Dissertation Proposal

A Comprehensive and Comparative Examination of Healthcare Data Breaches: Assessing Security, Privacy, and Performance

Mohammed Alkinoon

Date: July 18, 2023

Department of Computer Science
The University of Central Florida
Orlando, FL 32816

Doctoral Committee:
Dr. David Mohaisen (Chair)
Dr. Cliff Zou
Dr. Sung Choi
Dr. Xueqiang Wang

Mohammed Alkinoon

Department of Computer Science, University of Central Florida (UCF)
4000 Central Florida Blvd., R1-368, Orlando, FL 32816-2362 USA

Education

Ph.D. in Computer Science (2020 – Current)
University of Central Florida

M.Sc. in Computer Science (2020 – 2022)
University of Central Florida
CGPA: 3.61

B.Sc. in Computer Science (2017 – 2019)
Eastern Florida State College
CGPA: 3.75

Peer-Reviewed Publications

1. **Mohammed Alkinoon**, Sung J Choi, and David Mohaisen. *Measuring healthcare data breaches*. In Information Security Applications: 22nd International Conference, WISA 2021. (Acceptance Rate: 38%)

2. **Mohammed Alkinoon**, Marwan Omar, Manar Mohaisen, and David Mohaisen. *Security Breaches in the Healthcare Domain: A Spatiotemporal Analysis*. In Computational Data and Social Networks: 10th International Conference, CSoNet 2021.

3. **Mohammed Alkinoon**, Abdulrahman Alabduljabbar, Hattan Althebeiti, Rhongho Jang, Dae-Hun Nyang, and David Mohaisen. *Understanding the Security and Performance of the Web Presence of Hospitals: A Measurement Study*. At the 32nd International Conference on Computer Communications and Networks (ICCCN 2023). (Acceptance Rate: 30%)

4. Mohammed Alqadhi, **Mohammed Alkinoon**, Jie Lin, Ahmed Abdalaal, and David Mohaisen, *Entangled Clouds: Measuring the Hosting Infrastructure of the Free Contents Web*. (To be submitted)

5. Mohammed Alqadhi, **Mohammed Alkinoon**, Jie Lin, and David Mohaisen. *The Infrastructure Utilization of Free Contents Websites Reveal their Security Characteristics: A Correlation Analysis*. (To be submitted)

6. Mohammed Alqadhi, **Mohammed Alkinoon**, Jie Lin, and David Mohaisen. *Demystifying the Network Characteristics of Free Contents Hosting Infrastructure*. (To be submitted)

7. Hattan Althebeiti, **Mohammed Alkinoon**, Manar Mohaisen, Saeed Salem, DaeHun Nyang, and David Mohaisen. *Enhancing Vulnerability Reports with Automated and Aug-mented Description Summarization*. IEEE Transactions on Big Data.

Contents

Abstract

The healthcare industry is among the most crucial sectors in our lives, considering the importance of its life-saving services. Moreover, healthcare is vital for our well-being and for improving the quality of our lives and communities. Additionally, healthcare is considered the fastest-growing industry in the united states and worldwide. According to a recent study by Grand View Research, the healthcare industry is estimated to reach $81 billion compound annual growth rate (CAGR) by 2026 [57]. The recent transformation of healthcare medical records from traditional paper-based to a digitized era using electronic medical records (EHR) provided many benefits by improving efficiency, patient safety, accessibility and availability, and cost savings. Despite these benefits, the implementation of EHR transformed healthcare into a digital infrastructure by introducing new technological systems and processes which can potentially increase risk exposure, such as data breaches. Several factors can increase the risk of exposure in healthcare, such as increasing volume and accessibility, interconnected systems and networks, insider threats, and cybersecurity threats. Data breaches are an eternal threat to the healthcare industry and can occur internally and externally.

First, to analyze and understand data breaches in the healthcare industry, we start by conducting a detailed measurement-based study of the VERIS (Vocabulary for Event Recording and Incident Sharing) dataset. Among other characteristics, we temporally analyze data breaches and their growth over time. To understand the attacks' intent, we analyze the type of breaches over various security attributes and characterize the threat actions, highlighting the attack vector employed for the breach. We hope those characterizations shed light on the trend and the attack vectors, thus providing directions for mitigating those breaches.

Second, we conduct a spatiotemporal analysis of healthcare data breaches by examining the geographical and temporal distribution of several attributing characteristics. We provide a detailed measurement-based study of the VERIS (Vocabulary for Event Recording and Incident Sharing) and the Office of Civil Rights (OCR) datasets. To understand attackers' intents and motives, we analyze the type of assets targeted during breaches over various characteristics to investigate their effect. We further analyze data breaches considering multiple views looking at their distribution, affected entities, breached information, location of the breach, etc.

Finally, given the lack of systematic work on understanding the characteristics of hospitals' presence on the web and their associated security and performance attributes, we explore hospitals' web presence across a range of attributes. Moreover, through a comparative analysis, we uncover the differences and similarities between the Government, Non-profit, and Proprietary hospitals in the United States. Our analysis is conducted across three dimensions: security, contents, and domains. By exploring these dimensions, we aim to uncover the similarities and differences among these types of hospitals in terms of their online presence.

1 Introduction

A data breach can be defined as ”any security incident in which unauthorized parties gain access to sensitive data or confidential information, including personal data (Social Security numbers, bank account numbers, healthcare data) or corporate data (customer data records, intellectual property, financial information)” [33]. In healthcare, data breaches cause devastating damage to both healthcare organizations and patients due to the sensitive and personal information involved in this sector. Data breaches expose highly personal and confidential information such as medical records, diagnoses, treatment plans, insurance, and financial information. Cybercriminals can exploit the breached information maliciously, such as through identity theft and financial fraud, by using personal information such as Social Security Numbers (SSNs) and insurance details to perform fraudulent activity.

Healthcare data breaches cause severe financial consequences to health organizations and patients. Data breach victims of theft or fraudulent activities may face financial losses and difficulties resolving after the breach. Additionally, healthcare organizations may face legal consequences, financial penalties, loss of reputation, and potential litigation resulting from breaches. The healthcare industry continues to be a primary target for cybercriminals, as evidenced by recent trends and statistics. In 2023, the healthcare sector remains highly vulnerable to cyber threats. According to the Office for Civil Rights (OCR), the first three months of 2023 alone witnessed 145 data breaches [25]. To put this into perspective, in 2022, there were a staggering 707 reported incidents, resulting in the compromise of approximately 51.9 million records [48].

The Health Insurance Portability and Accountability Act (HIPAA) [50] establishes privacy and security standards to protect patient’s health information, securing its handling and protection by healthcare providers, health plans, and healthcare clearinghouses. HIPAA categorizes data breaches and other compliance failures into four tiers, which determine the severity of the incident and the corresponding penalties. Unlike some other data protection laws, HIPAA fines are assessed per violation, which can pertain to specific areas of non-compliance or individual compromised records. HIPAA classifies data breaches and compliance failures into four tiers, each associated with its own range of penalties. The penalties for Tier 1 breaches typically range from $100 to $50,000 per violation. Tier 2 penalties can be anywhere from $1,000 to $50,000 per violation. Tier 3 breaches, resulting from intentional disregard but corrected within a specified timeframe, may incur penalties between $10,000 and $50,000 per violation. Finally, the most severe Tier 4 breaches, characterized by sustained noncompliance, can lead to penalties starting at $50,000 or more per violation [29]. The specific penalty amount for HIPPA breaches varies based on the circumstances and the organization’s response.

Given the continuous and rapid increase in the number of data breaches in the healthcare sector, we find it crucial to emphasize the importance of conducting thorough investigations and analyses of these breaches.

1.1 Statement of Research

In this dissertation, we propose three comprehensive studies that delve into the analysis of data breaches in the healthcare industry. Each study focuses on specific aspects providing a deep understanding of the critical risk of data breaches. We further elaborate on each study in the following.

***Measuring Healthcare Data Breaches* (§ 3).**
Recently, healthcare data breaches have grown rapidly. Moreover, throughout the COVID-19 pandemic, the level of exposure to security threats increased as the frequency of patient visits to hospitals has also increased. During the COVID-19 crisis, circumstances and constraints such as the curfew imposed on the public have resulted in a noticeable increase in Internet usage for healthcare services, employing intelligent devices such as smartphones. The Healthcare sector is being targeted by criminals internally and externally; healthcare data breaches impact hospitals and patients alike. To examine issues and discover insights, a comprehensive study of health data breaches is necessary. To this end, this study investigates healthcare data breach incidents by conducting measurements and analysis recognizing different viewpoints, including temporal analysis, attack discovery, security attributes of the breached data, attack actors, and threat actions. Based on the analysis, we found the number of attacks is decreasing, although not precluding an increasing severity, the time of attack discovery is long across all targets, breached data does not employ basic security functions, threat actions are attributed to various vectors, e.g., malware, hacking, and misuse, and could be caused by internal actors. Our study provides a cautionary tale of medical security in light of confirmed incidents through measurements.

***Security Breaches in the Healthcare Domain*: A Spatiotemporal Analysis (§ 4).**
Over the past several years, data breaches have grown and become more expensive in the healthcare sector. Healthcare organizations are the main target of cybercriminals due to sensitive and valuable data, such as patient demographics, SSNs, and personal treatment records. Data breaches are costly to breached organizations and affected individuals; hospitals can suffer substantial damage after the breach while losing customer trust. Attackers often use breached data maliciously, e.g., demanding ransom or selling patient information on the dark web. To this end, this study investigates data breaches incidents in the healthcare sector, including community, federal, and non-federal hospitals. Our analysis focuses on the reasoning and vulnerabilities that lead to data breaches, including the compromised information assets, geographical distribution of incidents, size of healthcare providers, the timeline discovery of incidents, and the discovery tools for external and internal incidents. We use correlation to examine the impact of several dimensions on data breaches. Among other interesting findings, our in-depth analysis and measurements revealed that the average number of data breaches in the United States is significantly higher than in the rest of

the world, and the size of the health provider, accounting for factors such as the population and number of adults in a region, highly influences the level of exposure to data breaches in each state.

***Understanding the Security and Performance of the Web Presence of Hospitals*: A Measurement Study (§ 5).**

The recent transformation of healthcare medical records from paper-based to digital and connected systems raises concerns regarding patients' security and online privacy. For instance, sensitive personal information, such as patients' names, addresses, and social security numbers, may be targeted due to the lack of proper security and privacy mechanisms. Using a total of 4,774 hospitals categorized as government, non-profit, and proprietary hospitals, this study provides the first measurement-based analysis of hospitals' websites and connects the findings with data breaches through a correlation analysis. We study the security attributes of three categories, collectively and in contrast, against domain name-, content-, and SSL certificate-level features. We find that each type of hospital has a distinctive characteristic of its utilization of domain name registrars, top-level domain distribution, and domain creation distribution, as well as content type and HTTP request features. Security-wise, and consistent with the general population of websites, only 1% of government hospitals utilized DNSSEC, in contrast to 6% of the proprietary hospitals. Alarmingly, we found that 25% of the hospitals used plain HTTP, in contrast to 20% in the general web population. Alarmingly too, we found that 8%-84% of the hospitals, depending on their type, had some malicious contents, which are mostly attributed to the lack of maintenance.

We conclude with a correlation analysis against 414 confirmed and manually vetted hospitals' data breaches. Among other interesting findings, our study highlights that the security attributes highlighted in our analysis of hospital websites are forming a very strong indicator of their likelihood of being breached. Our analyses are the first step towards understanding patient online privacy, highlighting the lack of basic security in many hospitals' websites and opening various potential research directions.

2 Related Work

Previous studies have extensively examined healthcare data breaches in recent years, providing insights into their frequency and underlying causes. However, there has been relatively little research focused on analyzing the security configurations of healthcare providers. To provide a comprehensive understanding of our work, we conducted a review of previous studies pertaining to website content, security analysis, and healthcare data breach analysis.

2.1 Data Breaches Analysis

Several studies have recently investigated data breaches in the healthcare industry [5, 6, 60, 68]. For instance, Seh *et al.* [60] conducted a comprehensive analysis of HIPPA data breach reports. Their study highlights that hacking incidents, unauthorized access (internal), theft or loss, and improper disposal of unnecessary data are the main disclosure types of protected healthcare information. Moreover, the authors applied the Simple Moving Average (SMA) and Simple Exponential Smoothing (SES) time series methods on the data to determine the trend of healthcare data breaches and their cost to the healthcare industry. Choi *et al.* [14] estimated the link between data breaches and hospital advertising spending, studying the period of the two years following the breach and finding hospitals had much higher advertising expenditures. Siddartha *et al.* [61] found that the healthcare industry is being targeted for two main reasons: being a rich source of valuable data and its weak defenses.

Siddartha and Ravikumar [61] suggested that the security techniques employed in the healthcare industry miss data analysis improvements, e.g., data format preservation, data size preservation, and other factors. Luis *et al.* [68] defined DNS queries and TLS/SSL connections to identify the dangers encountered inside a hospital environment without disrupting the functioning network using two years of collected data. Another line of work, the 2022 Data Breach Investigations Report (DBIR) [23], investigates healthcare breaches among other industries. Based on the report, healthcare suffered 849 incidents, with 571 confirmed data disclosure in 2022. The report summarized various findings and determined that external actors are behind 61% of data breaches while 39% of data breaches involved internal actors. Furthermore, according to the same report, financial gain is the highest motive for attackers at 95%, followed by espionage at 4%. Raghupathi *et al.* [55] conducted a recent study that investigated data breach occurrences in healthcare provider environments specifically related to patient data using the U.S. Department of Health and Human Services publicly available dataset. Their study found a correlation between the occurrence of data breaches, breach locations, breach types, and the presence of business associates. Moreover, their study identified Hacking is the most common type of data breach, and Network servers are the most popular location for information breaches.

2.2 Hospitals Websites Analysis

Over the past few years, there has been a drastic increase in the development and utilization of online services and web applications. Paralleled with this rise has been an increasing concern over the privacy and security of these online services and applications, e.g., different components can be compromised, putting their users at risk. Chung *et al.* [15] offered the first in-depth analysis of incorrect certifications in the online Public Key Infrastructure (PKI), showing that most PKI certificates are invalid. The same study scrutinized the origin of the invalid SSL certificates and summarized that the preponderance of the invalid certificates was generated by end-user devices, with a periodical renewal of new self-signed certificates. SSL certificates have been investigated for website risk and vulnerability analysis [4, 7, 10–12, 15, 16, 37–40, 47, 78]. For instance, Meyer *et al.* [46] examined the SSL certificates' content and information to distinguish between phishing and benign websites.

Alabduljabbar *et al.* [2, 3] explored the SSL certificate-based structural differences between free and premium content websites and highlighted that 35.85% of the free websites' certificates have significant security issues, with 17% invalid, 7% expired, and 12% with mismatched domain names. Bach *et al.* [9] examined the content of hospital websites in three different countries and assessed them as information repositories or as interactive online communication means in three countries: (Bosnia-and-Herzegovina), recent (Croatia) and established EU member countries (Slovenia). In a recent study conducted by Yu *et al.* [77], a comprehensive analysis was performed on 19,483 hospital websites from 152 countries and provincial jurisdictions across Asia, Europe, North America, Latin America, Africa, and Oceania. The researchers utilized these crawled websites to investigate the presence of trackers, ultimately revealing some concerning findings. The findings showed that 53.5% of these websites used tracking scripts/cookies. Additionally, 33 websites were flagged as malicious, and 699 sites transmitted sensitive data to external servers through session replay services. This highlights concerns about privacy, security, and data protection on hospital websites.

3 Measuring Healthcare Data Breaches

3.1 Summary of Completed Work

In this work, we conducted an in-depth analysis of healthcare data breaches using an authentic dataset. We focused on analyzing the timeline of data breaches and their impact on health organizations. This work provides valuable insights into the nature of data breaches in the healthcare sector, highlighting the importance of addressing security vulnerabilities to protect patients, customers, and employees. The findings underscore the need for robust cybersecurity measures in health organizations to mitigate the risks posed by internal and external threats.

3.2 Introduction

The United States Department of Health and Human Services defines a data breach as an intentional or non-intentional use or disclosure of confidential health information. A data breach compromises privacy and security, resulting in a sufficient risk of reputation, financial, and other harm to the affected individuals [75]. Over the past few years, concerns related to healthcare data privacy have been mounting, since healthcare information has become more digitized, distributed, and mobile [36]. Medical records have transformed from paper-based into Electronic Health Records (EHR) to facilitate various digital system possesses. Medical EHR can be described as "a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. Included in this information are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports" [45]. EHR enhances patient care by enhancing diagnostics and patient outcomes, improving patient participation, enhancing care coordination, practicing efficiencies, and cost savings [52]. Despite the numerous benefits of EHR, the transformation has inflated the security and privacy concerns regarding patients' information. The growing usage of the Internet of Things (IoT) and intelligent devices affects the methods of communication in hospitals and helps patients quickly access their medical treatment whenever necessitated.

Nevertheless, the usage of such technologies is a fundamental factor that can cause security risks and lead to data breaches [63]. Broadly, healthcare data breaches are external and internal. External breaches are malicious, including at least one or more threat actions from cyber criminals, such as hacking, malware, and social attacks. On the other hand, internal data breaches typically occur due to malfeasance by insiders, human errors, and negligence from employees. Data breaches have increased in the past decade. In comparison with other industries, healthcare is the worst affected [42]. Cybercriminals are targeting healthcare for two fundamental reasons: it is a rich source of valuable data, and its defenses are weak [21]. Medical records contain valuable information such as victims' home addresses and Social Security Numbers (SSNs). Adversaries utilize such information for malicious activities and identity theft or exchange those medical records for financial profit on the dark web.

Contributions. For a better understanding of the landscape of healthcare data breaches against various attributing characteristics, we provide a detailed measurement-based study of the VERIS (Vocabulary for Event Recording and Incident Sharing) dataset. Among other characteristics, we temporally analyze data breaches and their growth over time. To understand the attacks' intent, we analyze the type of breaches over various security attributes and characterize the threat actions, highlighting the attack vector employed for the breach. We hope that those characterizations will shed light on the trend and the attack vectors, thus providing directions for mitigating those breaches.

