specification. Hospitals were not forced to adopt encryption as a regular practice if their internal risk analyses did not justify it (Beaver and Herold, 2004).[7]

[6]Many law firms specializing in security laws have developed their own lists of laws which they share with clients. These lists seem frequently to be outdated and prone to error, so we examined the texts of the laws ourselves.

[7]45 CFR 164.312(a)(2)(iv) and 164.312(e)(2)(ii).

Table 2: State Laws and Encryption Exceptions

State	Bill	HIPAA Exception	Blanket Encryption Exception	Effective Date
Arizona	Ariz. Rev. Stat. 44-7501	Yes	All encrypted personal information excluded	12/31/2006
Arkansas	Ark. Code 4-110-101 et seq.	No	All encrypted personal information excluded	3/31/2005
California	Cal. Civ. Code 56.06, 1785.11.2, 1798.29, 1798.82	Yes	All encrypted personal information excluded	7/1/2003
Colorado	Colo. Rev. Stat. 6-1-716	No	All encrypted personal information excluded	9/1/2006
Connecticut	Conn. Gen Stat. 36a-701(b)	No	All encrypted personal information excluded	1/1/2006
Delaware	Del. Code tit. 6, 12B-101 et seq.	No	All encrypted personal information excluded	6/28/2005
District of Columbia	D.C. Code 28- 3851 et seq.	No	Encrypted data not explicitly excluded	1/1/2007
Florida	Fla. Stat. 817.5681	No	All encrypted personal information excluded	7/1/2005
Hawaii	Haw. Rev. Stat. 487N-2	Yes	Encrypted data not explicitly excluded	1/1/2007
Idaho	Idaho Code 28-51-104 to 28-51-107	No	Encrypted data not explicitly excluded	7/1/2006
Illinois	815 ILCS 530/1 et seq.	No	All encrypted personal information excluded	1/1/2006
Indiana	Ind. Code 24-4.9 et seq., 4-1-11 et seq., 2009 H.B. 1121	Yes	Encrypted personal information excluded unless key is compromised	6/3/2006
Iowa	Iowa Code 715C.1 (2008 S.F. 2308)	No	All encrypted personal information excluded	7/1/2008
Kansas	Kan. Stat. 50-7a01, 50-7a02	No	All encrypted personal information excluded	1/1/2007
Louisiana	La. Rev. Stat. 51:3071 et seq.	No	Encrypted data not explicitly excluded	1/1/2006
Maine	Me. Rev. Stat. tit. 10 1347 et seq., 2009 Public Law 161	No	All encrypted personal information excluded	1/31/2006
Maryland	Md. Code, Com. Law 14-3501 et seq.	No	All encrypted personal information excluded	1/1/2008
Massachusetts	Mass. Gen. Laws 93H-1 et seq.	No	Encrypted personal information excluded unless key is compromised	2/3/2008
Michigan	Mich. Comp. Laws 445.72	Yes	Encrypted personal information excluded unless key is compromised	7/2/2007
Minnesota	Minn. Stat. 325E.61, 325E.64	No	All encrypted personal information excluded	1/1/2006
Montana	Mont. Code 30-14-1701 et seq., 2009 H.B. 155, Chapter 163	No	All encrypted personal information excluded	3/1/2006
Nebraska	Neb. Rev. Stat. 87-801, -802, -803, -804, -805, -806, -807	No	All encrypted personal information excluded	7/16/2006
Nevada	Nev. Rev. Stat. 603A.010 et seq.	No	All encrypted personal information excluded	1/1/2006
New Hampshire	N.H. Rev. Stat. 359-C:19, -C:20, -C:21		Encrypted personal information excluded unless key or other password or code is compromised	1/1/2007
New Jersey	N.J. Stat. 56:8-163	No	All encrypted personal information excluded	7/2/2006
New York	N.Y. Gen. Bus. Law 899-aa	No	Encrypted personal information excluded unless key is compromised	12/8/2005
North Carolina	N.C. Gen. Stat 75-65	No	Encrypted personal information excluded unless key is compromised	12/1/2006
North Dakota	N.D. Cent. Code 51-30-01 et seq.	No	All encrypted personal information excluded	6/1/2005
Ohio	Ohio Rev. Code 1347.12, 1349.19, 1349.191, 1349.192	No	Definition of a security breach is broad enough to include encrypted personal information for which the key has been compromised.	2/17/2006
Oregon	2007 S.B. 583, Chapter 759	No	Encrypted personal information excluded unless key is compromised	10/1/2007
Pennsylvania	73 Pa. Stat. 2303	Yes	All encrypted personal information excluded	6/30/2006
Rhode Island	R.I. Gen. Laws 11-49.2-1 et seq.	Yes	All encrypted personal information excluded	3/1/2006
Tennessee	Tenn. Code 47-18-2107	No	All encrypted personal information excluded	7/1/2005
Texas	Tex. Bus. & Com. Code 521.03	No	All encrypted personal information excluded	9/1/2005
Utah	Utah Code 13-44-101, -102, -201, -202, -310	No	All encrypted personal information excluded	1/1/2007
Vermont	Vt. Stat. tit. 9 2430 et seq.	No	All encrypted personal information excluded	1/1/2007
Virginia	Va. Code 18.2-186.6	No	Encrypted personal information excluded unless key is compromised	7/1/2008
Washington	Wash. Rev. Code 19.255.010	No	All encrypted personal information excluded	7/24/2005
West Virginia	W.V. Code 46A-2A-101 et seq.	No	Encrypted personal information excluded unless key is compromised	6/26/2008
Wisconsin	Wis. Stat. 134.98 et seq.	No	All encrypted personal information excluded	3/16/2006
Wyoming	Wyo. Stat. 40-12-501 to -501	No	Encrypted data not explicitly excluded	7/1/2007

Based on the text of laws supplied in Alexander (2009). This was then verified against information provided by the National Conference of State Legislatures.

2.4 Data on Hospitals

One of the advantages of studying publicized hospital-sector data breaches is that there are comprehensive financial and customer data about the number of patients (in terms of admissions and outpatient visits), employee compensation, and spending on capital investments. Table 3 summarizes the variables we use in our specifications. The analysis uses data from the AHA hospital survey from 2005-2007. For the year 2008, the AHA has not yet released new hospital data, so we use data from the previous year.

The annual American Hospital Survey covers more than 6,000 hospitals. We matched these to the HIMSS database using Medicare ID numbers where available. We were able to match all but 193 of our the hospitals in the HIMSS database. The hospitals we could not match from the HIMSS database were largely hospitals that were split into two campuses in the HIMSS database but reported as a single campus in the AHA database. There were, however, over 1,000 hospitals in the AHA database for which there were no data. These unmatched hospitals had 137 beds as compared to 215 beds for the matched hospitals. This implies that our results should be interpreted as a study of publicized data breaches at larger hospitals. In all, after combining the two datasets we were left with 4,325 hospital observations in each year.