3.3 Data Source and Temporal Analysis

The object of this paper is to conduct a measurement of healthcare data breaches to understand trends and motives. To accomplish that, we used trusted and reliable data called VERIS. In the past, there were numerous initiatives to accumulate and share security incidents. Nonetheless, commitment and participation have been minimal. Reasons behind that are many, including (i) the difficulty of categorization, and (ii) the uncertainty of what to measure [69]. To facilitate data collection and sharing, VERIS is established as a nonprofit community designed to accommodate a free source of a common language for describing security incidents in a structured and repeatable way [69]. Due to the prevailing lack of helpful information, the VERIS dataset is an effective solution to the most critical and persistent challenges in the security industry. VERIS tackles this problem by offering organizations the ability to collect relevant information and share them responsibly and anonymously.

VERIS and Incident Attributes. VERIS's primary purpose is to create an open-source database to design a foundation that constructively and cooperatively learns from their experience to ensure a more reliable measurement and management risk system. VERIS is a central hub whereby information and resources are shared to maximize the benefits of contributing organizations. During the incident collection process, the VERIS community focuses on successfully implementing an intersection, namely the 4A's, which indicate the following: who is behind the incidents (actors), the action used by the adversary (actions), devices affected (assets), and how are they effected (attributes). An example of the 4A's for an incident can be as follows: internal (actor), hacking (action), network (asset), and confidentiality (attribute). VERIS designers estimate the needed information to be collected about an incident based on the level of threat, asset, impact, and control. Understanding these risk aspects enables organizations to improve their management systems and make informed decisions. The power of VERIS is the collection of evidence during and after the incidents, besides providing helpful metrics to maximize risk management.

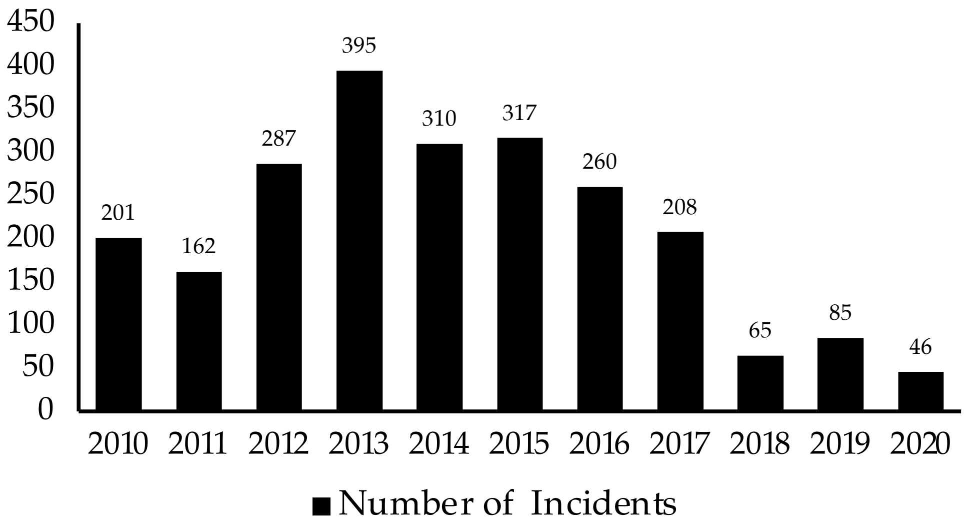


Figure 1: The yearly distribution of Data Breach Incidents.

3.3.1 Distribution of the Incident's Timeline

We analyzed the timeline mapping of the incidents across the years. The VERIS dataset contains incidents that took place between the years 1971 and 2020. While a long period of time is considered, the time frame from 1971 until 2010 seemed to contain a low number of incidents, with only 272 (total) of them, per the VERIS dataset. Thus, to understand the actual trend in the active region, we limit ourselves to the years 2010 onward. This analysis is essential because it provides us with insights into the active period of breaches and attacks, and could hint at the underlying ecosystem. To this end, and upon this analysis, we found out the following (1) the per year number of incidents follows a normal distribution, with the peak in 2013. (2) 2013 was the highest year in the number of incidents, with 395 (16%), followed by 2015 with 317 (13%), and 2014 with 310 ($\approx$13%). (3) Contrary to the common belief that the number of attacks is increasing, we found that the number of breaches has been decreasing since 2013, per VERIS reporting, as shown in Figure 1.

Takeaway. *There is a decrease in the number of incidents per year, possibly due to the lax reporting. This decrease, however, does not preclude the possibility that each of those breaches is getting more severe than past breaches.*

3.3.2 Timeline Discovery for Data Breaches

Health organizations encounter various difficulties in attempting to keep patients' medical records safe. The *timeline discovery* affects both the patient and the hospital. The longer it takes for an organization to discover a data breach, the more significant harm it can cause. The damage cannot only result in data loss or the disclosure of information but also includes businesses. In the literature, it was shown that organizations take 197 days to identify a data breach and 69 days to contain it, on average [32]. That amount of time to detect a data breach is considered long and costs organizations millions of dollars. An organization containing the data breach incident in less than 30 days from the date it happened can save up to $1 million compared to others who fail to do so, per the same study. In healthcare, hospitals and organizations can suffer many consequences due to a data breach, including lawsuits from the affected individuals, as well as reputation and trust loss. In addition, healthcare organizations incur significant costs in fixing the problem and protecting patients from additional harm.

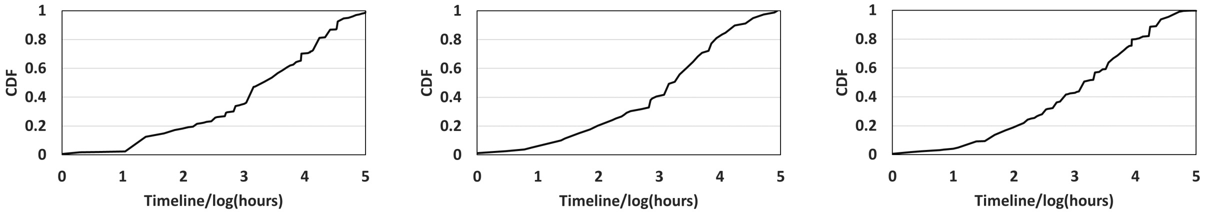
We examined the response time for incidents affecting victims, patients, customers, and employees. In the following, we present the results and contrast.

Results. We began by converting the timeline discovery into one unit (hours). Then, we calculate the cumulative distribution function (CDF) of incidents. Due to the extensive range of timelines, we used the logarithmic function to the discovery time range for simplicity and visibility, as shown in Figure 2. Based on this analysis, we noticed that the discovery time of incidents for employees is significantly faster than for customers and patients. As we can see in Figure 2, we discovered that 20% of the incidents for employees were discovered within four days or less. It took five days or less to discover the same percentage for customers, and up to six days to discover that for patients. Such results indicate the difference between the different categories of breach discovery time, and perhaps the priorities associated with their discovery and protection, although all are relatively high. To further establish that, for 50% of the incidents, the discovery time was 2, 2.5, and 3 months for customers, employees, and patients, respectively. The patients represent most victims with 41%, and the discovery time for their data breaches extends to years (14 years to discover 100% of all incidents). While discovering 100% of incidents for customers require a longer time: up to 21 years. On the contrary, the discovery time of incidents for employees is much less because discovering 100% of the incidents for this category is about ten years.

Takeaway. *Incidents discovery, even for most protected victims, can take many years, highlighting the lax security posture of healthcare organizations.*

3.4 Security Attributes

The VERIS dataset uses pairs of the six primary security: confidentiality/possession, integrity/authenticity, and availability/utility as an extension of the CIA triad. In this section, we attempt to investigate the compromised security attributes during the incidents by conducting the following: (i) analyzing the confidentiality leakage that occurred during data breaches, (ii) presenting the different data types, and noting which is the most targeted by adversaries, (iii) determine the state of



(a) Patients

(b) Employees

(c) Customers

Figure 2: CDF for the timeline discovery of different victim types.

the compromised data at the time of the incidents.

Data Confidentiality. Confidentiality refers to the limit of observations and disclosure of data [69]. We start by examining the data confidentiality leakage that occurred during data breaches. This analysis is necessary because it examines the amount of compromised data and their varieties throughout the incidents. Using the VERIS dataset, we found that 1,045 out of total data, 1,937 incidents had *information disclosure*, representing 54% of the total incidents, 882 had a *potential information disclosure*, representing 46%, while only two incidents that had *no information disclosure* at all and eight incidents are *unknown*.

We analyzed data that attackers often target. Based upon this analysis, we discovered the following: medical information exposed to higher disclosure compared to the other types of information, encompassing 1,413 incidents, representing 73%, while personal information appeared in second place, with 345 incidents, representing 18%. Lastly, payment information appeared in third place, having 61 incidents, representing 3%. Other targeted information include *unknown* (44; 2%), *banking* (33; 2%), *credentials* (23; 1%), and *others* (18; 1%).

Takeaway. *Despite their variety in breaches, medical and personal information are the most targeted, with 91% of the incidents combined.*

Status of Breached Data. During the exposure or compromise process, we investigated the state of the data and whether it was encrypted, transmitted, or stored unencrypted during the attack. This categorization aims to understand the security controls while the data is at rest or in motion due to transformation. As a result of this investigation, we noticed 36% of the data was *stored unencrypted*, 30% *stored*, 25% *unknown*, 3% *printed*, 2% *transmitted unencrypted*, and 4% with other attributes.

Takeaway.*The majority of breached data does not employ basic security functions, making it an easy target to adversaries for exploitation at rest or in transit.*

Data Integrity and Authenticity. Integrity refers to an asset or data to be complete and unchanged from the original state, content, and function [69]. Example of loss of integrity includes but is not limited to unauthorized insertion, modification, and manipulation. We wanted to discover the varied nature of integrity loss. Each time incidents occur, there can be at least one integrity attack. However, many losses can be associated with a single incident. Following the analysis, we noticed that most data integrity losses are due to altering behaviors containing 93 incidents, representing 31% of the overall. Software installation comes in second with 91 incidents, representing 30% of the known reasons. Other integrity related attacks include *fraudulent transmission* (18%), *data modification* (11%), *re-purposing* (3%), and *others* (6%).

Authenticity refers to the validity, conformance, correspondence to intent and genuineness of the asset (or data). Losses of authenticity include misrepresentation, repudiation, misappropriation, and others. Short definition: Valid, genuine, and conforms to intent [69]. Based upon this analysis, we observed that the authenticity state was poorly reported at the time of the incidents.

Data Availability. Availability refers to an asset or data being present, accessible, and ready for use when needed [69]. A loss to availability includes deletion, destruction, and performance impacts such as delay or acceleration. We will show the variety of data available that might happen during the incidents. This analysis is necessary to understand the nature or type of availability or utility loss. Based on this analysis, we found that 769 incidents contained a loss of data regarding their effect on availability, representing 90% of the total incidents with the reported attribute. *Obfuscation*, and *interruption* are reported as remaining causes affecting availability, with 9% and 1% of all incidents, respectively.

Takeaway. *Despite limited reporting, more than 20% of all the studied incidents suffer from integrity and authenticity attacks, due to a range of factors, magnifying the potential of attacks without data leaving the organization.*

3.5 Analyzing the Threat Actors

Threat actors are entities that can cause or contribute to an incident [69]. Each time an incident happens, there can be at least one of the three threat actors involved, but on some occasions, there can be more than one actor involved in a particular incident. Threat actors' actions can be malicious, non-malicious, intentional or unintentional, causal or contributory [69]. VERIS classifies threat actors into three main categories, namely: external, internal, and partner. This classification excludes the contributory error that unintentionally occurs. For instance, if an insider unintentionally misconfigures an application and leaves it vulnerable to an attack. The insider would not be considered as a threat actor if the applications were successfully breached by another actor [69].

On the other hand, an insider who deliberately steals data or whose inappropriate behavior (e.g., policy violations) facilitated the breach would be considered a threat actor in the breach [69]. This

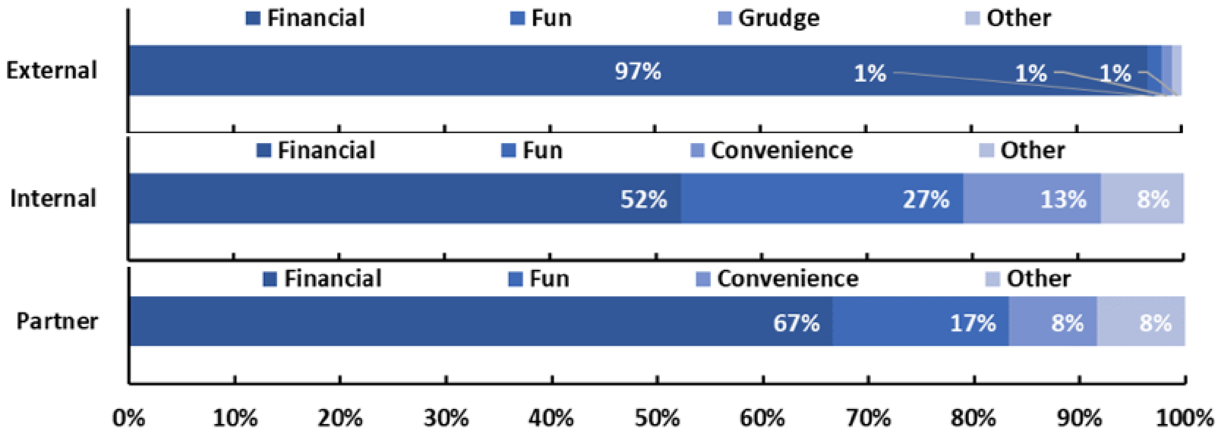


Figure 3: Threat actor's motives for external, internal, and partner actors.

section will explain and analyze each category of threat actors with their presence in incidents from our dataset. This analysis is essential because of the following reasons: (i) it provides us with an understanding of the reasons or motives that can lead actors to act, (ii) the analysis can provide knowledge for organizations to consider proper precautions to defend against how threat actors operate. Several motives can be a reason for a data breach, such as fear, ideology, grudge, espionage, convenience, fun, and financial. Based upon this analysis, we noticed that the financial motive is the primary motive for adversaries, followed by looking for fun.

External Actors. External threats originate from outside of an organization and its third-party partners [69]. Examples include criminal groups, lone hackers, former employees, and government entities. It is also comprised of God (as in "acts of"), "Mother Nature," and random chance. Typically, no trust or privilege is implied for external entities. We found out that 97% of the external actor motives are financial, and 1% are for fun. Figure 3 shows the different motives of the actor's external motives.

Internal Actors. Internal threats originate from within the organization, which encompasses full-time company employees, independent contractors, interns, and other staff. Insiders are trusted and privileged (some more than others). Upon further analysis, we found that 52% of the internal motives for adversaries are financial, while 27% are for fun. Figure 3 presents the distribution of motives for internal motives.

Partner Actors. Partners include any third party sharing a business relationship with the organization, including suppliers, vendors, hosting providers, outsourced IT support, and others. Some level of trust and privilege is usually implied between business partners [69]. Based on this analysis, we found out that most of the motives behind the incidents are financial 67%; fun and convenience are 17% and 8%, respectively. The remaining results for the internal motives distribution are shown in Figure 3.

Results: Data Breaches Victims. We analyzed the most targeted victims from adversaries according to the number of incidents. Reasons often differ as to why these victims have been targeted, and it also depends on several other aspects, such as location, specific personal information, or a high number of patients in a hospital. We found that most of the targeted victims were patients (88%), the customer came in second (5%), and 5% for employees. Other types of victims include students (interns) working inside healthcare organizations or third-party companies that share data with a specific entity. Figure 4 shows the most targeted victims in the incidents.

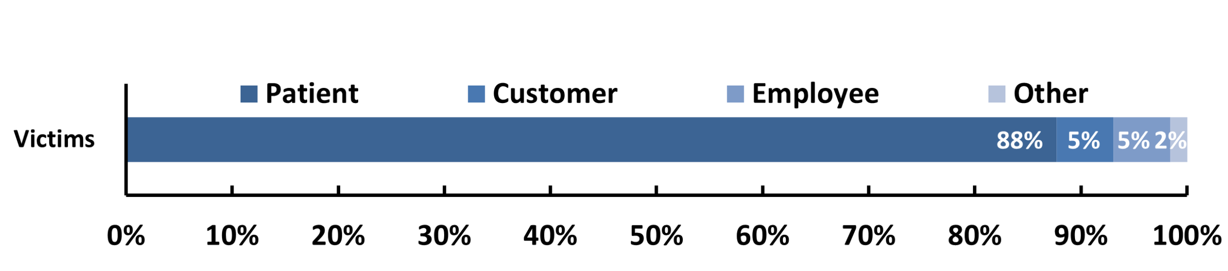


Figure 4: Distribution of incidents by targeted victims.

Takeaway. *For the three threat actors combined, financial gain is the primary motive for adversaries to launch their attacks.*

3.6 Analyzing Threat Actions

In this section, we introduce our measurement and analysis of the threat actions used by adversaries during a data breach. This investigation intends to provide insight and the causes of threat actions and their occurrences in our dataset. The following section discusses the two types of threat actions: the different action varieties and the most used vectors by adversaries during an attack. The VERIS dataset classifies threat actions into seven primary categories: malware, social, hacking, misuse, physical, error, and environmental. Analyzing threat actions is essential due to the amount of risk associated with each of them every time an incident occurs. Generally, an incident usually contains a least one of the threat actions; however, most of the incidents will comprise multiple actions that often come with numerous categories.

Terminology Definitions. Below, we define several types of threat actions.

Malware Malicious software or malware is a computer code designed to disable, disrupt, or take control of the computer system by altering its state or function without the owner's informed consent [69]. Malware exploits technical flaws or vulnerabilities in hardware or software.

Hacking Refers to all attempts to intentionally access or harm information assets without (or exceeding) authorization by circumventing or thwarting logical security mechanisms. It includes brute force, SQL injection, cryptanalysis, denial of service attacks, etc. [69].

Social Social engineering criminals strive to exploit the users of these technologies by pretending to be something they are not to persuade others. Attackers utilize the trust to their advantage by misleading users into disclosing information that compromises data security. Social engineering tactics employ deception, manipulation, intimidation, and other techniques to exploit the human element, or users, of information assets, including pretexting, phishing, blackmail, threats, scams, etc. [69].

Misuse The use of entrusted organizational resources or privileges for any purpose or manner contrary to the intended is considered misuse. It includes administrative abuse, use policy violations, use of non-approved assets, etc. [69]. These actions can be malicious or non-malicious.

Physical Encompass deliberate threats that involve proximity, possession, or force. These include theft, tampering, snooping, sabotage, local device access, assault, etc. [69]. Natural hazards and power failures are classified into physical actions. However, VERIS restricts these events to intentional incidents only caused by human actors.

Error Error broadly encompasses anything done (or left undone) incorrectly or inadvertently. It includes omissions, misconfigurations, programming errors, malfunctions, etc. [69]. It does not include any intentional incidents.

Environmental The environmental category includes natural events such as earthquakes and floods and hazards associated with the immediate environment or infrastructure in which assets are located. The latter encompasses power failures, electrical interference, pipe leaks, and atmospheric conditions.

Results: Threat Actions Analysis. We measured the existence of each threat action category by calculating their varieties and vectors used in the incidents. We observed that ransomware represents 82% of the malware threat, followed by others 8%. VERIS "other" to define any enumeration not represented by one of the categories in the data set. For the social threat actions category, with a percentage of 69%, phishing plays a large part in threat actions. The use of stolen credentials represents 80% of the hacking threat actions. With an increase in the number of employees, errors increased. Loss errors represent the main factor in this threat actions category representing 28%, followed by a disposal error of 27%. It is worth noticing that theft is in the physical threat actions category with a percentage of 96%. Finally, privilege abuse in the misuse category, with a rate of 59%, is behind most of the threat actions in these two categories.

On the other hand, when we analyze the threat action vectors as shown in Figure 6, we found out that the direct install represents 45% of the malware threat actions. Email attachment is the second most common malware breach vector with 32%. The email vector represents 81% in the

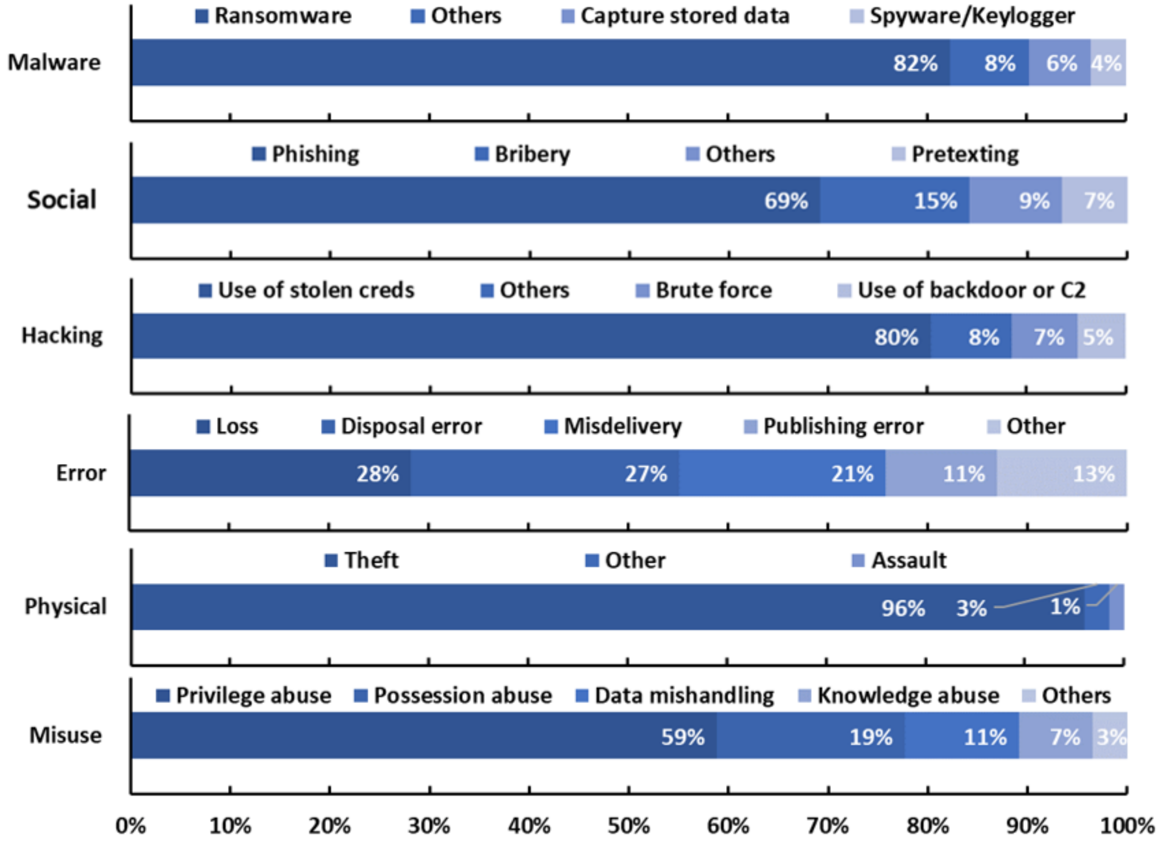


Figure 5: The variety of threat actions employed in data breach incidents.

social category for other categories of threat actions, and web applications represent the primary vector with 81% of all hacking threat actions.

Carelessness is the primary vector with 92% of the error category. Although most data breaches using hacking by threat actors involve brute force, or the use of lost or stolen credentials [22], At the same time, LAN access is the most effective vector in the misuse category with 65%. It is clear that email and web application vectors represent the highest percentages among other vectors, and this is associated with the shift of valuable data to the cloud, including email accounts and business-related processes [22].

Takeaway. *Despite the variety of vectors, ransomware is still the leading malware method involving 82% of the incidents.*

3.7 Summary of Completed Work

While analyzing the timeline of the data set that comprised all the data breaches, the results showed that the highest number of incidents occurred from 2010-2020. Moreover, this long-term study re-

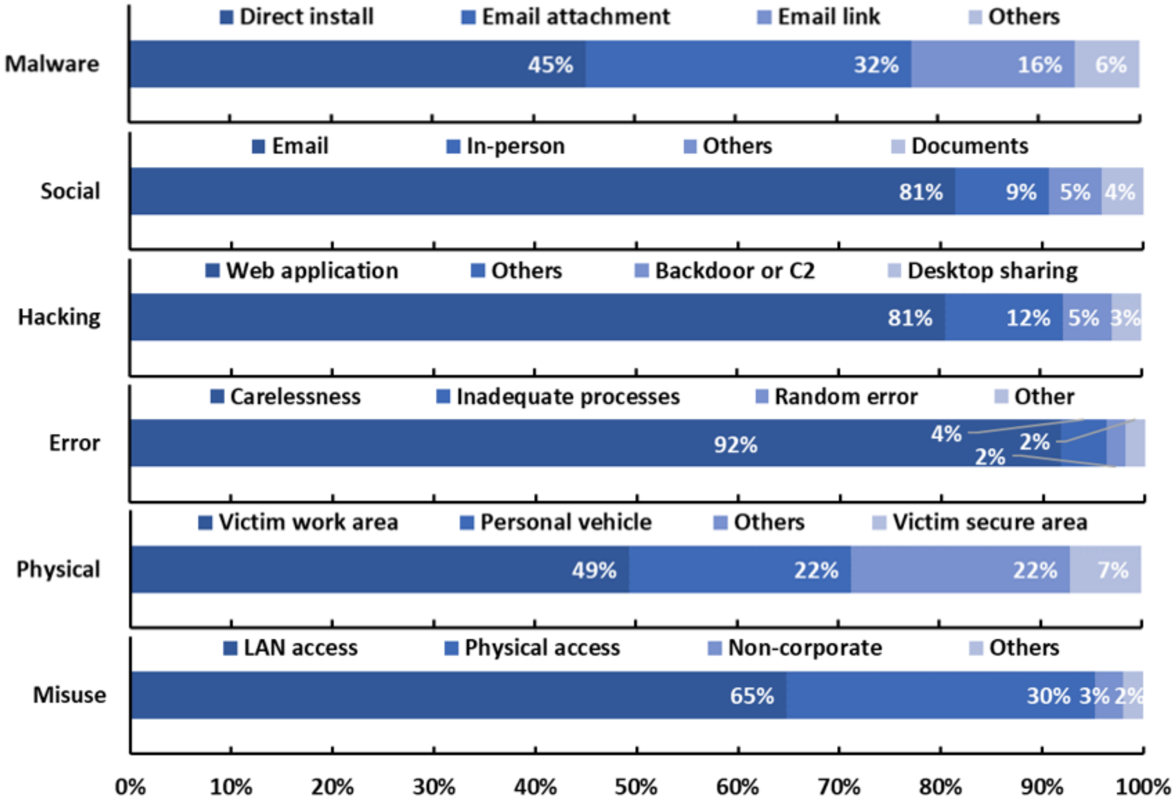


Figure 6: The threat action vectors employed in data breach incidents.

vealed that health organizations are exposed to internal, external, and partner attacks. The financial is the primary motivation for the external, internal, and partner attackers. Without a doubt, there is a high cost associated with data breaches; the price for each stolen health record increases with time. Based on a long-term analysis of the data set, the actions used by the threat actors are classified into seven categories: malware, hacking, social, misuse, physical, error, and environmental. Ransomware motivated 82% of malware threat actors, and 45% of malware threat actions are directly installed. In the future, it would be worthwhile examining the correlation between security breaches and other indicators, including GDP, hospital size, etc.

4 *Security Breaches in the Healthcare Domain*: A Spatiotemporal Analysis

4.1 Summary of Completed Work

In this work, we focused on studying the factors influencing data breach incidents in the healthcare sector. Our study revealed that the number of adults and the population of a state played a significant role in the exposure to data breaches, with California, Florida, and Texas being the primary targets. Additionally, our analysis revealed that the media group was the most breached asset, followed by the Server and User group. Interestingly, a majority of the incidents occurred in small-size organizations (57%), while (43%) of the incidents took place in large organizations, indicating that larger healthcare organizations tend to have better security systems. This work provides insights into the factors influencing data breach incidents in the healthcare sector. The findings emphasize the need for robust security measures in smaller healthcare organizations and highlight the need for prompt detection and response to mitigate the impact of data breaches.

4.2 Introduction

Electronic health records (EHR) can be described as "a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. Included in this information are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports" [45]. The adoption of EHR improves the healthcare industry and patients alike, and the transformation of healthcare organizations from paper-based to digital has increased healthcare quality by improving patient care and participation, care coordination, diagnostics, patient outcomes, and practice efficiency. However, despite the numerous benefits of EHR, this transformation has led to numerous privacy and security issues which may arise from vulnerabilities (e.g., software vulnerabilities, insider threats, human error, etc.), increasing the possibility of cyber-attacks [35]. The alarming surge in healthcare data breaches has caused huge concerns in the healthcare sector due to the illegitimate and unauthorized disclosure of private healthcare data [5, 60].

Healthcare Data breaches can be classified as either internal or external, and they can occur as a result of theft of private health records, hacking, loss of sensitive patient data, and unauthorized access to patient's private information [76]. External cybersecurity incidents are typically committed by cybercriminals operating in the dark web, while internal data breaches result from something internal to an organization, such as disgruntled employees, malicious insiders, employee negligence, and human error. Patient medical records and personal information are often targeted in healthcare data breaches due to their sensitivity and value. External attacks aim to steal those records and demand a ransom or sell those records for hundreds of dollars per single patient on the dark web [62].

Data breaches are devastating and can cause significant damage to healthcare organizations; all the research in this domain demonstrates that the healthcare industry is the most targeted sector due to the attractive financial return of selling sensitive patient records on the dark web [73]. Additionally, the lenient security controls deployed by healthcare organizations further complicate matters and make the healthcare domain a favorite target for hackers. The cost of recovering from such breaches varies greatly by the nature of the incident and the number of compromised health records. To better understand the cost aspect, we can break down the cost of data breaches for healthcare entities into two categories: direct costs and indirect costs. Direct expenses include activating incident response teams, engaging forensic experts, outsourcing hotline support, and providing free credit monitoring subscriptions and discounts for future products and services. On the other hand, indirect costs include in-house investigations and communication, as well as the extrapolated value of customer loss resulting from turnover or diminished customer acquisition rates [30]. Given these facts, it's compelling to conduct extensive research studies into the causes, effects, and consequences of healthcare data security incidents. Perhaps more importantly, gaining insights into the different trends and the landscape, and understanding, analyzing, and measuring the statistics in data breaches is crucial for combating such incidents. This is the motivation of this paper, and we also wish to motivate the research community in this space to extend the body of knowledge by conducting more studies to be able to understand data breaches better and propose solutions in the fight against cybercrimes.

Contributions. To understand the landscape of healthcare data breaches against several attributing characteristics, we provide a detailed measurement-based study of the VERIS (Vocabulary for Event Recording and Incident Sharing) and the Office of Civil Rights (OCR) datasets. To understand attackers' intents and motives, we analyze the type of assets targeted during breaches over various characteristics to investigate their effect. We also analyzed data breaches considering multiple views looking at their distribution, affected entities, breached information, location of the breach, etc.

4.3 Data Sources

One of the challenges with analyzing cybersecurity incidents, in general, and in the healthcare sector, in particular, is that most datasets are proprietary [72]. Additionally, most breached healthcare organizations shy away from disclosing their vulnerabilities after a breach due to a variety of concerns, including public image, reputation, and patient trust. The other challenge lies in the fact that each victim healthcare entity tends to take a different approach in analyzing and documenting a data breach [73]. This, in turn, complicates research efforts because data breach statistics are not stored in a central online repository and are thus inaccessible to the broader research community. To address the above challenges and conduct our measurements and analysis of data breaches, we turn to the largest publicly available datasets of cybersecurity incidents, namely, the VERIS dataset and the OCR dataset, which we describe below.

VERIS. We obtained a reliable data source to conduct our research, namely, the Vocabulary for Event Recording and Incident Sharing (VERIS). Veris provides a common language for reporting data breaches incidents in an organized and repeatable manner [43]. Thus, Veris plays a significant role in providing a solution to one of the most critical and persistent challenges in the security industry; lack of quality information. Veris contributes to the solution of this problem by helping organizations collect helpful incident-related details and share them anonymously and responsibly with others. Veris's primary goal is to lay a foundation to constructively and cooperatively learn from our experiences to ensure the proper measurements and managing risk [13].

Office of Civil Rights (OCR). Our second dataset is obtained from the U.S. Department of Health and Human Services Office of Civil Rights. The U.S. Department of Health and Human Services (HHS) Office for Civil Rights (OCR) enforces federal civil rights laws, conscience, and religious freedom laws, the Health Insurance Portability and Accountability Act (HIPAA) Privacy, Security, Breach Notification Rules, and the Patient Safety Act and Rule, which together protect your fundamental rights of nondiscrimination, conscience, religious freedom, and health information privacy [53]. The OCR has its breach portal, where data breaches are reported. The website contains data breaches that are currently under investigation within the last 24 months by the OCR. There is also an archived dataset where resolved data breaches and/or those older than 24 months are archived. All the data breaches reported by the OCR are in the U.S. only. Additionally, all records in the subsequent data breaches affect 500 or more individuals; the OCR does not report minor data breaches that affect less than 500 individuals.

4.4 Studied Dimensions and Variables

This study aims to examine healthcare data breaches considering different aspects of threat characterization and modeling.

- **Geographical mapping:** Section 4.5.1 analyzes the geographical mapping and distribution of incidents around the world. Analyzing the geographical mapping of the incidents is necessary for several purposes: (i) it provides us with an understanding of the areas most targeted by adversaries for an affinity characterization, (ii) identifying locations around the world where the number of incidents varies due to valuable medical information, particular age group, banking details, etc. We can use this analysis for correlation and prediction capabilities.

- **State-level distribution:** Section 4.5.2 measures the state distribution of incidents in the U.S.

This analysis is necessary for (i) identifying the hot spots targeted by attackers and (ii) conducting correlation analysis between states.

- **Compromised assets**: Section 4.5.3 details the targeted assets by breaches such as media, server, terminal, etc. Alongside, we will categorize the assets into groups, then dive into their varieties by an individual group against the number of incidents.

- **State-level correlation:** Section 4.5.4 carries a correlation analysis of the number of incidents within the top ten states with characteristics such as population, Gross Domestic Product (GDP), number of adults, etc. This correlation provides us with essential insights into the reasoning and bearings for each state.

- **Discovery methods:** Section 4.5.5 aims to identify the discovery mechanisms used by healthcare entities. Then, we will measure the reported tools and their use in data breaches in our dataset. This analysis can help with determining the appropriate tools needed to be implemented in organizations.

- **Adversary demography — The threat intent:** Section 4.5.6 measures the intention of attackers during data breaches. We intend to acknowledge whether the incidents are targeted or opportunistic.

4.5 Measurement Results and Discussions

4.5.1 The Global Distribution of Incidents

Mapping incidents is explicitly provided in our dataset. The dataset uses the ISO 3,166 country codes for each country variable [26], where the codes are generated based on the physical location of the hospital targeted by the attack. Based upon this analysis, we discovered that 1,955 incidents out of the total incidents (2,407) had taken place in the United States, representing 81% of the total incidents. The United Kingdom comes in second, with 157 incidents, representing 7%, and Canada comes in third with 152 incidents, representing only (6%). Figure 7 presents the results for the remaining highest ten countries, while the rest of the world represents (2%) comprising 58 incidents.