3 Empirical Analysis and Results

We start by analyzing the effect of security software on customer data breaches in a simple panel framework. We then move to a more complex framework that jointly models the endogenous adoption of security software alongside data breaches in section 3.1.

The initial specification takes the form of a probit, where the probability of a hospital i suffering from a publicized data breach in year t is captured by a binary variable $DataBreach_{it}$.[8]

[8]We treat $DataBreach_{it}$ as a binary variable because only one hospital had two publicized data breaches

Table 3: Summary Statistics for Full Sample

	Mean	Std. Dev.	Min	Max
Any Data Breach	0.019	0.14	0	1
Data Breach: Lost	0.0066	0.081	0	1
Data Breach: Theft	0.0079	0.089	0	1
Data Breach: Fraud	0.0037	0.061	0	1
Encryption	0.50	0.50	0	1
Physician Documentation	0.24	0.43	0	1
Firewall	0.59	0.49	0	1
Clinical Data Repository	0.66	0.47	0	1
Data Warehouse Financial	0.22	0.42	0	1
Data Warehouse Clinical	0.17	0.38	0	1
EMPI (Enterprise Master Person Index)	0.30	0.46	0	1
State Data Breach Law	0.50	0.50	0	1
Encryption Exception	0.39	0.49	0	1
Payroll Expense per Patient ($000)	7.55	9.03	0.0027	589.1
Capital Expense per Patient ($000)	18.0	21.6	0.0068	1549.7
Admissions (000)	7.68	9.32	0.012	108.6
# Hospitals in System	21.7	40.9	0	170
Average Pay in County ($000)	34.3	10.0	13.5	102.2
Total Outpatient Visits (000)	128.8	187.1	0	3282.5
Full Time Employees (000)	0.95	1.31	0.011	17.8
PPO	0.65	0.48	0	1
HMO	0.55	0.50	0	1

17,300 observations for 4,325 hospitals over 4 years.

$$\text{Prob}(DataBreach_{it} = 1 | Encryption_{it}, X_{it}) = \Phi(Encryption_{it}, X_{it}) \quad (1)$$

X_{it} is a vector of covariates that includes controls for the nature of the hospital and its IT infrastructure as well as both state and year fixed effects.

We control for heterogeneity at the hospital level using a rich set of hospital controls. We have also estimated a linear panel model with hospital-level fixed effects with similar results. However, we caution that hospital-level fixed effects are unlikely to be precisely estimated, given the fact we only observe a handful of repeated observations (Chamberlain, 1985) as the panel spans only 2005-2008. Despite this limitation, as evident in the results reported in Table A-1, the results are reassuringly similar.

We assume a normal distribution, implying a probit specification. We report heteroskedasticity-robust standard errors that are clustered at the state level to address potential correlations in the errors between different hospitals in the same state and serial correlations within states over time (Bertrand et al., 2004).

Table 4 reports results from this initial probit specification. To simplify interpretation the estimates are reported as marginal effects averaged across observations in the sample. Column (1) presents results for a simple panel model. A hospital having adopted encryption software is positively correlated with experiencing a publicized data breach, controlling for state and year. However, this may occur because a hospital is more likely to adopt encryption software because it is larger and consequently has more patient records to both protect and potentially use.[9] Therefore, Column (2) adds in controls for the hospital's characteristics. As expected, the coefficient on $Encryption_{it}$ becomes smaller. However, it still remains positively and significantly correlated with the hospital experiencing a publicized data breach. Many of the coefficients of the controls are statistically insignificant. The significant coef-

in the same year in our data.

[9]We also show that our results when we focus only on large hospitals in Table A-2 in the appendix.

ficients suggest that hospitals that pay their employees more are less likely to experience a publicized security breach. However, hospitals that are located in counties with higher wages for the general population are more likely to experience a publicized security breach. Hospitals with more outpatient visits are also more likely to experience a publicized security breach. Hospitals with PPO contracts are less likely to experience a publicized security breach than hospitals with HMO contracts.

Another possible explanation for the positive coefficient on the adoption of encryption software is that hospitals that adopt encryption software are also the ones which have complex IT systems that need to be protected from intruders, and that the existence of electronic data also makes publicized security breaches more likely. Column (3) adds controls for the technological environment. Three types of software systems are associated with an increase in likelihood of a security breach: a financial data warehouse, EMPI (Enterprise Master Person Index) software that allows hospitals to consolidate fragmented records under a master key, and a clinical data repository. For each of the software systems there is a viable explanation for this positive correlation. A financial data warehouse could facilitate the use of patient data to perpetrate financial fraud. EMPI software makes patient tracking within a hospital easier, but could also make it easier for those who wish to misuse medical data to identify it with a patient and also for this data to be meaningfully related to the customer's data in a newsworthy way if the data are lost. Similarly, a clinical data repository (CDR) is a real-time database that consolidates data from a variety of clinical sources to present a unified view of a single patient. This again may make it easier to consolidate data about a single patient. There are some forms of software systems, like firewall software and clinical data warehousing, that do not appear to be significantly related to experiencing a publicized data breach.

These findings are independently important because, as discussed by Miller and Tucker (2009), many of the fears related to electronic medical records that are couched in terms of

protecting consumer privacy are primarily about customer data security. Our results suggest that adoption of electronic medical records by a hospital is linked with a greater potential for a publicized loss of clinical data, and that this risk is primarily a function of the extent to which the hospital uses electronic systems that make it easy to consolidate data about a single patient. Since this ability to consolidate data is one of the major benefits of electronic medical records systems (Jha et al., 2009), it does suggest that there may be trade-offs with data security from the widespread adoption of electronic medical records.

Column (4) adds controls for whether or not there was a state-level data breach law in place in that year. The coefficient is insignificant and economically small, suggesting that laws like this have not been particularly effective at reducing publicized instances of data loss, and have not led to a large increase in the self-reporting of data-loss. It would be premature, however, to assume that there is no effect of the law on publicized data breaches because of the way the data are collected. It is possible that the presence of a law makes it more likely that an OSF volunteer who scours news feeds will find information about a data breach. If so, then there will be an underlying upwards bias in our estimates of how the law affects data breach, that may mask the potential for the law to have reduced the actual number of reported and unreported data breaches.

In general, the magnitudes of these estimates suggest that there was an increase in the likelihood of a data breach of around 0.4 percentage points if a hospital had installed encryption software (the 95 percent confidence interval is between 0.14 percentage points and 0.65 percentage points). This is quite sizeable relative to a mean of a 1.9 percent chance each year that a hospital would be embroiled in a data breach.

We evaluate different types of data breaches, and how their occurrence is correlated with the adoption of encryption software. We divide the occurrences based on information surrounding their cause of data loss into three types of data breach from Table 1: Data breaches due to loss of equipment, theft of data (either physical or remote), and fraud.

Table 4: Single Equation Specification

	(1) Data Breach	(2) Data Breach	(3) Data Breach	(4) Data Breach	(5) Data Breach: Lost	(6) Data Breach: Theft	(7) Data Breach: Fraud
Encryption	0.012***	0.0062***	0.0040***	0.0040***	0.00084**	0.000076	0.0000017***
	(0.0015)	(0.0012)	(0.0013)	(0.0013)	(0.00032)	(0.00018)	(0.0000043)
Payroll Expense per Patient		-0.00052***	-0.00047***	-0.00046***	-0.000037*	-0.000082*	-0.000000082**
		(0.00016)	(0.00015)	(0.00015)	(0.000027)	(0.000047)	(0.00000023)
Capital Expense per Patient		0.0000065	0.0000021	0.0000026	0.00000083	-0.0000060	0.000000010**
		(0.000041)	(0.000039)	(0.000038)	(0.000013)	(0.0000079)	(0.000000030)
Admissions (000)		0.00017**	0.00014**	0.00014**	0.0000097	0.000015*	0.000000016
		(0.000080)	(0.000075)	(0.000075)	(0.000021)	(0.000018)	(0.000000048)
# Hospitals in System		-0.0000056	-0.0000064	-0.0000063	-0.0000027	-0.0000031***	6.0e-09**
		(0.0000054)	(0.0000054)	(0.0000053)	(0.0000041)	(0.0000028)	(0.000000015)
Average Pay in County		0.00019***	0.00017***	0.00016***	0.000018*	0.000013***	0.000000032***
		(0.000042)	(0.000038)	(0.000039)	(0.000015)	(0.0000096)	(0.000000084)
Total Outpatient Visits (000)		0.0000058***	0.0000050***	0.0000050***	0.00000082*	0.00000049**	-1.1e-09
		(0.0000021)	(0.0000020)	(0.0000020)	(0.00000032)	(0.00000031)	(2.4e-09)
Full Time Employees		-0.00061	-0.00049	-0.00048	0.000055	-0.000023	0.000000027
		(0.00061)	(0.00059)	(0.00058)	(0.00018)	(0.000083)	(0.00000016)
PPO		-0.0064***	-0.0059***	-0.0058***	-0.00070*	-0.00081***	0.00000039
		(0.0016)	(0.0015)	(0.0015)	(0.00036)	(0.00050)	(0.0000011)
HMO		0.0042***	0.0036***	0.0036***	0.00012	0.00064***	-0.00000032
		(0.0015)	(0.0013)	(0.0013)	(0.00030)	(0.00050)	(0.00000100)
Physician Documentation			-0.0011*	-0.0011*	-0.00035	-0.00019**	0.00000014
			(0.00067)	(0.00066)	(0.00027)	(0.00016)	(0.00000043)
Firewall			0.0015	0.0015	0.00032	0.00020	-0.00000076**
			(0.0013)	(0.0013)	(0.00044)	(0.00030)	(0.0000021)
Clinical Data Repository			0.0020**	0.0020**	0.00042	0.00011	0.00000052**
			(0.00094)	(0.00093)	(0.00041)	(0.00013)	(0.0000014)
Data Warehouse Financial			0.0018***	0.0018***	-0.00051	-0.000012	0.0000011**
			(0.00070)	(0.00071)	(0.00040)	(0.00024)	(0.0000027)
Data Warehouse Clinical			-0.00085	-0.00083	0.00024	-0.00034**	0.00000041*
			(0.00077)	(0.00076)	(0.00047)	(0.00023)	(0.0000011)
EMPI			0.0015**	0.0015**	0.00094***	0.00038**	-0.0000015***
			(0.00071)	(0.00070)	(0.00027)	(0.00020)	(0.0000037)
State Data Breach Law				0.00069	0.00057	-0.000081	0.0000041***
				(0.0014)	(0.00092)	(0.00031)	(0.000011)
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	17300	17300	17300	17300	17300	17300	8888
Log-Likelihood	-1344.5	-1254.3	-1241.5	-1241.4	-480.3	-537.3	-201.6

Panel data from 2005-2008 for 4,325 hospitals in the US. Probit specification. Marginal effects reported are averaged across observations in the sample. Robust standard errors clustered at the state level. * $p < 0.10$, ** $p < 0.05$,*** $p < 0.01$ Dependent variable in Columns (1)-(4) is an indicator variable for whether there was any data breach at the hospital. Dependent variable in Columns (5)-(7) are indicator variables for whether there was a data breach due to equipment loss, theft or fraud.

These labels are the labels given the breaches by the OSF volunteers. Lost equipment refers to mislaid or misplaced equipment. Theft of equipment refers to occasions when there is definite information that the equipment was stolen by external parties. Since there were only eight instances of data breaches linked to 'Hackers' in the period we study, and every hospital that experienced hacking had adopted encryption software, there was not sufficient variation to be able to run the probit for this kind of data breach. As summarized in Table 3, in some cases the number of publicized data breaches in each of these areas is small, so our results should be treated with that limitation in mind.

The results for different kinds of data breaches are reported in Columns (5)-(7) in Table 4. There are suggestive differences in the relative magnitudes of the positive correlations between encryption software adoption and the likelihood of that kind of data breach. Our results suggest that adoption of encryption software is more likely to be associated with instances of data loss due to equipment loss and fraud, but not more likely to be associated with an increase in the theft of data. The positive effect of encryption on data breaches due to loss of equipment could suggest that employees become more careless with equipment if they feel that the data on it are secure. This is a speculative, however, and we lack precise data to pinpoint the mechanism.

The lack of a negative effect for encryption software on the instances of theft of data may be because thieves were not aware that data would be encrypted on the laptop they were stealing. The significant and positive correlation between encryption software and fraud is suggestive about the continued threat from a firm's own employees. One potential explanation is that firms that deploy encryption software, feeling that they have secured their data from external sources, may become complacent about data access procedures within the firm. Another explanation is that firms are less concerned that data loss will lead to litigation or harm their reputation among customers if they have taken the precaution of installing encryption software. The data do not support this belief, however, because all

nine of the cases in our data that mention consumer lawsuits happened at hospitals that had adopted encryption software.

There are some other interesting differences with the correlates of the different kinds of data breach. Large hospital systems are more likely to suffer publicized data breaches due to fraud, but less likely to experience direct theft of data. Hospitals with PPO contracts are less likely to experience publicized data breaches due to equipment loss or data theft. Hospitals with HMO contracts are more likely to experience publicized cases of data theft. The establishment of a financial and clinical data warehouse is more likely to be associated with an increase in data fraud than other types of data breaches. Indeed, maintaining a data clinical warehouse reduces the chance of data being stolen, presumably because the data are no longer being stored on local machines. EMPI software allows the easy tracking of patients, and is associated with increases in data breaches due to equipment loss or data theft, but it seems also to be associated with a reduction in internal fraud. Speculatively, this could be because having a master key makes it easier to prevent fraud by monitoring who is accessing data. However, it may also make data breach cases more likely to be newsworthy, as the data can be more readily identified back to an individual patient. It also appears that state breach data reporting laws are correlated with a smaller number of publicized data breaches involving fraud. This may be because firms are more likely to be because are more likely to invest trying to prevent this particularly salient kind of data breach with a law in place.