As a result of the geographical mapping analysis, we decided to conduct our in-depth analysis study on the United States since most incidents occurred in this country. Several reasons explain why the majority of the incidents are in the United States. First, the Health Insurance Portability and Accountability Act (HIPAA) requires healthcare entities to notify the Department of Health and Human Services (DHHS) whenever a data breach occurs. Second, covered entities must notify affected individuals following the discovery of a breach of unsecured protected health information [53]. In addition to that, covered entities must notify the Secretary of breaches of unsecured

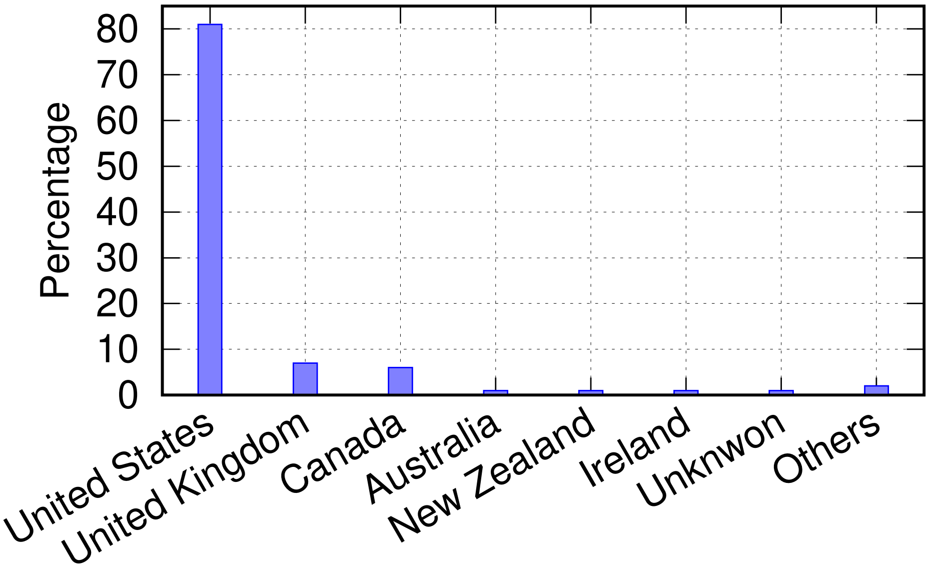


Figure 7: The geographical distribution of data breach incidents by country.

protected health information if the affected individuals are 500 or more [53]. Third, covered entities that experience a breach affecting more than 500 residents of a State or jurisdiction are, in addition to notifying the affected individuals, required to provide notice to prominent media outlets serving the State or jurisdiction [53]. Moreover, breach notification is also required for vendors, and third-party service providers under the Health Information Technology for Economic and Clinical Health Act (HITECH) [44]. Finally, the HIPAA Security Rule requires healthcare organizations to create a risk management plan protecting all personal health data against security incidents (Office of Civil Rights 2015), which may explain the significant number of reported incidents in the United States [1].

4.5.2 Number of Incidents by State

Following the global distribution of incidents, we moved into the mapping of incidents on the state level. We analyzed the number of incidents by state. As a result of this analysis, we noticed that California is the highest state, with the number of incidents comprising 241 incidents, representing 24% of the overall. Florida comes in second with 147 incidents, representing 15%, and Texas with 145 incidents, representing 14%. Figure 8 shows the remaining results of this analysis.

4.5.3 Analyzing the Compromised Assets

This section investigates the compromised information assets in the Veris dataset. We harnessed the power of Natural Language Processing (NLP) models to help with analyzing the data gathered

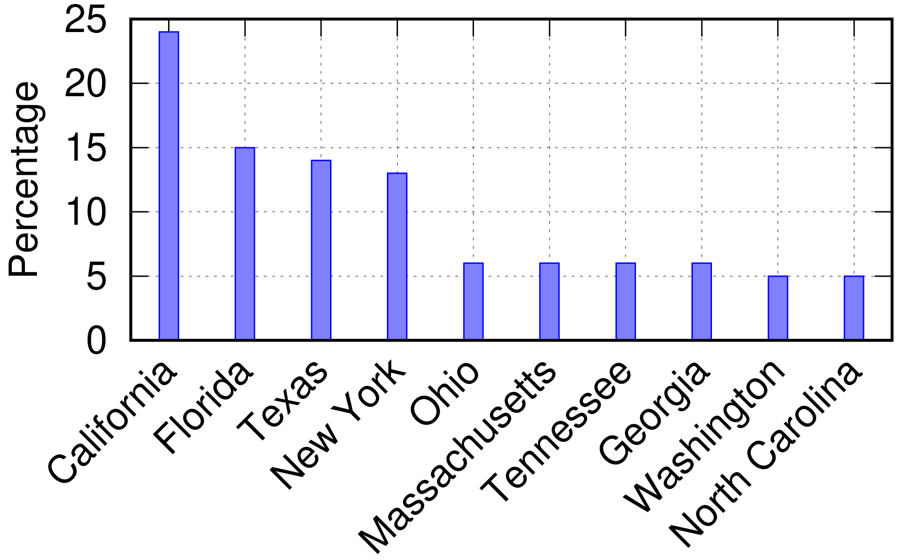


Figure 8: The distribution of data breach incidents across different states.

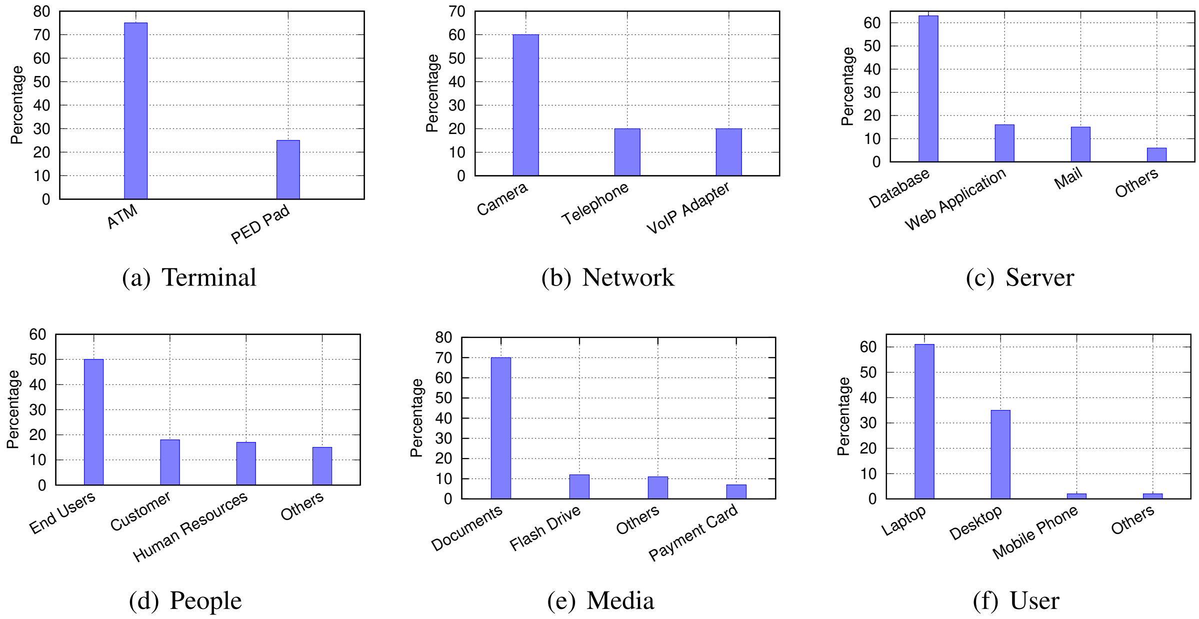
from breaches. Information assets fall into six main groups: media, server, terminal, network, user, and people. Each group comprises different varieties [59]. First, the network group includes access control readers such as badge and biometrics, camera or surveillance system, firewall, intrusion detection systems (IDS) or intrusion prevention systems, and others. Second, the media group comprises disk media such as CDs or DVDs, flash drives or cards, hard disk drives, identity smart cards, and others. Third, the people group includes the administrator, auditor, cashier, customer, former employee, guard, and others. Fourth, the server includes authentication, backup, database, Dynamic Host Configuration Protocol (DHCP), DNS, mail, and others. Fifth, the terminal group includes an automated Teller Machine (ATM), detached PIN pad or card reader, gas "pay-at-the-pump" terminal, self-service kiosk, and others. Finally, the user group includes an authentication token or device, desktop or laptop, media player or recorder, mobile phone or smartphone, and many others.

The existence of assets depends on several reasons and conditions during each incident. We will measure each asset group based on their occurrences in the incidents, and then, we get into the measurement of their varieties to look into the most targeted type of each asset group. This analysis is essential, and its primary purpose is to adequately describe the incidents, assess control weaknesses and vulnerabilities, determine impact, and identify mitigation strategies.

Usually, during a data breach incident, one or more assets get compromised by hackers [27]. A compromised asset refers to any loss of confidentiality, integrity, or availability during or after the incidents. In the following section, we seek to analyze and measure the asset groups and the total incidents for each group; then, we move to their different asset groups. Based on this analysis, we

Table 1: The distribution of incidents by asset group type during data breaches.

Asset Group Type	# Incidents	Percentage
Media	564	33.97%
Server	560	33.73%
User	493	29.69%
People	34	2.04%
Network	5	0.30%
Terminal	4	0.24%
Overall	1660	100 %



(a) Terminal

(b) Network

(c) Server

(d) People

(e) Media

(f) User

Figure 9: The varieties of information asset groups involved in data breach incidents.

noticed that media assets are the clear leader comprising 564 incidents out of the overall, representing 33.97%, and the server comes in second, comprising 560 incidents, representing 33.73%. Table 1 shows the remaining asset categories and their number of incidents.

After measuring the number of incidents for each asset group as a whole, we moved into measuring their varieties. Based on the analysis done, we found that 61% of the incidents in the user group are through laptops, followed by the terminal group with 75% of the incidents through ATMs. In the server asset group, we found out that 63% of the incidents happened through exploiting the database. While for the people asset group, 50% of the incidents are because of the end-user. Most of the incidents that happen in the network are throughout cameras, representing 60%. Lastly, 70% of the incidents in the media group are through documents. In Figure 9, we present the remaining results for the other asset groups and their varieties.

Table 2: State level correlation. Numbers of incidents (I), hospitals (H), employees (E), staffed beds (B), GDP (G), population (P), and adults (A) are considered.

	I	**H**	**E**	**B**	**G**	**P**	**A**
I	1.00						
H	0.88	1.00					
E	0.92	0.91	1.00				
B	0.94	0.92	0.97	1.00			
G	0.95	0.86	0.92	0.89	1.00		
P	**0.96**	0.95	0.94	0.94	0.95	1.00	
A	**0.96**	0.88	0.94	0.96	0.89	0.90	1.00

4.5.4 State Level Correlation

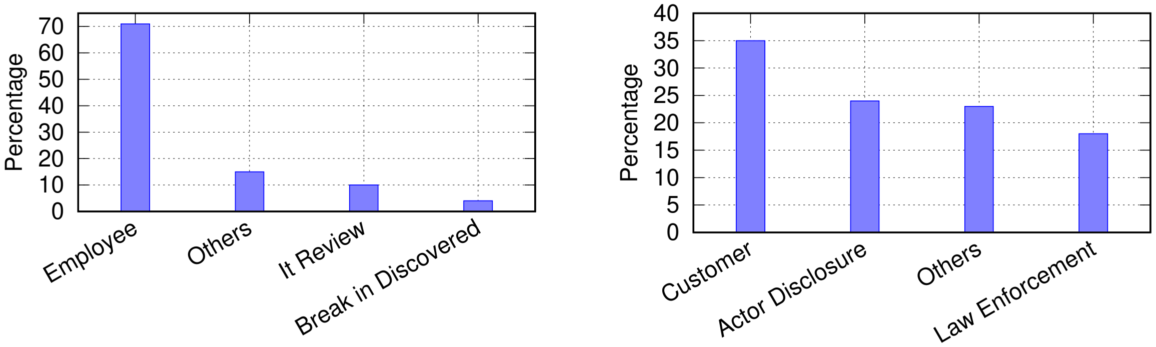
This section will conduct a state-level correlation between the number of reported incidents and hospitals, staffed beds, population, and gross domestic product (GDP) for the top 10 states. GDP is the gross domestic product and is represented in billion U.S. dollars. To address the following question, we conducted a state-level analysis considering these factors related to the reported incidents in our dataset. We decided to run this analysis on the highest ten states in terms of the number of reported incidents. We started by collecting the specified statistics for each state, including population, GDP, staffed beds, and hospitals. The relationship between two variables can be a positive relationship (1), no relationship (0), and an inverse relationship (-1). Upon this analysis, we discovered that the population and adults are highly correlated with the number of incidents (0.96). Followed by the GDP (0.95). The remaining results of the correlation are shown in Table 2.

4.5.5 Internal and External Discovery Methods

Discovery methods fall into two main categories; internal and external. Organizations use several tools to discover an incident depending on the type of data breach. External and internal data breaches are different, and each one of them requires special discovery tools. First, healthcare organizations use numerous tools to discover incidents for internal incidents, such as Host IDS or file integrity monitoring, network IDS, and IPS alerts. In contrast, practices including law enforcement, actor disclosure, and customer notifications can help discover external incidents. Our analysis found that most of the internal incidents are discovered by employees, representing 71% of the total incidents. In contrast, customers discover 35% of the external incidents, and actor disclosure comes in second, representing 24%. The remaining results of this analysis are shown in Figure 10.

4.5.6 Targeted vs Opportunistic

To understand the nature of the data breach incidents and whether they are intentional or non-intentional, we conducted a measurement analysis to investigate the number of targeted incidents



(a) Internal discovery.

(b) External discovery.

Figure 10: Comparison of data breach discovery methods.

and opportunistic ones. This classification is uniquely relevant to deliberate and malicious actions. There are two main categories: targeted and opportunistic. First, opportunistic incidents occur when the victim exhibits a weakness that the actor has the knowledge to exploit. Second, targeted incidents happen when the adversary chooses the victim as a target, and then the actor will investigate possible vulnerabilities to exploit. Using our exclusively given records in our dataset, we found that more than half of healthcare data breaches are opportunistic, representing 80%, while, on the other hand, 20% are targeted.

4.6 Analysis of the OCR Dataset

Type of Breach. We analyzed the causes of healthcare data breaches based on the reported incidents and observed that most incidents occur due to hacking or IT-related disclosure, comprising 1,069 incidents, representing 31% of the overall incidents. Unauthorized access and disclosure came in second, holding 934 incidents overall, representing 27%. Finally, the theft category came in third place, comprising 909 incidents, accounting for 26% of the total incidents.

State Distribution. The following section addresses the distribution of the incidents for the U.S. states. Using the OCR data, we measured the incidents for each state; this analysis is essential for trends and comparison. Following this analysis, we have observed that states with a large population, high Gross Domestic Product (GDP), and large adult population are more targeted than others, as shown in section 4.5.4. California was the most affected, totaling 357 incidents, followed by Texas with 279 incidents, while Florida was the third largest with 215 incidents.

Distribution of Incidents by Year. Using the ORC dataset, and over the period between 2009 and the time of conducting this study in 2021, we measured the reported incidents in the dataset affecting 500 or more victims and reported them to the HHS OCR. Following this analysis, we notice that the number of incidents surged over time, indicating a lack of implementing stringent

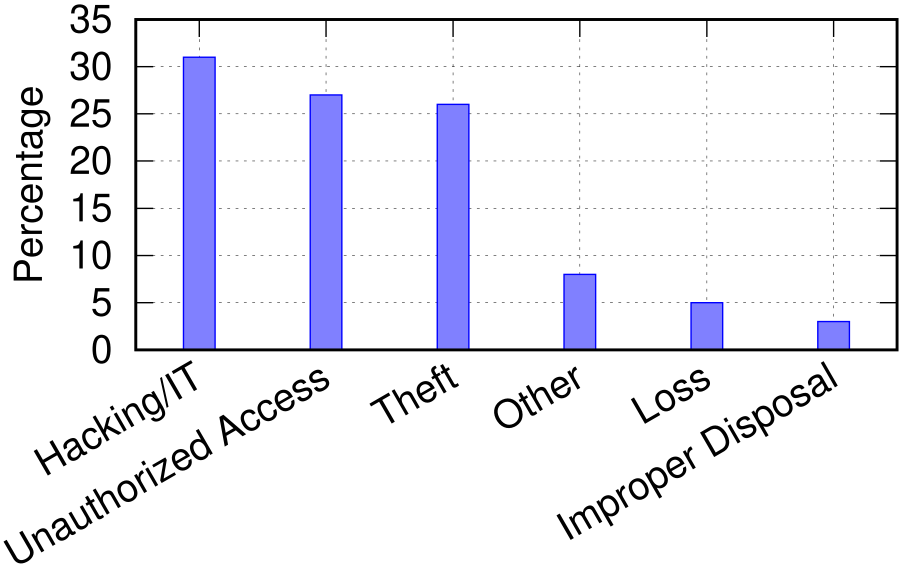


Figure 11: The distribution of breach types within the healthcare sector.

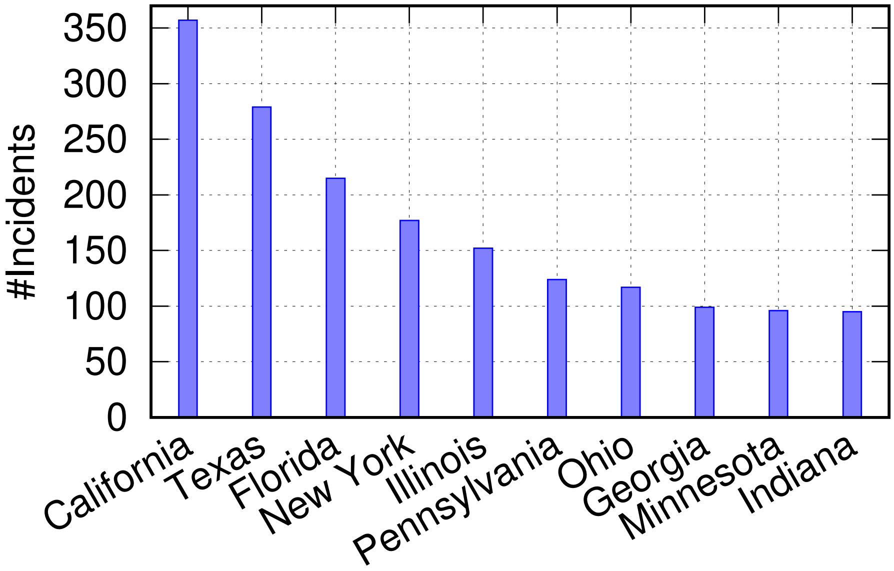


Figure 12: The distribution of data breach incidents across different states.