3.1 Endogenous Technology Adoption

Even with controls for observable heterogeneity for hospitals that have adopted encryption software, there may still be unobservable heterogeneity that can jointly explain the loss of data and the adoption of encryption software. To address this, we move to a model that explicitly treats the binary decision to adopt encryption software as endogenous by separately estimating an equation that captures adoption. That is, in addition to estimating (1), we

also estimate simultaneously an equation (2) for the decision to adopt encryption software, allowing for correlations in the normally distributed error terms.

$$\text{Prob}(Encryption_{it} = 1 | EncryptionException_{it}, Z_{it}) = \Phi(EncryptionException_{it}, Z_{it}) \quad (2)$$

Z_{it} is a vector of covariates that as well as including controls from the AHA data for hospital characteristics also includes both state and year fixed effects. We estimate the model using maximum likelihood. Using this bivariate probit approach allows us to control for endogeneity when both the dependent variable and the endogenous variable are discrete.

Wilde (2000) clarifies that the bivariate probit model is identified so long as each equation includes at least one varying exogenous regressor. Nevertheless, rather than relying on the non-linear functional form as our sole source of identification, we also impose an exclusion restriction on the main equation and implement an instrumental variables approach to estimate the impact of encryption software on data breach. Specifically, we include the $EncryptionException_{it}$ indicator in the adoption model but exclude it from the breach model. This approach resembles in spirit traditional linear instrumental variables approaches for continuous data, in that we assume that the existence of a $EncryptionException_{it}$ provision for encryption software in data breach laws is a plausibly exogenous motivator for the adoption of encryption software. The key difference is that we use a model that reflects the fact that both variables are binary.

Angrist and Pischke (2008, pp. 199-205) uses data from Angrist and Evans (1998) to show a bivariate probit specification and a traditional linear probability with instrumental variables model produce similar results when the means of the dependent variables are not close to 0 or 1. In our setting, the bivariate probit model is attractive because it constrains the dependent variables to be between 0 and 1. Since we have few positive observations for

Table 5: The effect of breach notification encryption exceptions on encryption software adoption

	Encryption adoption before law	Encryption adoption after law	Difference	T-stat	P-value
States with no encryption exception	0.50	0.54	-0.038	-2.79	0.0052
States with encryption exception	0.38	0.52	-0.13	-12.1	3.3e-33

our main dependent variable, a linear probability model may be biased, since it is unlikely to predict within the correct 0 to 1 range (Horrace and Oaxaca, 2006).[10]

As with any instrumental variables specification, it is important that the instrument be correlated with the potentially endogenous variable of encryption software adoption.

Table 5 gives some descriptive statistics that indeed suggest that incorporating an exception for encryption does encourage hospitals to adopt encryption software, relative to hospitals in states that did not have a blanket exception. There is still a small increase in adoption in states that passed laws that did not allow for a blanket exception. This is to be expected, because some state laws offered a limited but not a full safe harbor for organizations that encrypt data (see Table 2). However, for our identification strategy it is only the strength of a blanket exception for encrypted data relative to a limited exception for encrypted data that is important for identification. We also repeated our estimation in a classic linear model that allowed for familiar instrument-strength testing. According to the Anderson-Rubin Wald F-test statistic of 7.66, our instrument is significant at the ($p < .01$) level.

Table 6 reports results for the bivariate probit. We report marginal effects averaged across observations in the sample. Column (1) reports our main results, while Column (2)-(4) report the results for the different types of publicized data loss. The main results in

[10]In earlier versions of this paper we estimated a linear two-stage least squared probability model. While the results are directionally similar, due to this bias the coefficients are implausibly large.

Table 6: Biprobit Specification

	(1) Data Breach	(2) Data Breach: Lost	(3) Data Breach: Theft	(4) Data Breach: Fraud	(5) Data Breach: All, No Public Records
Loss of Data					
Encryption	0.015***	0.0026	0.0061	0.0014***	0.012***
	(0.0024)	(0.0015)	(0.0063)	(0.0022)	(0.0022)
Payroll Expense per Patient	-0.00026*	-0.00017	-0.00062	-0.000072**	-0.00019
	(0.00016)	(0.00043)	(0.00097)	(0.0000080)	(0.00016)
Admissions (000)	-0.0000027	0.000050	0.000067	0.000015	0.000040
	(0.000061)	(0.00012)	(0.00019)	(0.000014)	(0.000043)
Average Pay in County	0.00010***	0.000080	0.000094	0.000028***	0.000077***
	(0.000046)	(0.00021)	(0.00016)	(.)	(0.000028)
Total Outpatient Visits (000)	0.0000026	0.0000037***	0.0000034	-0.00000093	0.0000028**
	(0.0000017)	(0.0000068)	(0.0000067)	(0.0000011)	(0.0000013)
PPO	-0.0043**	-0.0030*	-0.0064**	0.00034	-0.0038**
	(0.0019)	(0.0057)	(0.0090)	(0.00043)	(0.0014)
HMO	0.0017	0.00061	0.0047	-0.00027	0.0020**
	(0.0015)	(0.0025)	(0.0075)	(0.00034)	(0.00097)
Clinical Data Repository	0.0014	0.0018	0.00082	0.00046**	0.00078
	(0.0011)	(0.0040)	(0.0020)	(0.000015)	(0.00100)
Data Warehouse Financial	0.0015	-0.0022	-0.000060	0.00097**	0.0016**
	(0.00096)	(0.0046)	(0.0020)	(0.000035)	(0.00080)
Data Warehouse Clinical	-0.00063	0.0010	-0.0027*	0.00036*	-0.00028
	(0.0011)	(0.0029)	(0.0033)	(0.00017)	(0.00094)
State Data Breach Law	0.00044	0.0025	-0.00070	0.0042***	-0.0012
	(0.0031)	(0.0061)	(0.0022)	(0.00046)	(0.0023)
Encryption Software Adoption					
Encryption Exception	0.11***	0.077***	0.067***	0.082***	0.12***
	(0.016)	(0.025)	(0.014)	(0.0031)	(0.011)
State Fixed Effects	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
Other Hospital Controls	Yes	Yes	Yes	Yes	Yes
Observations	17300	17300	17300	17300	17261
Log-Likelihood	-12020.4	-11300.4	-11356.5	-11051.5	-11849.2

Panel data from 2005-2008 for 4,325 hospitals in the US. Bivariate probit specification estimates and standard errors reported. Marginal effects reported for marginal probability in the first equation averaged across observations in the sample. Dependent variable in the second equation is an indicator variable for whether the hospital has adopted encryption software. Additional control variables for encryption software adoption equation included but not reported for readability. Dependent variable for first equation in Columns (1) is an indicator variable for whether there was any publicized data breach at the hospital. Dependent variable for first equation in Columns (2)-(4) are indicator variables for whether there was a publicized data breach due to equipment loss, theft or fraud. Column (5) excludes reports of data-loss stemming from official sources.
Robust standard errors clustered at the state level. * $p < 0.10$, ** $p < 0.05$,*** $p < 0.01$.
Only variables where the coefficient was significant in one of the regressions are reported.