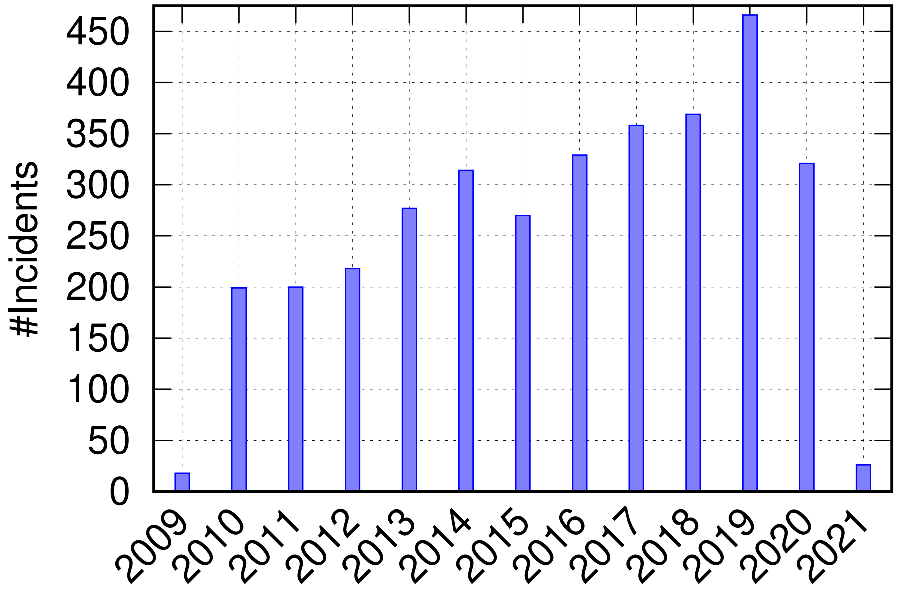


Figure 13: The yearly distribution of Data Breach Incidents.

security controls by organizations in the healthcare industry. As shown in Figure 13, there was a massive increase in the number of incidents in 2019, as it was the year with the highest number of breaches in the whole dataset.

Covered Entity. We analyzed the distribution of incidents by organization type. According to the OCR dataset, there are three main targeted entities. First, healthcare entities that provide healthcare services and engage in professional review activity through a formal peer review process for the purpose of furthering quality health care, a committee of that entity, a professional society, a committee or agent thereof, including those at the national, state, or local level, physicians, dentists, or other health care practitioners that engage in professional review activity through a formal peer review process to further quality health care [56]. Second, a business associate is a person or entity that performs certain functions or activities that involve the use or disclosure of protected health information on behalf of or provides services to a covered entity [31]. Third, health plan, which constitutes individual or group health plans that provide or pay the cost of medical care [41]. Following this analysis, we observed that healthcare entities are most targeted during the incidents, having 2,450 incidents which represent 73% of the total incidents; business associates and healthcare plans came in second and third, comprising 451 and 439 incidents and representing 14% and 13%, respectively. Figure 14 depicts the results of this analysis.

Business Associates. We further analyzed the existence of incidents when a business associate is present or not. According to HIPAA, any covered entities and business associates enter into a contract to ensure the safety of protected healthcare information. A business associate may use or disclose protected health information only as permitted or required by its business associate

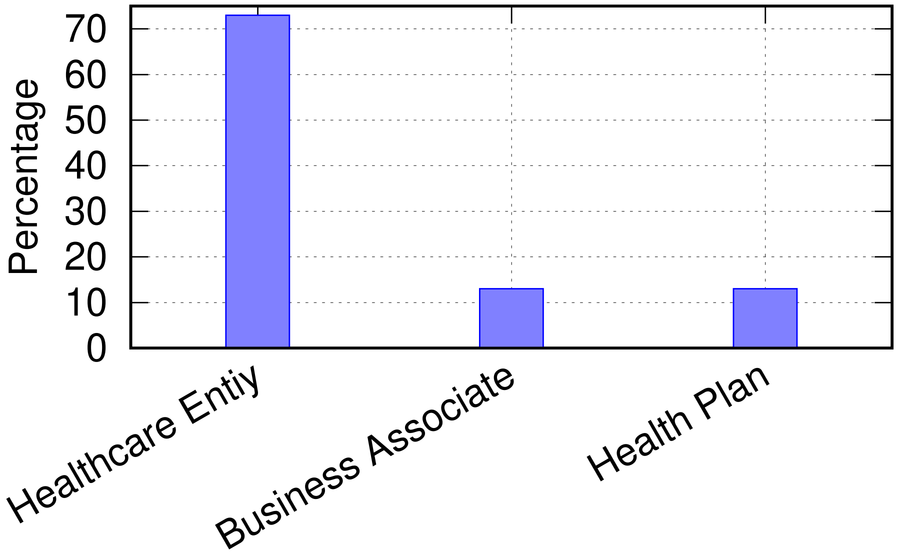


Figure 14: The distribution of covered entities.

contract or as required by law [66]. Our analysis revealed that 2,532 incidents had no business associates included, representing 76%, while only 819 incidents had a business associate, representing 24% of the incidents as shown in Figure 15.

Location of Breached Information. When a data breach occurs, private and confidential patient information gets disclosed due to either unauthorized access or human error. The healthcare system keeps records of valuable information and medical records containing sensitive personally identifiable information (PII) such as address history, financial information, social security numbers, and patient medical treatment records. Hackers often target this sensitive information due to its outstanding value. Hackers can easily use that data to set up a line of credit or take out a loan under patients' names. Unfortunately, healthcare organizations often lack the stringent security measures (e.g., encryption, robust anti-virus software, multi-factor authentication, etc.) required to secure medical records. To this end, we analyzed the most targeted information to gain insight into the type of medical and personal data prioritized by hackers in healthcare data breaches. We observed that paper/films are the most breached information comprising 662 of the overall incidents, representing 20%. Closely, the network server came in second, comprising 643 incidents, accounting for 19%. The other category came in third, comprising 641 incidents, representing 19% as well. The remaining attributes and results of this analysis are presented in Figure 23.

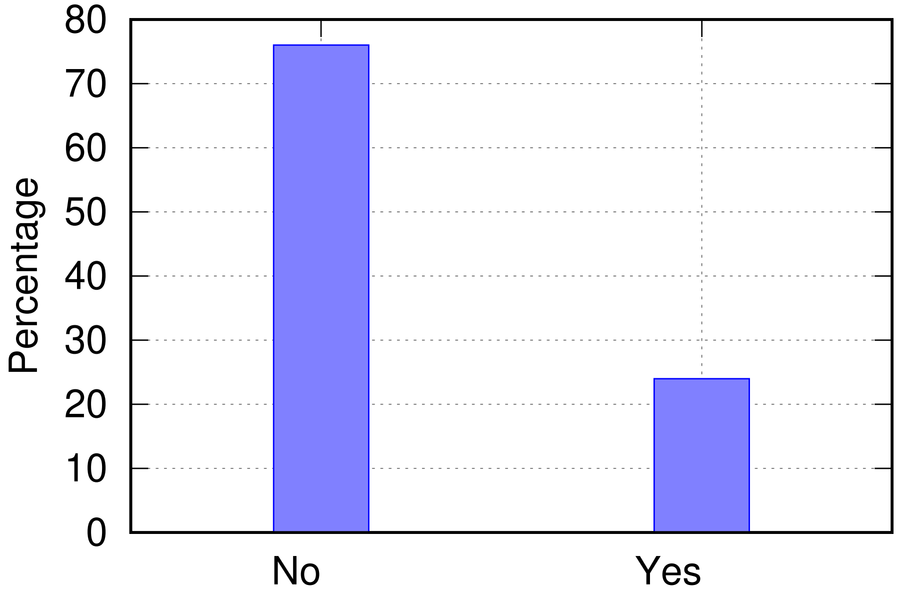


Figure 15: The distribution of business associates.

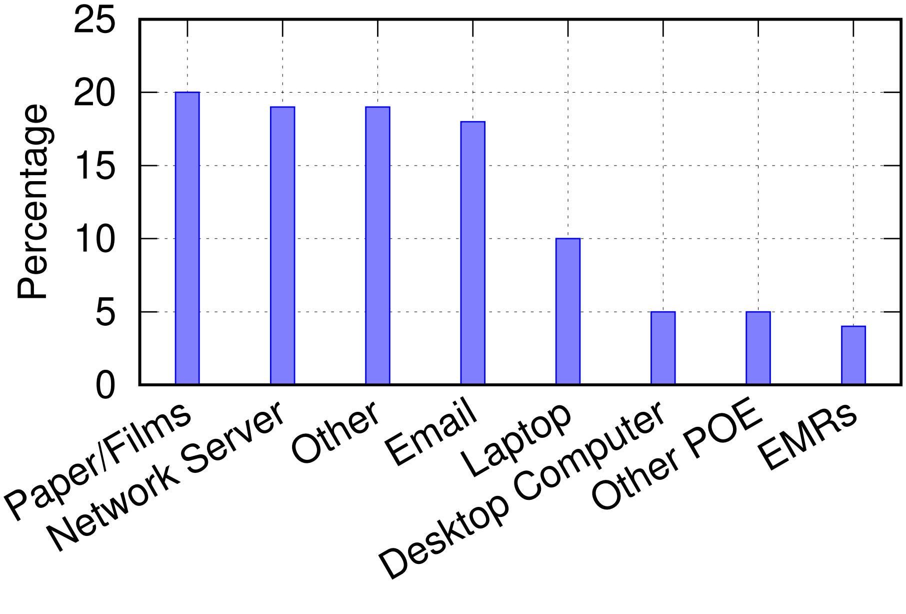


Figure 16: The distribution of breached information.

4.7 Summary and Work to be Completed

Our study revealed that the number of adults and the state population highly influence the exposure to data breach incidents, with California, Florida, and Texas being the lead targets. We show that the media group was the most breached asset, followed by the Server and User group. Our timeline discovery revealed that most of the incidents, approximately 52%, were discovered within months, while 15% of the incidents took years to be discovered. Employees discovered the majority of the incidents for internal incidents. In the future, it would be interesting to conduct research harnessing the power of machine learning to enable information sharing on data breaches.

Our future work intends to further investigate data breaches by exploring the relationship between the size of healthcare entities and their contribution to data breaches. This analysis aims to analyze how the size of healthcare organizations influences the likelihood and severity of data breaches. Additionally, we will study the timeline discovery of reported incidents. By examining the timeline of data breach incidents, our goal is to uncover patterns and trends in the detection and reporting of these security breaches within healthcare organizations

5 *Understanding the Security and Performance of the Web Presence of Hospitals*: A Measurement Study

5.1 Summary of Completed Work

This work analyzes the online presence of hospitals and investigates their websites in light of the increasing trend of attacks targeting hospital networks. We categorize hospitals into government, non-profit, and proprietary sectors, providing a comparative study that examines various structural and security features. We investigate the SSL certificate validity and related issues observed on hospitals' websites, as well as identify any malicious associated behaviors. Furthermore, we collect attributes and utilize them as features. The study aims to highlight the most important indicators of websites associated with data breach incidents and improve understanding of their security posture.

5.2 Introduction

Electronic Health Records (EHR) are longitudinal electronic patient health information records, which include patient demographics, progress notes, health problems, medications, vital signs, medical history, immunizations, laboratory data, etc. [45]. The adoption of EHRs has led to improved accessibility of healthcare information for patients and providers, resulting in higher quality patient care and more efficient coordination between hospitals. However, despite these benefits, the transformation to EHRs has also raised privacy and security concerns, particularly when EHR data is retrievable through website systems. EHRs centralize sensitive patient data, which can make them a prime target for cybercriminals seeking to steal or exploit this information. Additionally, EHRs can be accessed and shared across multiple healthcare providers, increasing the risk of data breaches and unauthorized access. Moreover, the implementation of EHR systems can introduce new vulnerabilities that may be exploited by cybercriminals.

For instance, vulnerabilities and software exploits in the healthcare domain have become a central focus of targeted cyber attacks [35], which can result in devastating data breaches. Given the sensitivity of private healthcare data, the unauthorized and illegitimate disclosure of this information can have catastrophic consequences [5, 60].

Understanding the effect of healthcare data breaches is essential, and efforts in the literature classified those breaches into internal and external breaches. Internal breaches are commonly caused by human errors, particularly among healthcare employees. In contrast, external breaches, which are the more critical type, are caused by an unauthorized third party involved in the theft of private health records through hacking of the web-based user and healthcare provider-facing systems [76]. Cybercriminals typically commit these incidents, making their effect an open question, with no accurate assessment of their cost. For instance, adversaries involved in external breaches may aim to steal sensitive records and demand a ransom or sell those records for hundreds of dollars per single patient on the dark web [62, 73].

Given the importance of understanding data breaches in healthcare and the role of web technologies in enabling a significant part of the attack surface, this study is dedicated to analyzing the commonalities and differences among three types of hospitals: government public hospitals, non-profit hospitals, and proprietary hospitals. Namely, we analyze the websites and patients' portals for security configurations and common privacy practices. We note that compromising patients' portals allows the attacker to obtain sensitive information regarding the patient's records, including diagnoses, treatment records, hospital visits, and future appointments, alongside personal information. To the best of our knowledge, this work is the first in this direction, associating actual hospital potential exploitations and data breaches with website security and privacy configurations.

Our analysis is based on a total of 4,774 hospital websites grouped into three major hospital categories: government public hospitals, non-profit hospitals, and proprietary (private) hospitals. For our measurement assessments, we conduct both domain-level and content-level analyses to understand the similarities and differences among website attributes.

Our analysis is multi-faceted and covers a range of features by examining and comparing the website's domain SSL certificates, creation date, HTTP requests, page size, content type, average load time, and malicious activity association. The features explored in this analysis are particularly lightweight and do not require deep analyses of contents but rather focus on meta-attributes, making our analysis techniques more generalizable to large-scale measurements. We further investigated the security attributes of these websites by exploring their association with malicious behaviors, including an assessment of the domain-based and content-based malicious behaviors of those websites and associated trends and characteristics.

To understand the implications of those characteristics, we further study their correlation with a manually vetted dataset of recently disclosed data breaches provided by the U.S. Department of Health and Human Services, the Office for Civil Rights (OCR) [25]. Leveraging information regarding the websites and associated breaches, we extracted the commonalities among hospitals' websites targeted with those data breach attacks towards their modeling and characterization. We believe that this work is the first step towards understanding website attributes that may lead to breaches and enable future research on vulnerability prediction and detection.

Research Questions. We aim to answer an overarching single question: **Is there any difference between the different categories of hospital websites with respect to the studied features across content, performance, and security?** We break the question down into the following quantifiable questions.

- **RQ1.** How different are different hospitals with their use of domain, content, and transport layer features? We answer this question by comparatively exploring the domain-level features (section 5.4.1), including the domain name registrar, top-level domain distribution, domain creation distribution, and content-level features, including the content type and HTTP request features (section 5.4.2).

- **RQ2.** What are the main security characteristics of hospital websites, and how do they differ across types? We answer this research question by exploring the SSL certificate features and properties (section 5.4.3), maliciousness characteristics against various engines (section 5.4.4), and data breaches association (section 5.5).

Contributions and Findings. Given the lack of any systematic work on understanding the characteristics of hospitals' presence on the web and their associated security and performance attributes, this study sets out to explore these hospitals' web presence across a range of attributes. Moreover, through a comparative analysis, this work uncovers the differences and similarities between the Government, Non-profit, and Proprietary hospitals in the United States. Our analysis is conducted across three dimensions: security, contents, and domains. To this end, our contributions are as follows:

1. **Domain-level Analysis (§5.4.1).** Domain names are the gateway to websites, and they are essential to understanding various coarse-grained and easy-to-obtain features of those domains and entities behind them. To this end, we conduct a domain name registrar and top-level domain analysis to uncover websites/hospitals' characteristics and to contrast them. We uncover the affinities in the choice between websites and registrars, top-level domain choice, and domain creation dates. Among other interesting findings, we observe that the number of websites for government and non-profit hospitals has been declining in recent years, hinting at the aggressive proprietary healthcare system.

2. **Content-level Analysis (§5.4.2).** We examined the contents of hospitals' websites for a deeper look into their utilized content types, size, and employed security features. Through this analysis, we found the gap in employing different content types, such as images and scripts, in those websites, which affects the various performance metrics, including loading times. More interestingly, and rather surprisingly, we found that 6% of proprietary hospitals use the Domain Name System Security Extension (DNSSEC), in comparison to less than 1% of the government and non-profit hospitals.

3. **SSL Certificate-level Analysis (§5.4.3).** Certificates are essential for website authentication and to facilitate web content encryption at the transport layer, providing a secure application medium. We investigate the HTTPS protocol configurations, associated SSL features, and the SSL certificate validity of hospitals' websites. We categorize websites based on certificate authority affinity, utilized algorithms, and certificate validity. Among other interesting findings, our investigation uncovers that more than 25.25% of hospital websites are still using

the insecure HTTP protocol. Further, among websites that utilize HTTPS, up to 23% of the SSL certificates are invalid.

4. **Malicious Activities Analysis (§5.4.4).** Because of their complex nature, unintended weaknesses might emerge due to the agglomeration of third-party code and the utilization of various shared pieces of infrastructure in hospital websites. To understand this dimension, we utilize various scanning tools to explore those websites' malicious activities at the domain and content levels. Among other interesting findings, we uncover that a large portion of websites contains malicious content and are associated with malicious behaviors.

5. **Data Breaches Analysis (§5.5).** Data breaches are inevitable. But what (in the correlation sense) makes a website prone to data beach? We explore this question by correlating and associating the hospitals to recently reported data breach incidents and uncover that non-profit hospitals are more likely to be involved in data breach incidents. We demonstrate the most important attributes contributing to data breach incidents, including hosting malicious codes, a large number of images, etc.

5.3 Dataset, Pipeline and Research Questions

For this study, we utilized an authentic dataset of U.S. hospitals obtained from the Homeland Infrastructure Foundation-Level Data (HIFLD) [51]. The dataset contains hospitals distributed among the 50 U.S. states, Washington D.C., and U.S. territories of Puerto Rico, Guam, American Samoa, Northern Mariana Islands, Palau, and the Virgin Islands.

We categorized the hospitals in the dataset into three categories: government, non-profit, and proprietary. The government hospitals include federal, district, local, and state hospitals. Non-profit hospitals are those operated using charities according to the Internal Revenue Service (IRS) [28]. The proprietary hospitals are those owned and operated for profit by individuals, partnerships, or, in most cases, corporations [64]. Overall, we had 1,034 governmental hospitals, 2,187 non-profit hospitals, and 1,550 proprietary hospitals.

Websites Preprocessing and Crawling. Our study involved the use of website crawling to systematically gather information from websites and incorporate it for our further analysis. Figure 17 illustrates the overall process we followed, starting with website enumeration to identify the websites we needed to crawl. Next, we conducted website preprocessing, which involved removing irrelevant websites (i.e., websites with irrelevant contents to the scope of the study) and non-functioning websites. Finally, we performed feature extraction to extract attributes from the websites that we used for our further analysis.

To conduct our analysis, we introduced a data augmentation step. Upon crawling the websites associated with each hospital, we enriched the collected data with additional attributes, including

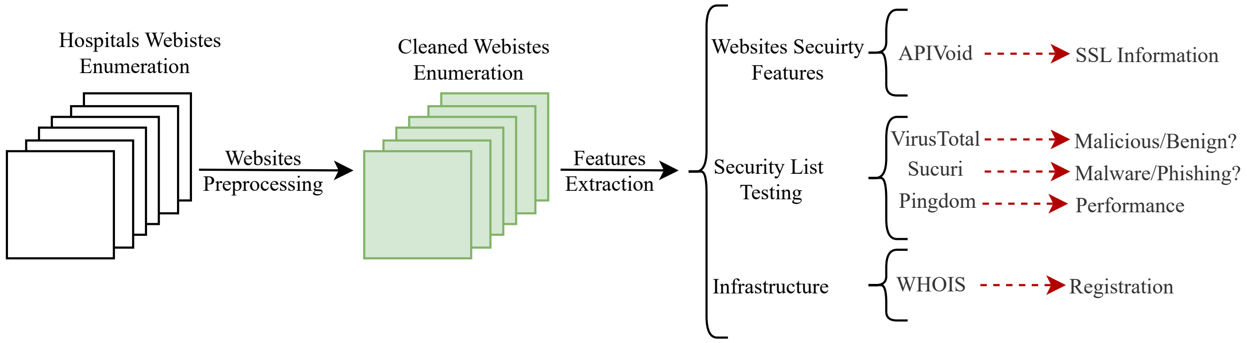


Figure 17: Our pipeline with the steps taken in website crawling and data augmentation against various dimensions: SSL, maliciousness, vulnerability, performance, and domain attributes.

the following.

- **SSL Attributes.** To extract SSL certificate information such as mismatched domains, SSL expiration dates, and certificate validity, we utilized APIVoid [8], a framework that offers cyber threat analysis and detection capabilities.
- **Maliciousness Attributes.** To analyze the maliciousness of hospital online content, we employed VirusTotal API [70], an online service that aggregates results from over 70 scanning engines.
- **Vulnerability Attributes.** To examine each website for vulnerabilities and identify any malicious code, we utilized Sucuri [65], a service that tests websites against multiple known malware and blacklisting lists.
- **Performance Attributes.** To evaluate website performance and availability, we utilized Pingdom [54], a global monitoring software for websites.
- **Domain Attributes.** To determine ownership and DNSSEC information for each website, we utilized WHOIS [74], an Internet resource ownership database, and queried each website's creation date.

Overall, the steps of websites crawling and data augmentation allowed us to extract two types of information: ① website content data such as images, fonts, HTML, CSS, scripts, XHR, and redirects, and ② performance metrics such as page size, load time, and the number of requests.

5.4 Websites Analysis

To understand the online presence and structural differences between hospitals of different categories, we conducted a range of analyses of their websites: domain-, content-, SSL certificate-, and malicious activities-based analyses. In the following, we review the findings from those analyses.

Table 3: The hospitals' website URLs correspond to domain registrar organizations. Notice that *Network Solutions* and *GoDaddy* are the most prominent in the list, with up to 67.65% associated URLs.

Domain Registrar	Government		Non-profit		Proprietary	
	#	%	#	%	#	%
Network Solutions_LLC	330	35.22	956	44.90	345	22.91
GoDaddy.com_LLC	314	33.51	621	29.17	527	34.99
MarkMonitor_Inc	1	0.11	19	0.89	156	10.36
eNom_LLC	28	2.99	51	2.40	159	10.56
Register.com_Inc	24	2.56	43	2.02	14	0.93
NAMECHEAP_Inc	13	1.39	37	1.74	24	1.59
CSC CORPORATE_Inc	2	0.21	51	2.40	19	1.26
Tucows_Inc	37	3.95	26	1.22	10	0.66
Other	188	20.06	325	15.27	252	16.73

Table 4: Top-Level Domain comparison between the Government, Non-profit, and Proprietary hospitals.

Type	.org	.com	.gov	.net	.mil	.edu	.us
Government	48.45%	36.56%	4.35%	4.35%	2.90%	2.42%	0.48%
Non-profit	67.49%	28.81%	0.05%	1.92%	0.00%	1.37%	0.27%
Proprietary	7.81%	87.35%	0.00%	3.48%	0.00%	0.13%	0.45%

5.4.1 Domain-level Analyses

The domain name is a crucial asset for any organization, serving as a key element in their branding efforts and providing them with a strong online presence and Search Engine Optimization (SEO) benefits. Therefore, in order to kickstart our website analysis, we begin by examining the domain name details, including the domain name registrar, the top-level domain, and the domain creation date.

Domain Name Registrar. The domain name registrar is an organization that manages the reservation of Internet domain names, as well as the assignment of IP addresses for those domain names [17], and certain registrars tend to be more lax with their security provisions and policies [19, 20, 67]. Analyzing the domain name registrar is crucial in evaluating a website's overall security and reliability. This is because the registrar provides important information about the organization's online presence and security measures. The level of security provisions and policies of the registrar can vary, which can impact the website's security and trustworthiness. Examining the domain name registrar can provide valuable insights into the organization's security approach and help assess potential risks associated with the domain name. In addition, understanding the registrar can shed light on the organization's online strategy and web hosting arrangements, which can further inform the analysis of the website's structure and performance. Table 3 shows the break-down of domain registrar organizations by hospital type. Notably, *Network Solutions* and *GoDaddy* are the most prominent registrars, accounting for up to 67.65% of the domains. Additionally, we observe that although *Mark Monitor* and *eNom LLC* are relatively absent from government and non-profit websites, they contribute to 20.92% of proprietary hospital websites.

Top Level Domain. The Top-Level Domain (TLD) is the "extension" of a domain name. Besides branding, TLD plays an essential role in the Domain Name System (DNS) lookup and helps classify and communicate the purpose of domain names. Examples of TLDs include ".com," ".org," ".net," and ".edu." The TLD provides information about the website's purpose, organization type, or geographic location. For instance, ".com" is commonly used for commercial websites, ".org" for non-profit organizations, ".edu" for educational institutions, and ".gov" for government websites. Understanding the TLD can provide insights into the website's intended audience and the type of content or services it provides. In addition, analyzing the TLD can help identify potential security risks associated with the website. For example, certain country-specific TLDs are known to be associated with malicious activities, and websites using such TLDs may be more likely to pose security threats. Moreover, some websites may use TLDs that are misspelled or similar to well-known TLDs in an attempt to deceive visitors and carry out fraudulent activities. By analyzing the TLD, we can identify such risks and take appropriate measures to mitigate them. Therefore, analyzing the TLD is a crucial step in evaluating a website's overall security and reliability. The Internet Corporation for Assigned Names and Numbers (ICANN) is responsible for the authority over all TLDs on the Internet and delegates these TLDs' responsibility to various organizations [18] Table 4 shows the TLD comparison between the hospital categories in our dataset. We observe that *".org"* is the most dominant TLD for the government (48.45%) and non-profit (67.49%), while it is relatively absent in the proprietary hospitals (only 7.81%). On the other hand, *".com"* is dominant for proprietary hospitals (87.35%) compared to (36.56% and 28.81%) in government and non-profit hospitals, respectively. We also notice that 92.15% of the hospitals' websites, in the aggregate, have *".com"* or *".org"*. Despite our common beliefs, we surprisingly uncover that only 4.35% of the government websites use the *".gov"* TLD.

Domain Creation. The domain creation date refers to the date on which a specific domain name was initially registered with a domain name registrar. It is a crucial piece of information because it can provide insights into the website's history and online presence. A website that has been registered for a longer duration may be more established and have a more significant online presence than a newer website. Analyzing the domain creation date can be useful in identifying potentially fraudulent or malicious websites. For instance, a website that has been recently registered may be more likely to be a part of a phishing scam or a fraudulent scheme. Therefore, analyzing the domain creation date can be a valuable step in evaluating the website's overall credibility and potential security risks. Figure 18 shows the domain creation date of hospitals in different categories. As shown, both government and non-profit hospitals' websites emerged in a similar period (1995 –

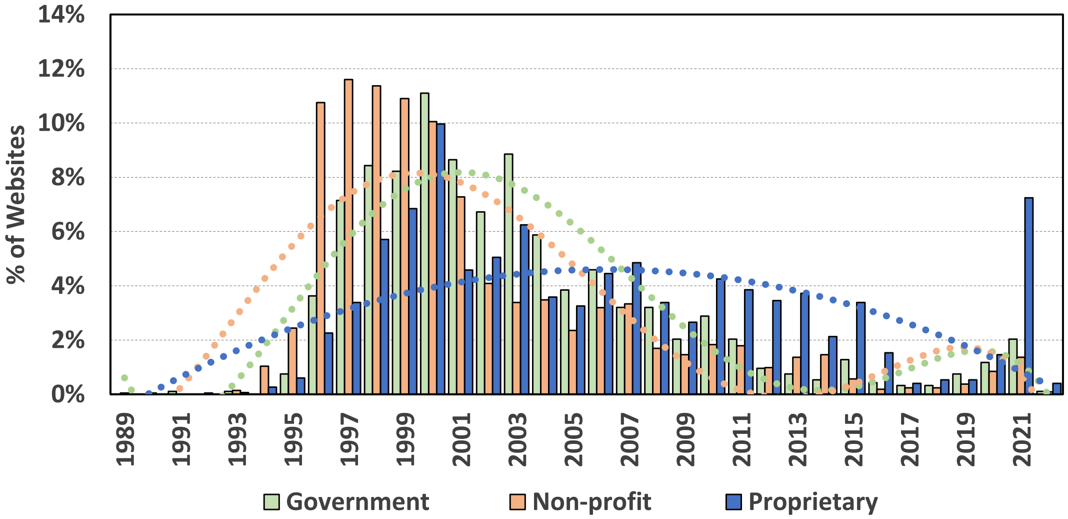


Figure 18: The domain creation date temporal analysis between the three hospital categories. Dot lines are moving averages.

Table 5: Content-type comparison between the Government, Non-profit, and Proprietary hospitals.

Category	CSS	Font	HTML	Image	Redirect	Script	XHR
Government	6.05%	8.18%	2.71%	40.55%	17.37%	20.34%	4.79%
Non-profit	5.30%	7.80%	2.86%	38.01%	16.11%	27.01%	2.91%
Proprietary	6.63%	9.61%	3.53%	28.05%	13.76%	33.12%	5.29%

2009), with a declining trend after 2009. However, the emergence of proprietary hospital websites is steady, with a rapid increase in their numbers in 2021.

> **Takeaway:** (**RQ1.**) While the number of websites for government and non-profit hospitals has been declining in recent years, proprietary hospitals have been growing significantly, particularly in 2021. Moreover, despite being government-supported, most government hospitals do not have *".org"* top-level domain.

5.4.2 Content-level Analyses

To analyze the content differences between the different categories of hospitals, we crawled the hospitals' websites using Pingdom [54], obtaining the HTTP request information and all associated files; scripts, images, HTML, and CSS files.

Content Type. On the structural level, Table 5 shows the distribution of the file type among the three hospital categories. XHR is an API used as an object to interact with servers and exchange data between servers and web browsers. Containing *"XHR"* is prominent among the government and proprietary hospital websites, with 10.08% combined. Overall, the file type distribution is similar for all categories, except for *"Image"* and *"Script"*. The *"Script"* content, which is defined

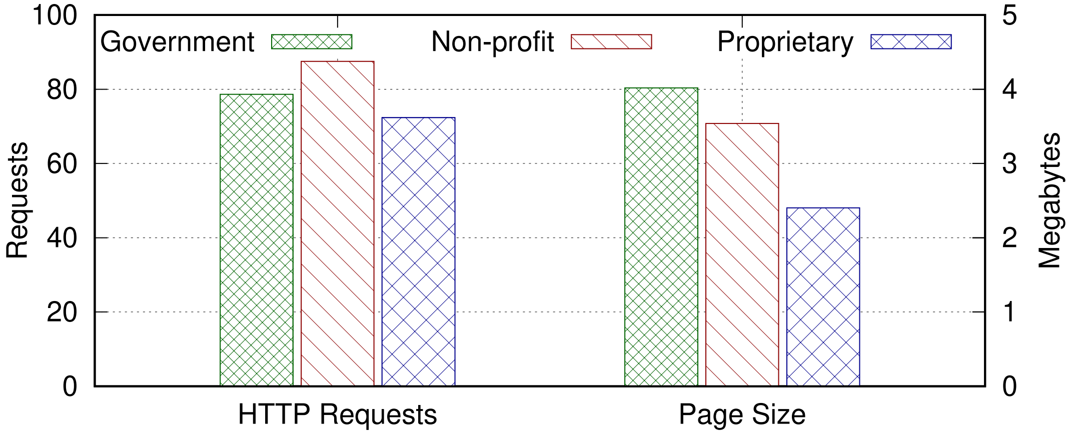


Figure 19: Request and response size comparison.

as a computer program for adding dynamic capabilities to a website, is used most among the proprietary hospitals with 33.12%. The *"Redirect"* content, on the other hand, which is a website feature that sends a user from the current URL to another server, is applied more in the government (17.37%) and non-profit (16.11%) and the least in the proprietary category (13.76%).

HTTP Request. The HTTP (Hypertext Transfer Protocol) request is a message that is sent from a client (such as a web browser) to a server, requesting a particular resource or action. The request typically includes a URL (Uniform Resource Locator) that specifies the resource or action being requested, along with any additional information needed by the server to fulfill the request, such as headers and cookies.

By Analyzing HTTP requests, businesses and organizations can gain insights into the performance of their websites, including, for example, the speed of page load times, the number of requests per page, and the size of files being requested. This information can help identify areas where website performance can be further optimized.

Figure 19 shows the average HTTP requests per website across the three different categories of analyzed hospitals. We found that most websites generated 65 to 90 HTTP requests per visit, with the non-profit hospitals being the highest. Despite having relatively similar HTTP requests, the proprietary hospitals' average page size was 45% less than the government hospitals. Upon further analysis, we found that the proprietary hospitals' websites contain the least percentage of images in contrast to the government and non-profit hospitals, which help explain this trend.

> **Takeaway:** (**RQ1.**) Structurally, the content type distribution of the hospitals' categories are similar, except for *"Image"* and *"Script"* content types. Although the HTTP requests were relatively similar among hospitals, the average page size of the proprietary hospitals was 45% smaller than that of the government hospital.

Table 6: The corresponding certificate issuer organizations for the hospitals' websites. Notice that *Let's Encrypt* is the most prominent certificate issuer organization, with up to 25.21% associated URLs.

Issuer Organization	Paid	Government		Non-profit		proprietary	
		#	%	#	%	#	%
Let's Encrypt	✗	321	31.04	559	25.56	295	19.03
GoDaddy.com_Inc.	✓	131	12.66	175	8.00	132	8.51
Cloudflare_Inc.	✗	25	2.41	108	4.93	179	11.54
Sectigo Limited	✓	67	6.47	133	6.08	38	2.45
cPanel_Inc.	✗	47	4.54	72	3.29	111	7.16
DigiCert Inc.	✓	35	3.38	154	7.04	14	0.90
Trustwave Holdings, Inc.	✓	1	0.09	9	0.41	148	9.54
Entrust L1K	✓	9	0.87	86	3.93	43	2.77
Other	-	148	14.31	385	17.60	149	9.61
No SSL Certificate Found	-	250	24.17	506	23.13	441	28.45

5.4.3 HTTPS and SSL Certificate Analysis

The HTTP protocol is responsible for transferring website content from the web server to the endpoint browser. However, this protocol is insecure, exposing content to unauthorized access. Therefore, most websites have moved to using HTTPS, a secure version of HTTP, on top of the Secure Sockets Layer (SSL), which, among other functions, implements an encryption mechanism to protect the transferred content between web servers and endpoint browsers. Healthcare websites often require users to enter sensitive personal information such as health-related data, insurance information, and medical history. HTTPS can help protect this information from being intercepted by attackers, thereby ensuring that patient data remains confidential.

To this end, we next look into SSL-related configurations: certificate authority, signature algorithm, and certificate validation. Among the studied hospitals, we noticed that 25.25% of the websites are still using HTTP, in contrast to only around 20% in general web [71]. While there is a 5% of difference in the number of websites, that number is alarmingly high given the type of data associated with hospitals (i.e., at least one out of four hospitals uses an insecure protocol).

Certificate Authority. A certificate authority (CA) is an organization that validates the identities of entities, including websites, email addresses, etc., by binding entities to cryptographic keys through the issuance of electronic documents. Investigating the certificate authority organization (i.e., the issuer of the certificate), Table 6 shows that the majority of hospitals are using the free *Let's Encrypt* services [34] (i.e., free SSL certificates)., with up to 31.04% for the governmental hospital group. We also notice that hospitals' websites widely use free SSL certificates. Surprisingly, we did not find SSL certificates in 24.17% of government, 23.13% of non-profit, and 28.45% of proprietary websites.