Column (1) suggest a similar pattern to before. Organizations are more likely to experience a publicized security breach after installing encryption software than before it. The estimates for Encryption suggest a magnitude of around 1.5 percentage points (with a 95 percent confidence interval from 1.0 to 1.9 percentage points). The fact that these estimates for the effect of encryption are larger than those from the single-equation model may be because we are measuring a local average treatment effect, or in other words we are measuring only the effect of adoption that was provoked by the enactment of a security breach law (Imbens and Angrist, 1994). If hospitals adopt encryption software because of legal 'carrots,' it may be implemented in a less comprehensive and rigorous way than adoption of encryption software where the incentives arise organically from a desire to protect customer data within the organization.

Estimates for the other variables in the equation for data loss are similar to Table 4. In the equation where we estimate the determinants of the adoption of encryption software, as expected, the estimate for the excluded safe-harbor variable 'Exception for Encryption' is positive and significant. This reflects the pattern of Table 5, which shows that firms are responding to these data breach laws by installing encryption software, if the law allows an explicit exception for encryption. We also checked whether the insights about different kind of data breaches held from Table 4. Comparing across column (2)-(4) in Table 6 is suggestive about differences in the relative magnitudes of the positive correlations between encryption data adoption. The results again show that adoption of encryption software is more likely to be associated with instances of data loss and data fraud than with data theft or the loss of data remotely. The correlation in the errors between the equation for the adoption of encryption software and instances of data loss was not significant, suggesting that firms with unobserved traits that make them more likely to adopt encryption software do not also have unobserved traits that increase their risk for experiencing data breaches.

Table 7: Lack of pre-trend in health sector data loss incidents for states where laws were passed

	States with no law	States with law before law passed	Difference	T-stat	P-value
Proportion of hospitals affected by data loss annually	0.019	0.016	0.0024	0.18	0.86

3.2 Validity of Instrumental Variables

For instrumental variable approaches to be valid does not require merely a correlation between the excluded variable and the endogenous variable as shown in Table 5. The estimation also has to meet the exclusion restriction.

For the encryption exception indicator to meet the exclusion restriction, the incorporation of such encryption exceptions in state laws needs to be unrelated to instances of medical data loss in the state, except through the mechanism of giving incentives to hospitals to install encryption software. This seems plausible in this case, because the inclusion of an encryption exception in these laws is not motivated by instances of medical data loss at the state level. As described by Miller and Tucker (2009), generally the security of medical data is addressed in separate sections of the state's regulations. Data breach notification laws generally are motivated by concerns about data security in the banking and retail sector. The state fixed effects should capture time-invariant unobserved heterogeneity that relates differences in data loss rates to the enactment of legal regulation. To alleviate concerns about time-variant heterogeneity, we verified that there was no systematic difference in states' levels of publicized data loss before they enacted these laws compared to states that did not enact laws. As shown in Table 7, there was no significant difference in the annual per-hospital incidences of publicized data loss in states that passed legislation versus states that did not pass legislation.

Another concern is that the way that our data are collected may affect the validity of the instrument. It is possible that after the states passed data breach notification laws (even

those with the encryption exception), the OSF volunteers were more likely to observe a data breach, simply because the data breach notification law meant that the hospital was required to publicize it. Empirically, the results of Column (4), Table 4 suggest that this is not the case, since the coefficient on the enactment of a general breach notification law is economically and statistically insignificant. However, we still recognize that this is a concern. Similarly, it is possible that in a state with encryption exceptions, fewer hospitals disclosed that a data breach had occurred and consequently the volunteers who were collecting the data were less likely to find a news story about it. However, this does not appear to be true in the data. Instead, there is actually a positive correlation between an encryption exception and the publicization of a data loss, rather than the negative correlation one would expect if this story were true. However, there still may be a concern that because data breach laws often require reporting of the loss to the state authorities, this may have actually facilitated the process of volunteers finding out about the data loss. To check the robustness of our research to this concern, we repeated the exercise excluding 39 observations where losses were publicized because they appeared in an official state database that was published online. As reported in column (5) of Table 6, the main results are unchanged.

3.2.1 Falsification Checks

Since we cannot directly test the exclusion restriction, we instead present the results of a falsification check and other empirical evidence that suggests that the exclusion restriction is valid in our setting. Specifically, we examined whether there was still a bump in adoption of encryption software in states whose breach disclosure rules exempted HIPAA organizations including hospitals.

Table 8 shows that in states where hospitals were excluded from data breach reporting requirements, there was a similar increase in the adoption of encryption software to states without these exemptions.

Table 8: The effect of breach notification encryption exceptions on encryption software adoption in states where there were HIPAA exemptions

	Encryption adoption before law	Encryption Adoption after law	Difference	T-stat	P-value
States with no encryption exception	0.50	0.58	-0.074	-3.87	0.00011
States with encryption exception	0.47	0.54	-0.079	-2.68	0.0076

This similar increase is not statistically different from the increase observed for states that did not have HIPAA exemptions and passed laws with no encryption exceptions in Table 5. It is, however, statistically smaller than the increase in encryption software adoption observed in states with encryption exceptions and no exclusion for entities covered under HIPAA.

This suggests that the relative increase in adoption of encryption software that we observe in states that gave encrypted data a safe harbor in Table 5 is linked to the presence of an encryption exception stipulation in the law, rather than to unobserved differences across hospitals in states that enacted the kind of laws that gave safe harbor to encrypted data.

Another concern is that rather than the exception for encryption *per se*, our results are picking up the fact that laws with blanket exceptions for encryptions are less tough in other dimensions than laws that have limited exemptions. If an encryption exemption simply implies that the law is weaker, the positive relationship between breaches and encryption may simply reflect a higher incidence of publicized data breaches in states with weaker security laws. To investigate this, we looked at other dimensions in which the laws differed. We found the states that had laws with blanket exemptions capped firm expenditures at a mean of $232,000 for breach notifications, while states that did not had caps on firm expenditures at a mean of $211,000. In other words, the states that had regulations that were less tough in that they allowed blanket exceptions, were if anything slightly tougher in other dimensions in terms of the expected financial liability of a firm.

Our estimation of the effect of encryption exceptions on the adoption of encryption software and publicized data breaches is also important because of controversy over the optimal policy approach. Many state laws include a blanket exception for encryption of data regardless of the security of the encryption key. However, it is quite common for encryption keys to be compromised. Getgen (2009) shows that 8% of organizations (both those that have experienced and those that have not experienced a security breach) have experienced problems with a lost encryption key. Generally, the breadth of the safe harbor given to encryption in such state laws has been a source of controversy. The security software industry has advocated aggressively for states to include broad safe harbor provisions in order to provide incentives for firms to adopt encryption software (Warmenhoven, 2006). Further, the potential for encryption software to avoid the costs of data breach notification is often touted in firm marketing materials. However, the inclusion of the general language that governs most safe harbors has been criticized by security experts as being possible to satisfy by even the most 'trivial' and insecure of algorithms (Carlson, 2005). The results in this paper suggest that while such blanket safe harbors do encourage the adoption of encryption software, safe harbors alone may not be sufficient to provide adequate data protection.