Table 7: SSL signature algorithms' comparison.

	Government		Non-profit		proprietary	
Algorithms	#	%	#	%	#	%
SHA256 with RSA	751	95.00	1,535	91.31	891	80.34
SHA256 with ECDSA	26	3.30	108	6.42	179	16.14
SHA384 with ECDSA	10	1.27	35	2.08	37	3.34
SHA1 with RSA	1	0.13	2	0.12	1	0.09
SHA384 with RSA	1	0.13	-	-	1	0.09
SHA512 with RSA	-	-	1	0.06	-	-

Algorithms. Table 7 shows the SSL signature algorithms used by the government, proprietary, and non-profit websites. As shown, *SHA256 with RSA* is the most used scheme with 95.00% for government, 91.31% for non-profit, and 80.34% for proprietary hospital websites, respectively. This is mainly because hospitals intend to use traditional go-to algorithms adopted by service providers. On the other hand, we notice that fewer hospitals website use *SHA256 with ECDSA* (Elliptic Curve Digital Signature Algorithm) algorithm that uses shorter keys for the same security level as in RSA with larger keys [3]. With 3.30% for government, 6.42% for non-profit, and 16.14% for proprietary hospital websites, respectively. We note ECDSA is a newer and more efficient algorithm and is mainly used in newer websites [3]. ECDSA is, however, more vulnerable to attacks than the older RSA under post-quantum adversaries, according to recent studies [58].

Certificate Validity. We further investigated the SSL certificate validity and potential issues. In the following, we discuss issues related to SSL certificate failures (see Figure 20).

SSL Mismatched Domain. A mismatched domain might be an indication of website impersonation or inconsistent website migration, and both highlight a lack of rigorous security practices. We found that 18.45% of the proprietary hospitals had SSL certificates with mismatched domains, versus 14.05% of the government hospitals and 17.67% of the non-profit hospitals. Even varying, all hospitals' websites had concerning ratios of mismatched domains.

SSL Expired. Our analysis uncovered that about 3.97% of non-profit hospitals have expired certificates, compared to 2.20% and 2.09% for government and proprietary hospitals. Similar to our previous analysis of out-of-date websites, this may lead to potential user information and data privacy risks.

SSL Invalid. The invalidity of SSL means that some fields in the certificate are incorrect. Surprisingly, all hospital categories had an alarming percentage of invalid SSL certificates, with 15.94% for government, 21.70% for non-profit, and 23.72% for proprietary hospital websites.

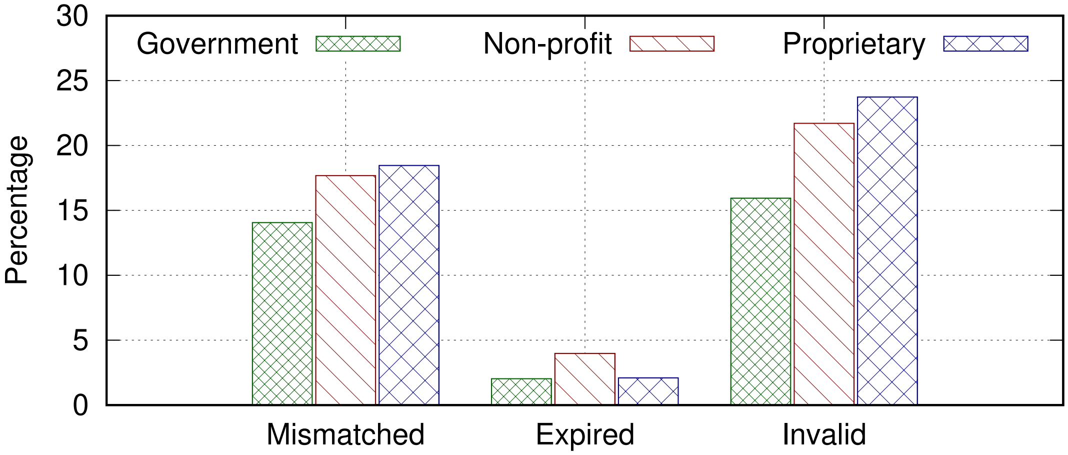


Figure 20: The SSL validity comparison of Government, Non-profit, and proprietary hospital websites.

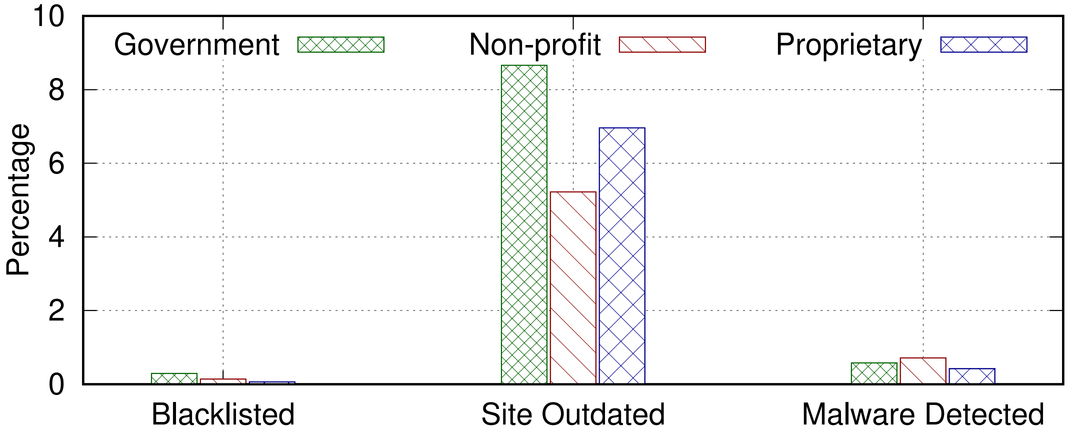


Figure 21: Comparing the maliciousness of Government, Non-profit, and proprietary hospitals' websites.

Takeaway: (**RQ2.**) More than 25% of hospitals' websites are using the plain HTTP protocol, which is alarmingly higher than ≈20% in the general websites [71]. Among websites that used HTTPS, 88.77% of them used *SHA256 with RSA*. Among the ≈75% hospitals with an SSL certificate, we found that 20.45% of the SSL certificates were invalid while 16.72% had a mismatched domain name, primarily in proprietary hospitals in both cases.

5.4.4 Malicious Activities Analysis

In addition to the structural differences and SSL certificate analyses, we study the malicious activities associated with the hospitals' websites. Malicious activities considered in this work include providing malicious or phishing content or the association of website resources with malicious attacks.

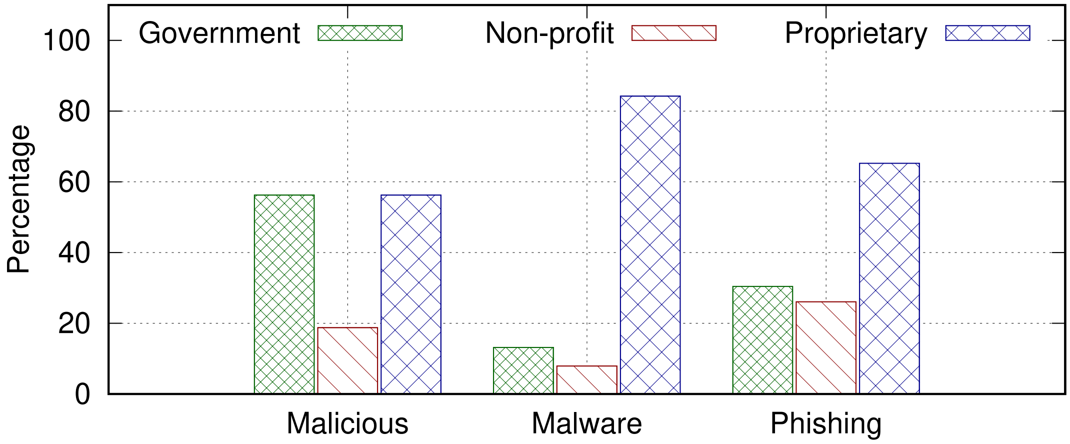


Figure 22: The potential maliciousness of Government, Non-profit, and proprietary hospitals.

Domain-based Malicious Activities. We leveraged Sucuri [65] to explore domain-based malicious activities. Figure 21 shows that although only a small portion of hospitals' URLs are blacklisted or labeled as malware, about 8.66% of government, 5.21% of non-profit, and 6.96% of proprietary hospital websites are outdated, which raises concerns of data leakage.

Content-based Malicious Activities. Next, we analyzed the website content using VirusTotal API [70]. Figure 22 shows that, among the three hospital categories, 84.21% of proprietary hospitals contained malware, compared to only 13.15% and 7.89% in the government and non-profit hospitals, respectively. Moreover, we observed that 65.21% of proprietary hospital websites are suspected of having phishing-like behaviors. We note that the percentage for the government (30.43%) and non-profit (26.08%) hospitals are significantly smaller than that of the proprietary hospitals but still noticeably high. Besides, we observed that 56.25% of government and proprietary hospitals and 18.75% of non-profit hospitals are associated with malicious activities.

> **Takeaway:** (**RQ2.**) Most hospitals have maliciousness features (domain or content), and many are vulnerable to data leakage due to a lack of maintenance (i.e., outdated websites). Among the compiled hospital websites, 8.66% of the websites are outdated. In addition, a concerning portion of websites contains malicious content and is associated with phishing and malicious behaviors.

5.5 Data Breaches Analysis

Analyzing the data breaches helps understand the correlation between web presence security and incidents. According to the Health Insurance Portability and Accountability Act (HIPPA), a data breach can be defined as an impermissible use or disclosure under the Privacy Rule that compromises the security or privacy of the protected health information [24]. In the healthcare domain, data breaches are devastating, as they cause damage to patients and healthcare organizations alike. Recent works have shown that healthcare is the most targeted industry by cyber criminals due to financial gain as attackers intend to sell patients' records on the dark web. To investigate historical data breach incidents in hospitals, we obtained the healthcare data breaches dataset from the U.S. Department of Health and Human Services, Office for Civil Rights (OCR). The OCR portal lists all data breaches of unsecured health information affecting 500 or more patients. The OCR portal categorizes data breaches into two categories; (i) incidents reported within the last 24 months and currently under investigation and (ii) the achieved breaches, which comprise the resolved breach reports older than 24 months. We note that it is challenging to associate the hospital names with the entities named in the data breaches, as they are not consistently organized (e.g., mixture or truncation). To resolve the issue, we started by using the hospitals' names as anchors and then leveraged Natural Language Toolkit (NLTK) [49] for punctuation removal, case normalization, stopwords removal, and lemmatization & stemming of the hospitals' names. A similar process was followed for the entity name among the data breaches dataset. Lastly, any hospital name and entity with two common words are filtered for manual analysis and vetting. Overall, we manually inspected 1,253 incidents, resulting in 414 accurate labeling of data breaches.

Hereafter, we analyze the data breaches, providing insights into the common online attributes of the breached hospitals.

Associated Hospitals & Individuals. Among the 414 data breach incidents in our dataset, 49 were government hospital-related, 156 were non-profit hospital-related, and 34 were proprietary hospital-related. It is worth mentioning that a hospital may be involved in several incidents. Our analysis indicates that the average number of affected individuals is 58,750 overall, including 60,458 for government, 64,977 for non-profit, and 50,815 for proprietary hospitals. Remarkably, the proprietary hospitals are involved in the least number of incidents and affected individuals.

Data Breach Surface. As shown in Figure 23, *"paper/films"* are the most commonly targeted for government (32%) and proprietary (21%) hospitals, despite only 6% for non-profit hospitals. Then, *"emails"* are mostly targeted in non-profit hospitals (34%), despite not being heavily targeted in government and proprietary hospitals (1%). Overall, *"network server"* is the second most common target after *"paper/films"*, inferring the importance of hospitals' online security.

Data Breach Type. As shown in Figure 24, we found the majority of incidents are *"hacking/IT"* representing 45.75% in the government hospitals, followed by proprietary and non-profit hospitals representing 30.22% and 22.58%, respectively. Further, we observed that *"unauthorized access/disclosure"* is the most common data breach type within the non-profit hospitals representing 50%, while 37.91% for proprietary and 32.98% for government hospitals.

Data Breach Online Presence Attribution. To understand the relationship between online presence, security properties, and data breach incidents, we used a gradient boosting model with non-

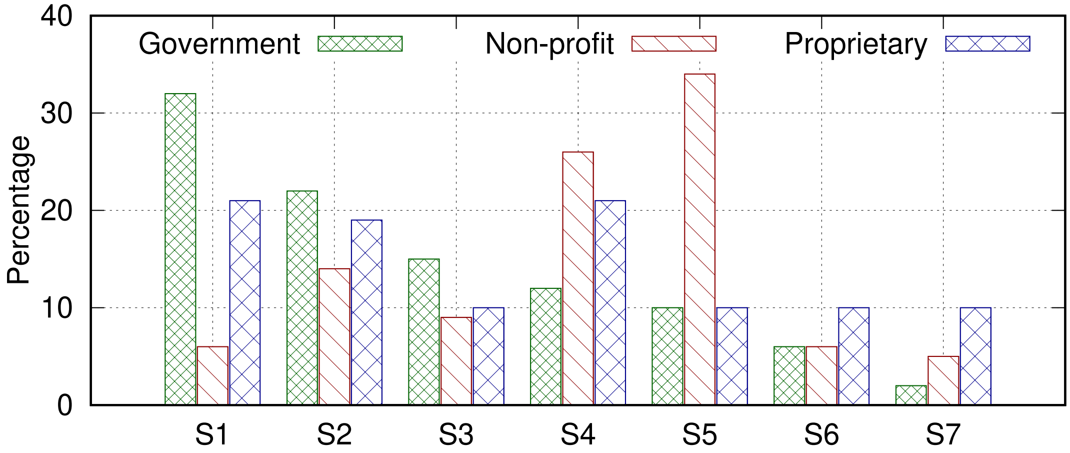


Figure 23: Comparing the data breach surfaces of Government, Non-profit, and proprietary hospitals. S1= Paper/Films, S2=Network Server, S3=EMR, S4=Other, S5=Email, S6=Laptop, S7=Desktop Computer.

negativity constraint (i.e., monotonously constraint) to learn important attributes of breach incidents.

Table 8 illustrates the 21 attributes used in our model, and Figure 25 shows the ten attributes that are directly (and mostly) correlated with the breached websites. As shown in the figure, when the website contains malware software (F20: Websites detected as malware by Sucuri API), it is (naturally) more likely to be involved in a data breach incident. Other features that highly correlated with websites' data breaches are (F15: The percentage of images retrieved by Pingdom API, and F13: The percentage of font retrieved by Pingdom API).

Takeaway: (**RQ2.**) 156 non-profit hospitals were associated with reported breach incidents, which is significantly higher than the remaining two categories. Moreover, hospitals' online presence security features play a clear role in their potential to be targeted, according to the top 10 attributes of hospital websites indicative of data breach incidents. We also notice that proprietary hospitals are the least susceptible to breaches.

5.6 Summary and Work to be Completed

Recent reports showed an increasing trend of attacks targeting hospital networks to compromise sensitive patient data. In this paper, we investigated the online presence of hospitals by analyzing their websites. Benefiting from a categorization into government, non-profit, and proprietary hospitals, we conduct a comparative study that sheds light on various structural and security features. Of particular note, we investigated the SSL certificate validity, the related issues among hospitals' websites, and malicious associated behaviors. Leveraging the collected attributes as features, we demonstrate the most important attributes indicative of websites associated with data breach inci-

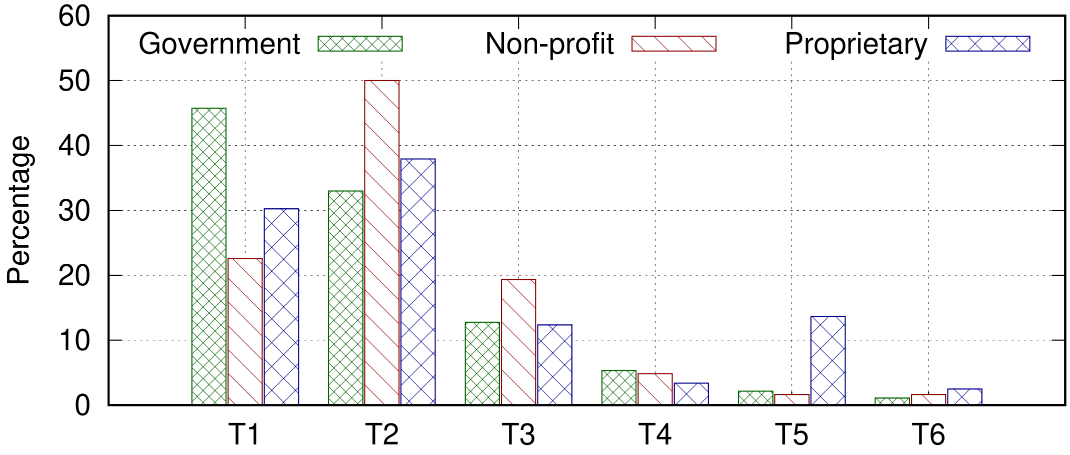


Figure 24: Comparing the data breach types of Government, Non-profit, and proprietary hospitals. T1=Hacking/IT, T2=Unauthorized Access, T3=Theft, T4=Loss, T5=Improper Disposal, T6=Other.