4 Implications

Collection and analysis of customer data is at the heart of many firms' IT systems. However, the loss of customer data can have substantial negative consequences for firms. The costs can stem from litigation or fines, or from negative publicity that harms the firm's reputation and erodes customer loyalty. This paper is the first quantitative study of the effect of the commonly-advocated data security policy of encryption on publicized incidents of data loss. Unexpectedly, we find that the adoption of encryption software increases the likelihood of experiencing a publicized case of data loss. This is driven by an increase in publicized cases of data loss associated with employee dishonesty (fraud) and employee carelessness (equipment

loss) after the adoption of encryption software.

The findings of this paper have public policy implications for the regulation of data security. A major emphasis of recent regulation has been to encourage encryption. However, encryption requires careful encryption key management, and the underlying algorithm itself must be strong to protect data, so blanket provisions exempting encrypted data are inadequate. Further, many instances of electronic data loss are due to the insider threat rather than direct instances of hacking or theft. Encryption does not protect organizations against this insider threat. Our research suggests that policy makers should expand the breadth of security measures to encompass other technologies such as user-access controls that are better able to address the insider threat.

The findings of this paper also suggest that digitization of patient records may increase the likelihood of data breaches. This supports the emphasis in recent policies designed to encourage the digitization of patient data such as the 2009 HITECH act, on also addressing issues of data breaches and patient protection. However, our results also indicate that data breaches appear to be facilitated by a hospital using electronic systems that make it easy to consolidate data about a single patient. This suggests that future clarifications and improvements to the HIPAA data security rules should include particularly strong safeguards for the kind of systems that facilitate a 'master key' approach to patient data. We also find that large hospitals are more likely to lose data. This is understandable since they theoretically have more data to lose but does suggest that organizational or financial capacity is not sufficient to counter the underlying risk of data loss. In particular, our findings suggest that hospitals that are contemplating adopting EMR systems need to make sure that encryption is comprehensively applied and that employees comply with this policy. They also need to ensure that they have additional systems in place to address the potential threat to the security of data due to internal fraud.

Though our focus on data loss has been from a firm perspective, there are also implications

for our findings for customers. This is particularly important in the health sector setting that we study, because medical identity theft has grown faster than other types of identity theft in recent years (Mincer, 2009).[11] Our research suggests that, given the threat from employee negligence or fraud, consumers should not rely on firm statements about the encryption of data to protect their identities, but instead should themselves monitor their records for any unusual activity.

There are of course limitations to our findings. First, we only study the likelihood of publicized data loss rather than the harm that results from data loss. It is very likely that encryption software is useful at limiting harm when data is stolen. Firms are concerned with the negative publicity relating to any loss of data, so often managers' primary concern is to avoid any instance of data loss. It could be that the potential for expensive legal action as a result of identity theft would be reduced if encryption software were used. Analysis of the news stories gives anecdotal evidence, however, that this is not the case in our setting. Of the nine publicized cases of data breach in our dataset where the story relating to the data breach mentioned a consumer lawsuit, all nine of the hospitals had already adopted encryption software.

Second, our empirical analysis focuses on the health sector, a sector of the economy where data losses are likely to include sensitive personal data and also which has been criticized for its low level of penetration of technology.

Third, the kind of encryption software that we study and situations where it is employed is typically used for data stored on disks. We do not study the effect of encryption for remote communications, such as is often used on websites.

Fourth, we do not have data on other commonly advocated security policies such as training and awareness programs; manual procedures and controls; and identity and access-

[11] Most anecdotes describe medical identity theft perpetuated by firm employees. Mincer (2009) describes a front-desk clerk at a medical clinic in Weston, Fla. who downloaded the personal information of more than 1,100 Medicare patients and gave it to a cousin, who then made $2.8 million in false Medicare claims.

management deployments. Therefore, while our results suggest that encryption by itself is not enough to lower risks of security breaches, we cannot evaluate whether these other policies used in conjunction with encryption will be effective in lowering the risk.

Fifth, we speculate that our result that the adoption of encryption software is positively associated with more instances of publicized data losses, because it encourages people to be careless, or makes internal data breaches in the form of fraud easier to conduct because of the false sense of security given by the encryption software. This is in line with behavioral theories of a 'risk thermostat' proposed by Adams (1999) who suggests that most people and organizations are governed by a finely balanced risk thermostat. Containing and minimizing one dimension of risk can lead individuals and organizations to behave in a more risky way in other dimensions. The most cited example of this is that drivers who wear seat-belts tend to take more risk when driving, but there are obvious parallels that encryption may lull organizations and employees into a false sense of security which means they fail to take appropriate precautions along other dimensions. More research is needed to evaluate the potential for such behavioral mechanisms that may undermine security practices in organizations.

References

Acquisti, A., A. Friedmann, and R. Telang (2006). Is there a cost to privacy breaches? An Event Study. In *Proceedings of the Twenty Seventh International Conference on Information Systems and Workshop on the Economics of Information Security,*, pp. 1–23.

Adams, J. (1999). The management of risk and uncertainty. *Policy analysis 335*, 50.

Alexander, P. (2009). *Data Breach Disclosure Laws: A State-by-State Perspective* (2 ed.). Thomson West.

Anderson, R. and T. Moore (2006). The economics of information security. *Science 314*(5799), 610–613.

Andrews, J. (2010, February 03). Balancing hospital security tricky. Technical report. Healthcare IT News.

Angrist, J. D. and W. N. Evans (1998, June). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review 88*(3), 450–77.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Press.

Beaver, K. and R. Herold (2004). *The practical guide to HIPAA privacy and security compliance*. CRC Press.

Bertrand, M., E. Duflo, and S. Mullainathan (2004, February). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Bulgurcu, B., H. Cavusoglu, and I. Benbasat. (2010). Information security compliance: an empirical study of rationality-based beliefs and information security awareness. *forthcoming, MIS Quarterly*.

Carlson, C. (2005, May 31). Storm brews over encryption safe harbor in data breach bills. *Eweek.com*.

Cavusoglu, H., B. Mishra, and S. Raghunathan (2004). The effect of internet security breach announcements on market value: Capital market reactions for breached firms and internet security developers. *International Journal of Electronic Commerce 9*(1), 70–104.

Chamberlain, G. (1985). *Longitudinal analysis of labor market data*, Chapter Heterogeneity, omitted variable bias, and duration dependency, pp. 3–38. Cambridge University Press, Cambridge.

Dolan, P. (2010, Feb 22). Data security breaches often triggered by carelessness. Technical report, American Medical News.