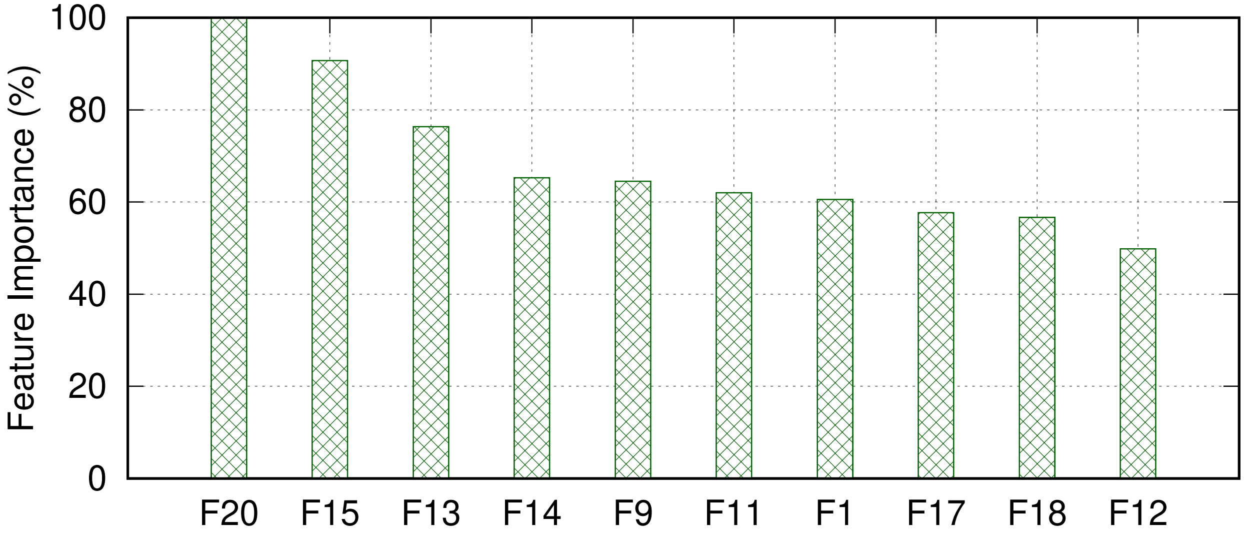


Figure 25: The domain- and content-level attributes importance (%) in distinguishing hospital websites associated with data breaches. The titles and descriptions of all features are shown in Table 8

dents and helpful in understanding their security. Our findings are among the first steps toward achieving patient security, alarmingly highlighting the lax security in many hospitals' websites.

In the future, we aim to further investigate the security of hospital websites by exploring the prevalence of DNSSEC (Domain Name System Security Extensions) implementation in hospital websites. Our goal is to assess the adoption and implementation of DNSSEC within the healthcare sector's web infrastructure, specifically focusing on hospital websites. Moreover, we plan to examine the HTTP header attributes for security evaluation to assess the security measures and configurations in hospital websites. By conducting such analysis, we can provide insights into the

Table 8: Attributes extracted for data breach analysis.

	Title	Description
F1	Certificate Invalid	The browser fails to verify website certificate
F2	Certificate Unmatched	The website name does not match SSL certificate
F3	Certificate_Expired	The website certificate becomes invalid
F4	Validity_Days_Left	The remaining validity days of website certificate
F5	Positives	The website domain is detected by VirusTotal API
F6	Malicious_Site	Websites detected as malicious by VirusTotal API
F7	Malware_Site	Websites detected as by malware VirusTotal API
F8	Phishing_Site	Websites detected as by VirusTotal API phishing
F9	Page_Size (MB)	The website average page size of in MegaByte
F10	Load_Time (S)	The website average page load time of in seconds
F11	Number of Requests	The website average number of requests
F12	CSS	The percentage of CSS retrieved by Pingdom API
F13	Font	The percentage of font retrieved by Pingdom API
F14	HTML	The percentage of HTML retrieved by Pingdom API
F15	Image	The percentage of images retrieved by Pingdom API
F16	Redirect	The percentage of redirect retrieved by Pingdom API
F17	Script	The percentage of script retrieved by Pingdom API
F18	XHR	The percentage of XHR retrieved by Pingdom API
F19	Blacklisted Flag	Websites detected as blacklisted by Sucuri API
F20	Malware Flag	Websites detected as malware by Sucuri API
F21	DNSSec Flag	Websites detected using DNSSec

effectiveness of security protocols, such as HSTS (HTTP Strict Transport Security), CSP (Content Security Policy), and other security-related headers.

6 Concluding Remarks

In this dissertation, we explored three key areas related to data breaches and security in healthcare organizations. First, we conducted a long-term analysis of data breaches, revealing a significant increase in incidents from 2010 to 2020. Our study identified various attack types, including internal, external, and partner attacks, with financial motives being the primary driving factor. The high cost associated with data breaches, along with the classification of threat actor actions, provides valuable insights into the evolving threat landscape. Second, we investigated the influence of population factors on data breach exposure, highlighting the correlation between the number of adults, state population, and incident rates. The study also uncovers the most breached assets and provides a timeline analysis, emphasizing the importance of prompt incident discovery. Finally, we focused on the hospitals' online presence and website security. We examined SSL certificate validity, associated issues, and malicious behaviors. The findings shed light on the vulnerabilities and lax security practices observed in many hospital websites, emphasizing the need for improved security measures. Overall, this dissertation provides a comprehensive understanding of data breaches and security in healthcare organizations. It offers insights into trends, risks, and mitigation strategies, guiding future research to enhance cybersecurity in the healthcare sector.

References

[1] A. O. Adebayo. A foundation for breach data analysis. *Journal of Information Engineering and Applications*, 2(4):17–23, 2012.

[2] A. Alabduljabbar, R. Ma, S. Alshamrani, R. Jang, S. Chen, and D. Mohaisen. Poster: Measuring and Assessing the Risks of Free Content Websites. In *Network and Distributed System Security Symposium,(NDSS'22), San Diego, California*, 2022.

[3] A. Alabduljabbar, R. Ma, S. Choi, R. Jang, S. Chen, and D. Mohaisen. Understanding the Security of Free Content Websites by Analyzing their SSL Certificates: A Comparative Study. pages 19–25. ACM, 2022.

[4] A. Alabduljabbar and D. Mohaisen. Measuring the Privacy Dimension of Free Content Websites through Automated privacy policy analysis and annotation. In *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 860–867. ACM, 2022.

[5] M. Alkinoon, S. J. Choi, and D. Mohaisen. Measuring Healthcare Data Breaches. In *Information Security Applications - 22nd International Conference, WISA 2021, Jeju Island, South Korea, August 11-13, 2021, Revised Selected Papers*, volume 13009 of *Lecture Notes in Computer Science*, pages 265–277. Springer, 2021.

[6] M. Alkinoon, M. Omar, M. Mohaisen, and D. Mohaisen. Security Breaches in the Healthcare Domain: A Spatiotemporal Analysis. In *Computational Data and Social Networks - 10th International Conference, CSoNet 2021, Virtual Event, November 15-17, 2021, Proceedings*, volume 13116 of *Lecture Notes in Computer Science*, pages 171–183. Springer, 2021.

[7] O. Alrawi and A. Mohaisen. Chains of Distrust: Towards Understanding Certificates Used for Signing Malicious Applications. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 451–456. ACM, 2016.

[8] APIVoid. A framework provides JSON APIs useful for cyber threat analysis, threat detection and prevention, 2022.

[9] M. P. Bach, S. Seljan, B. Jakovic, A. Buljan, and J. Zoroja. Hospital Websites: From the Information Repository to Interactive. 2019.

[10] A. Bates, J. Pletcher, T. Nichols, B. Hollembaek, D. Tian, K. R. B. Butler, and A. Alkhelaifi. Securing SSL Certificate Verification through Dynamic Linking. In *Proceedings of the 2014*

ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014, pages 394–405. ACM, 2014.

[11] J. Berkowsky and T. Hayajneh. Security issues with certificate authorities. In *8th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2017, New York City, NY, USA, October 19-21, 2017*, pages 449–455. IEEE, 2017.

[12] Y. Chen and Z. Su. Guided differential testing of certificate validation in SSL/TLS implementations. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2015, Bergamo, Italy, August 30 - September 4, 2015*, pages 793–804. ACM, 2015.

[13] M. Chernyshev, S. Zeadally, and Z. Baig. Healthcare data breaches: Implications for digital forensic readiness. *Journal of medical systems*, 43(1):1–12, 2019.

[14] S. J. Choi and M. E. Johnson. Understanding the Relationship Between Data Breaches and Hospital Advertising Expenditures. *The American Journal of Managed Care*, 25(5), January 2019.

[15] T. Chung, Y. Liu, D. R. Choffnes, D. Levin, B. M. Maggs, A. Mislove, and C. Wilson. Measuring and Applying Invalid SSL Certificates: The Silent Majority. In *Proceedings of the 2016 ACM on Internet Measurement Conference, IMC 2016, Santa Monica, CA, USA, November 14-16, 2016*, pages 527–541. ACM, 2016.

[16] J. Clark and P. C. van Oorschot. SoK: SSL and HTTPS: Revisiting Past Challenges and Evaluating Certificate Trust Model Enhancements. In *2013 IEEE Symposium on Security and Privacy, SP 2013, Berkeley, CA, USA, May 19-22, 2013*, pages 511–525. IEEE Computer Society, 2013.

[17] CloudFlare. What is a domain name registrar?, year = 2022.

[18] CloudFlare. What is a top-level domain (TLD)?, 2022.

[19] S. E. Coull, A. M. White, T. Yen, F. Monrose, and M. K. Reiter. Understanding domain registration abuses. *Comput. Secur.*, 31(7):806–815, 2012.

[20] M. Cova, C. Leita, O. Thonnard, A. D. Keromytis, and M. Dacier. An Analysis of Rogue AV Campaigns. In *Recent Advances in Intrusion Detection, 13th International Symposium, RAID 2010, Ottawa, Ontario, Canada, September 15-17, 2010. Proceedings*, volume 6307 of *Lecture Notes in Computer Science*, pages 442–463. Springer, 2010.

[21] L. Coventry and D. Branley. Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. *PubMed*, April 2018.

[22] V. Enterprise. ”Introduction to the 2020 DBIR: Verizon Enterprise Solutions”. *Verizon Enterprise*, 2020.

[23] V. Enterprise. Verizon Data Breach Investigations Report, 2022.

[24] O. for Civil Rights (OCR). Breach Notification Rule, 2013.

[25] O. for Civil Rights (OCR). U.S. Department of Health and Human Services Office for Civil Rights Breach Portal, 2023.

[26] T. I. O. for Standardization (ISO). International Organization for Standardization: 3,166 COUNTRY CODES, 2021.

[27] K. Gwebu and C. W. Barrows. Data breaches in hospitality: is the industry different? *Journal of Hospitality and Tourism Technology*, 2020.

[28] D. Healthcare. What is the Difference Between Non-Profit and For-Profit Hospitals?, 2022.

[29] HIPPA. What are the penalties for hipaa violations? `https://tinyurl.com/bdh6aa47`, 2023.

[30] Health Sector Cybersecurity Coordination Center (HC3). A cost analysis of healthcare sector data breaches, 2019.

[31] U.S. Department of Health and Human Services (HHS). Business associates, 2019.

[32] I. B. M. C. (IBM). How much does a data breach cost?, 2021.

[33] I. B. M. C. (IBM). What is a data breach?, 2023.

[34] I. S. R. G. (ISRG). Let’s Encrypt.

[35] F. Kamoun and M. Nicho. Human and Organizational Factors of Healthcare Data Breaches: The Swiss Cheese Model of Data Breach Causation And Prevention. *Int. J. Heal. Inf. Syst. Informatics*, 9(1):42–60, 2014.

[36] F. Kamoun and M. Nicho. Human and organizational factors of healthcare data breaches: The swiss cheese model of data breach causation and prevention. *Int. J. Heal. Inf. Syst. Informatics*, 9(1):42–60, 2014.

[37] D. Kim, H. Cho, Y. Kwon, A. Doupé, S. Son, G. Ahn, and T. Dumitras. Security Analysis on Practices of Certificate Authorities in the HTTPS Phishing Ecosystem. In *ASIA CCS ’21: ACM Asia Conference on Computer and Communications Security, Virtual Event, Hong Kong, June 7-11, 2021*, pages 407–420. ACM, 2021.

[38] D. Kim, B. J. Kwon, and T. Dumitras. Certified Malware: Measuring Breaches of Trust in the Windows Code-Signing PKI. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, pages 1435–1448. ACM, 2017.

[39] D. Kim, B. J. Kwon, K. Kozák, C. Gates, and T. Dumitras. The Broken Shield: Measuring Revocation Effectiveness in the Windows Code-Signing PKI. In *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, pages 851–868. USENIX Association, 2018.

[40] B. J. Kwon, S. Hong, Y. Jeon, and D. Kim. Certified Malware in South Korea: A Localized Study of Breaches of Trust in Code-Signing PKI Ecosystem. In *Information and Communications Security - 23rd International Conference, ICICS 2021, Chongqing, China, November 19-21, 2021, Proceedings, Part I*, volume 12918 of *Lecture Notes in Computer Science*, pages 59–77. Springer, 2021.

[41] B. . E. law firm. HIPAA Regulations: General Provisions: Definitions: Health Plan - § 160.103, 2015.

[42] V. Liu, M. A. Musen, and T. Chou. Data Breaches of Protected Health Information in the United States. *JAMA*, 313(14):1471–1473, 04 2015.

[43] C. Makridis and B. Dean. Measuring the economic effects of data breaches on firm outcomes. *Journal of Economic and Social Measurement*, 43(1-2):59–83, 2018.

[44] A. McLeod and D. Dolezel. Cyber-analytics: Modeling factors associated with healthcare data breaches. *Decision Support Systems*, 108:57–68, 2018.

[45] N. Menachemi1 and T. H. Collum. Benefits and drawbacks of electronic health record systems, 2011.

[46] U. Meyer and V. Drury. Certified Phishing: Taking a Look at Public Key Certificates of Phishing Websites. In *Fifteenth Symposium on Usable Privacy and Security, SOUPS 2019, Santa Clara, CA, USA, August 11-13, 2019*. USENIX Association, 2019.

[47] M. A. Mishari, E. D. Cristofaro, K. M. E. Defrawy, and G. Tsudik. Harvesting SSL Certificate Data to Identify Web-Fraud. *Int. J. Netw. Secur.*, 14(6):324–338, 2012.

[48] O. Networks. Largest Healthcare Data Breaches Reported in February 2022 Confirms Need for Network Security Based on Zero Trust Microsegmentation, 2022.

[49] NLTK. Natural Language Toolkit, 2022.

[50] U. D. of Health and H. S. (HHS). Health Insurance Portability and Accountability Act, 2023.

[51] U. D. of Homeland Security. Homeland Infrastructure Foundation-Level Data (HIFLD), 2019.

[52] T. O. of the National Coordinator for Health Information Technology (ONC). Benefits of EHRs, 2017.

[53] Office for Civil Rights. Breach Notification Rule, 2013.

[54] Pingdom. Website Performance and Availability Monitoring, 2022.

[55] W. Raghupathi, V. Raghupathi, and A. Saharia. Analyzing Health Data Breaches: A Visual Analytics Approach. *AppliedMath*, (1):175–199, 2023.

[56] N. P. D. Rank. The National Practitioner Data Bank (NPDB), 2021.

[57] G. V. Research. Digital Health Market Size, Share & Trends Analysis Report, 2021.

[58] M. Roetteler, M. Naehrig, K. M. Svore, and K. E. Lauter. Quantum Resource Estimates for Computing Elliptic Curve Discrete Logarithms. In *Advances in Cryptology - ASIACRYPT 2017 - 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part II*, volume 10625 of *Lecture Notes in Computer Science*, pages 241–270. Springer, 2017.

[59] A. Sarabi, P. Naghizadeh, Y. Liu, and M. Liu. Risky business: Fine-grained data breach prediction using business profiles. *Journal of Cybersecurity*, 2(1):15–28, 2016.

[60] A. H. Seh, M. Zarour, M. Alenezi, A. Sarkar, A. Agrawal, R. Kumar, and P. R. Khan. Healthcare Data Breaches: Insights and Implications. *Healthcare*, 8:133, 05 2020.

[61] B. K. Siddartha and G. K. Ravikumar. Analysis of Masking Techniques to Find out Security and other Efficiency Issues in Healthcare Domain. In *Third International conference on I-SMAC*, pages 660–666, 2019.

[62] T. Smith. Examining Data Privacy Breaches in Healthcare. Technical report, Walden University, 2016.

[63] T. T. Smith. Examining Data Privacy Breaches in Healthcare. Technical report, 2016.

[64] B. Steinwald, , and D. Neuhauser. The Role of the Proprietary Hospital, 1970.

[65] Sucuri. Website security check and malware scanner, 2023.

[66] U.S. HHS. Business Associate Contracts, 2013.

[67] M. J. Van Eeten and J. M. Bauer. Economics of malware: Security decisions, incentives and externalities. 2008.

[68] L. Vargas, L. Blue, V. Frost, C. Patton, N. Scaife, K. R. B. Butler, and P. Traynor. Digital Healthcare-Associated Infection: A Case Study on the Security of a Major Multi-Campus Hospital System. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019.

[69] Veris community. The Vocabulary for Event Recording and Incident Sharing (VERIS), 2021.

[70] VirusTotal. Analyze suspicious files and URLs to detect types of malware, automatically, 2023.

[71] W3Techs. Usage statistics of Default protocol https for websites, 2022.

[72] S. Walker-Roberts, M. Hammoudeh, O. Aldabbas, M. Aydin, and A. Dehghantanha. Threats on the horizon: Understanding security threats in the era of cyber-physical systems. *The Journal of Supercomputing*, 76(4):2643–2664, 2020.

[73] S. Walker-Roberts, M. Hammoudeh, and A. Dehghantanha. A Systematic Review of the Availability and Efficacy of Countermeasures to Internal Threats in Healthcare Critical Infrastructure. *IEEE Access*, 6:25167–25177, 2018.

[74] WHOIS. Registration data lookup tool, 2023.

[75] S. Wikina. What Caused the Breach? An Examination of Use of Information Technology and Health Data Breaches. *Perspectives in health information management*, page 1h, 10 2014.

[76] S. B. Wikina. What Caused the Breach? An Examination of Use of Information Technology and Health Data Breaches. *Perspectives in health information management*, 11(Fall), 2014.

[77] X. Yu, N. Samarasinghe, M. Mannan, and A. M. Youssef. Got Sick and Tracked: Privacy Analysis of Hospital Websites. IEEE, 2022.

[78] L. Zhang, D. R. Choffnes, T. Dumitras, D. Levin, A. Mislove, A. Schulman, and C. Wilson. Analysis of SSL certificate reissues and revocations in the wake of heartbleed. *Commun. ACM*, 61(3):109–116, 2018.