Gal-Or, E. and A. Ghose (2005). The economic incentives for sharing security information. *Informations Systems Research 16*(2), 186–208.

Gaudin, S. (2007, May). Estimates Put T.J. Maxx Security Fiasco At $4.5 Billion. *InformationWeek*.

Getgen, K. (2009). 2009 encryption and key management benchmark survey. Technical report. Thales Group.

Goel, S. and H. A. Shawky (2009). Estimating the market impact of security breach announcements on firm values. *Information Management 46*(7), 404–410.

Horrace, W. C. and R. L. Oaxaca (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economic Letters 90*, 321–327.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Jha, A. K., C. M. DesRoches, E. G. Campbell, K. Donelan, S. R. Rao, T. G. Ferris, A. Shields, S. Rosenbaum, and D. Blumenthal (2009). Use of Electronic Health Records in U.S. Hospitals. *N Engl J Med 360*(16), 1628–1638.

MacKie-Mason, J. K. and H. R. Varian (1996). *Networks, Infrastructure and the New Task for Regulation*, Chapter Some economics of the Internet. University of Michigan.

Miller, A. R. and C. Tucker (2009, July). Privacy protection and technology adoption: The case of electronic medical records. *Management Science 55*(7), 1077–1093.

Mincer, J. (2009, November 29). Patient ID Theft Rises. Technical report, Wall Street Journal.

Nicastro, D. (2010, August 12). HITRUST: HIPAA Breaches Near $1 Billion. *HealthLeaders Media*.

Ponemon, L. (2008). Consumers report card on data breach notification. Technical report, Ponemon Institute.

Ponemon, L. (2009). Fourth Annual US Cost of Data Breach Study. Technical report, Ponemon Institute.

Roberds, W. and S. L. Schreft (2009). Data breaches and identity theft. *Journal of Monetary Economics 56*(7), 918 – 929.

Romanosky, S., R. Telang, and A. Acquisti (2008). Do Data Breach Disclosure Laws Reduce Identity Theft? *Mimeo, Carnegie Mellon*.

Schuman, E. (2009, July 12). Data breach bill's flawed assumptions. Technical report, CBS News.

Telang, R. and S. Wattel (2007). An empirical analysis of the impact of software vulnerability announcements on firm stock price. In *IEEE Transactions on Software Engineering.*, Volume 33 of *i8*, pp. 544–557.

Warmenhoven, D. (2006, June 8). Protect me, protect my data. *BusinessWeek*.

Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economic Letters 69*, 309–312.

Appendix

Table A-1: Linear Probability Model, Hospital-Level Fixed Effects. Alternative Results for Table 6

	(1) Data Breach	(2) Data Breach: Lost	(3) Data Breach: Theft	(4) Data Breach: Fraud
Encryption	0.0167***	0.00668***	0.00251	0.00743***
	(0.00302)	(0.00151)	(0.00221)	(0.00102)
Payroll Expense per Patient	-0.0000793	0.000115	0.000304**	-0.000404
	(0.000349)	(0.000116)	(0.000134)	(0.000272)
Admissions (000)	0.00132***	0.000182	0.000892***	0.000180
	(0.000453)	(0.000274)	(0.000264)	(0.000122)
Average Pay in County	0.000794***	0.000164*	0.000266**	0.000273***
	(0.000194)	(0.0000922)	(0.000125)	(0.0000820)
Total Outpatient Visits (000)	0.0000459**	0.0000198*	0.0000321*	-0.00000867***
	(0.0000231)	(0.0000112)	(0.0000165)	(0.00000316)
PPO	-0.0188***	-0.00566***	-0.0140***	0.00106
	(0.00453)	(0.00193)	(0.00318)	(0.000927)
HMO	0.00576	-0.000834	0.00771***	-0.00122
	(0.00407)	(0.00185)	(0.00293)	(0.00105)
Clinical Data Repository	0.00248	0.00214	0.00197	-0.00103
	(0.00250)	(0.00133)	(0.00163)	(0.000713)
Data Warehouse Financial	0.000937	-0.00515***	-0.00214	0.0107***
	(0.00385)	(0.00168)	(0.00263)	(0.00193)
Data Warehouse Clinical	-0.00293	0.00299	-0.00966***	0.00375*
	(0.00463)	(0.00209)	(0.00313)	(0.00217)
State Data Breach Law	-0.00773**	0.0000467	-0.00402***	-0.00311***
	(0.00319)	(0.00136)	(0.00124)	(0.000754)
State Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Other Hospital Controls	Yes	Yes	Yes	Yes
Observations	17300	17300	17300	17300
Log-Likelihood				

Panel data from 2005-2008 for 4,325 hospitals in the US. Dependent variable in Columns (1) is an indicator variable for whether there was any data breach at the hospital. Dependent variable for first equation in Columns (2)-(4) are indicator variables for whether there was a data breach due to equipment loss, theft or fraud.
Fixed effects at hospital level
Robust standard errors clustered at the state level. * $p < 0.10$, ** $p < 0.05$,*** $p < 0.01$.
Only variables where the coefficient was significant in one of the regressions are reported.

Table A-2: Single Equation Specification (Large Hospitals)

	(1) Data Breach	(2) Data Breach	(3) Data Breach	(4) Data Breach	(5) Data Breach: Lost	(6) Data Breach: Theft	(7) Data Breach: Fraud
Encryption	0.022***	0.015***	0.0088**	0.0087**	0.0016*	-0.00024	0.000014***
	(0.0034)	(0.0027)	(0.0036)	(0.0036)	(0.00074)	(0.00064)	(0.000028)
Payroll Expense per Patient		-0.0020***	-0.0018***	-0.0018***	-0.00018***	-0.00036**	-0.00000024
		(0.00066)	(0.00065)	(0.00064)	(0.00011)	(0.00016)	(0.00000050)
Capital Expense per Patient		0.000056	0.000071	0.000072	0.0000013	-0.0000071	0.000000077**
		(0.00013)	(0.00013)	(0.00013)	(0.000026)	(0.000026)	(0.00000019)
Admissions (000)		0.00027	0.00026	0.00026	-0.0000096	0.000032	0.00000017
		(0.00021)	(0.00020)	(0.00020)	(0.000044)	(0.000047)	(0.00000042)
# Hospitals in System		-0.0000025	-0.0000066	-0.0000066	-0.0000039	-0.000010***	0.000000054**
		(0.000016)	(0.000017)	(0.000016)	(0.0000056)	(0.0000075)	(0.00000011)
Average Pay in County		0.00041***	0.00037***	0.00037***	0.000024*	0.000036**	0.00000026***
		(0.000093)	(0.000088)	(0.000089)	(0.000018)	(0.000021)	(0.00000055)
Total Outpatient Visits (000)		0.000012**	0.000010**	0.000010**	0.00000089*	0.0000013*	-0.000000012
		(0.0000052)	(0.0000051)	(0.0000051)	(0.00000039)	(0.00000076)	(0.000000021)
Full Time Employees		-0.000049	-0.00014	-0.00014	0.00035	0.000055	-0.000000066
		(0.0016)	(0.0016)	(0.0016)	(0.00040)	(0.00027)	(0.0000015)
PPO		-0.020***	-0.019***	-0.019***	-0.0013***	-0.0024***	0.0000043
		(0.0046)	(0.0043)	(0.0043)	(0.00052)	(0.0013)	(0.000011)
HMO		0.012***	0.011***	0.011***	0.00034	0.0020***	-0.0000052
		(0.0042)	(0.0039)	(0.0039)	(0.00047)	(0.0014)	(0.000013)
Physician Documentation			-0.0023	-0.0023	-0.00070	-0.00057**	0.0000026
			(0.0019)	(0.0019)	(0.00040)	(0.00044)	(0.0000057)
Firewall			0.0066	0.0066	-0.000078	0.0012*	-0.0000086***
			(0.0044)	(0.0044)	(0.00092)	(0.0012)	(0.000019)
Clinical Data Repository			0.0036	0.0036	0.000093	0.00010	0.0000056
			(0.0030)	(0.0030)	(0.00060)	(0.00045)	(0.000011)
Data Warehouse Financial			0.0046**	0.0046**	-0.00051	0.000048	0.0000085**
			(0.0020)	(0.0020)	(0.00051)	(0.00079)	(0.000016)
Data Warehouse Clinical			-0.0043*	-0.0042*	-0.00022	-0.0013**	0.0000032*
			(0.0023)	(0.0023)	(0.00042)	(0.00076)	(0.0000076)
EMPI			0.0068***	0.0067***	0.0020***	0.0014**	-0.000013***
			(0.0022)	(0.0022)	(0.00062)	(0.00062)	(0.000026)
State Data Breach Law				0.0016	0.0010	-0.000060	0.000035***
				(0.0046)	(0.0012)	(0.0011)	(0.000080)
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8304	8304	8304	8304	8647	8647	4876
Log-Likelihood	-980.9	-920.7	-906.8	-906.7	-329.6	-391.5	-162.5

Panel data from 2005-2008 for 4,325 hospitals in the US. Probit specification. Marginal effects reported are averaged across observations in the sample.

Robust standard errors clustered at the state level. * $p < 0.10$, ** $p < 0.05$,*** $p < 0.01$ Dependent variable in Columns (1)-(4) is an indicator variable for whether there was any data breach at the hospital. Dependent variable in Columns (5)-(7) are indicator variables for whether there was a data breach due to equipment loss, theft or fraud.

Table A-3: Single Equation Specification (Omitting adoption in same year as data-loss)

	(1) Data Breach	(2) Data Breach	(3) Data Breach	(4) Data Breach	(5) Data Breach: Lost	(6) Data Breach: Theft	(7) Data Breach: Fraud
Encryption	0.011***	0.011***	0.0099***	0.0098***	0.0020	0.0016	0.0096*
	(0.0019)	(0.00074)	(0.0016)	(0.0017)	(0.00086)	(0.0030)	(0.0052)
Payroll Expense per Patient		-0.0011***	-0.0013***	-0.0013***	-0.00013**	-0.0014*	-0.00038*
		(0.00038)	(0.00047)	(0.00047)	(0.000082)	(0.00077)	(0.00020)
Capital Expense per Patient		0.000054	0.000037	0.000036	0.000041	-0.000093	0.00014*
		(0.000077)	(0.000094)	(0.000094)	(0.000033)	(0.00014)	(0.000075)
Admissions (000)		0.00034**	0.00033*	0.00033*	-0.000035	0.00027**	0.00023
		(0.00016)	(0.00019)	(0.00020)	(0.000076)	(0.00016)	(0.00021)
# Hospitals in System		-0.0000040	-0.0000025	-0.0000026	-0.0000036	-0.000048***	0.000049
		(0.000011)	(0.000014)	(0.000014)	(0.000016)	(0.000023)	(0.000031)
Average Pay in County		0.00032***	0.00036***	0.00036***	0.000057	0.00019**	0.00029*
		(0.000074)	(0.000093)	(0.000093)	(0.000070)	(0.000081)	(0.00014)
Total Outpatient Visits (000)		0.000010**	0.000011**	0.000011**	0.0000018	0.0000089*	-0.000010
		(0.0000047)	(0.0000056)	(0.0000057)	(0.0000022)	(0.0000047)	(0.0000065)
Full Time Employees		-0.0013	-0.0012	-0.0012	0.00055	-0.00052	-0.00076
		(0.0012)	(0.0015)	(0.0015)	(0.00067)	(0.0011)	(0.0017)
PPO		-0.0099***	-0.012***	-0.012***	0.0013	-0.015***	0.0012
		(0.0030)	(0.0039)	(0.0039)	(0.0012)	(0.0040)	(0.0011)
HMO		0.0083***	0.0093***	0.0093***	-0.00012	0.011***	-0.0016
		(0.0030)	(0.0037)	(0.0037)	(0.0012)	(0.0029)	(0.0014)
Physician Documentation			-0.0022	-0.0022	-0.00099	-0.0032*	0.0038*
			(0.0018)	(0.0018)	(0.0011)	(0.0017)	(0.0022)
Firewall			0.0028	0.0029	0.00040	0.0035	-0.0034*
			(0.0037)	(0.0038)	(0.0016)	(0.0048)	(0.0019)
Clinical Data Repository			0.0050**	0.0050**	0.00030	0.0031	-0.0012*
			(0.0026)	(0.0027)	(0.0014)	(0.0030)	(0.00066)
Data Warehouse Financial			0.0067***	0.0068***	-0.00022	-0.00065	0.013
			(0.0021)	(0.0021)	(0.0013)	(0.0045)	(0.0084)
Data Warehouse Clinical			-0.0018	-0.0019	0.0016	-0.0070**	0.0040
			(0.0020)	(0.0020)	(0.0016)	(0.0026)	(0.0035)
EMPI			0.0013	0.0014	0.0011	0.0077**	-0.013*
			(0.0017)	(0.0017)	(0.0013)	(0.0032)	(0.0072)
State Data Breach Law				-0.0043	-0.0013	0.0011	-0.0029*
				(0.0045)	(0.0035)	(0.0053)	(0.0014)
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13103	13103	13103	13103	14798	14798	14798
Log-Likelihood	-1092.3	-1034.4	-1023.8	-1023.2	-287.9	-479.7	19685.1

Panel data from 2005-2008 for 4,325 hospitals in the US. Probit specification. Marginal effects reported are averaged across observations in the sample. Robust standard errors clustered at the state level. * $p < 0.10$, ** $p < 0.05$,*** $p < 0.01$ Dependent variable in Columns (1)-(4) is an indicator variable for whether there was any data breach at the hospital. Dependent variable in Columns (5)-(7) are indicator variables for whether there was a data breach due to equipment loss, theft or fraud.