Journal of Medical Systems (2019) 43: 7
https://doi.org/10.1007/s10916-018-1123-2

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Healthcare Data Breaches: Implications for Digital Forensic Readiness

Maxim Chernyshev[1] · Sherali Zeadally[2] · Zubair Baig[3]

Received: 12 June 2018 / Accepted: 15 November 2018 / Published online: 28 November 2018

Abstract

While the healthcare industry is undergoing disruptive digital transformation, data breaches involving health information are not usually the result of integration of new technologies. Based on published industry reports, fundamental security safeguards are still considered to be lacking with many documented data breaches occurring as the result of device and equipment theft, human error, hacking, ransomware attacks and misuse. Health information is considered to be one of the most attractive targets for cybercriminals due to its inherent sensitivity, but digital investigations of incidents involving health information are often constrained by the lack of the necessary infrastructure forensic readiness. Following the analysis of healthcare data breach causes and threats, we describe the associated digital forensic readiness challenges in the context of the most significant incident causes. With specific focus on privilege misuse, we present a conceptual architecture for forensic audit logging to assist with capture of the relevant digital artefacts in support of possible future digital investigations.

Keywords Computer crime · Forensics · Health information management · Security · Threat

Introduction

Driven by the need to move away from incidental doctor-centered care towards more accessible patient-centered care, the healthcare industry is undergoing disruptive transformation [1]. The associated changes shall facilitate increased adoption of technology as part of the evolving health information technology (HIT) ecosystem. Technology-based innovation trends, specifically in the areas of (1) digital health, (2) big data and (3) precision health are instrumental in supporting the delivery of the future healthcare vision [2].

The ubiquitous Internet connectivity coupled with the increasing adoption of mobile, wearable and the Internet of Things (IoT) technologies will underpin solutions that handle unprecedented amounts of health information records. Fortunately for patients and practitioners, the increased availability of the data generated and collected shall enable more accurate and faster clinical and biomedical actions, including proactive life-saving interventions.

However, health information is also considered to be the most attractive target for cyber criminals. Depending on record completeness, a single patient's file can be sold for several hundreds of dollars on the dark web [3]. The transformational changes and integration of new technological elements into the HIT ecosystem will expand the attack surface of healthcare services [4]. The associated threat landscape suggests that data breaches involving health information are not generally the result of sophisticated attacks on contemporary technological building blocks, such as IoT sensors and wearable medical devices [5]. In contrast, widespread human errors, misuse and physical
actions such as loss and theft have been the major causes behind hundreds of publicly disclosed healthcare data breaches worldwide. This threat pattern is considered unique to the healthcare industry. Leveraging publicly available healthcare data breach information and industry reports, we examine the associated implications from a digital forensic perspective. We aim to identify key digital forensic readiness challenges associated with the primary causes of these breaches. The contributions of this paper are as follows:

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ Maxim Chernyshev
m.chernyshev@ecu.edu.au

[1] Edith Cowan University, Perth, Australia

[2] University of Kentucky, Lexington, KY 40506-0224, USA

[3] Commonwealth Scientific and Industrial Research Organisation (CSIRO), Data61, Melbourne, Australia

- We conduct an analysis of healthcare data breaches focusing on the location of the digital artifacts that can contain potential digital evidence.
- We identify relevant digital forensic readiness challenges that reflect the unique threat pattern of the healthcare sector.
- We present a conceptual architecture to address these challenges in investigations of incidents involving privilege abuse with a specific focus on electronic medical record (EMR) systems.

Despite widespread availability of healthcare data breach information and statistical analyses provided by industry, we are not aware of other works examining the associated implications with a specific digital forensic focus.

Data breaches in healthcare

Regulatory landscape overview

The highly sensitive nature of medical records has been recognized worldwide. There are several privacy frameworks and regulations in existence today in various countries. As shown in Table 1, several nations such as Australia, the United Kingdom (UK), and the European Union (EU) have enacted legislation that encompasses specific conditions pertaining to the handling and protection of health information under respective privacy laws [7–9]. The United States (US) has dedicated legislation as part of the long-standing Health Insurance Accountability and Portability Act (HIPAA) 1996 [6]. Although other nations do not necessarily have healthcare-specific provisions in their privacy legislation, health information is still usually covered under the broader definition of personal data.

Several regulations such as the HIPAA Breach Notification Rule in the US and the recently introduced Notifiable Data Breaches (NDB) scheme [10] in Australia mandate compulsory data breach notification requirements. Generally speaking, notifications are issued when the data breach poses high risk of harm to affected individuals, which is often the case with health information, given its highly sensitive nature.

To this effect, the US Department of Health and Human Services (HHS) data breach portal [14] lists over 2250 data breaches (including over 390 still under investigation) for the period between 2009 and 2018. The Australian NDB scheme received 63 submissions in just six weeks since its introduction, of which health service providers were responsible for the majority (24%) of all submissions.

Health information definition

The concept of health information requires specific attention. Sometimes, terms such as health records, medical records and health information are used almost interchangeably. Table 2 provides several definition summaries of these similar terms. There is no single universal description for what? for health information because we have slightly varying terms and definitions that have been adopted across the different contexts. Based on the key common aspects of these definitions, health information is considered to be:

Table 1 Major privacy frameworks and regulations

Country	Framework / Regulation	Healthcare-specific provisions
United States	Health Insurance Accountability and Portability Act (HIPAA) 1996 [6]	Dedicated
United Kingdom	Data Protection Bill 2017 [7]	Specific conditions [a]
European Union	General Data Protection Regulation (GDPR) 2018 [8]	Specific conditions [a]
Australia	Privacy Act 1988 [9], Notifiable Data Breaches (NDB) Scheme 2018 [10]	Specific conditions [b]
Singapore	Personal Data Protection Act (PDPA) 2012 [11]	Advisory guidelines
Canada	Personal Information Protection and Electronic Documents Act (PIPEDA) 2000 [12]	Not specified [c]
Japan	Act on the Protection of Personal Information 2005 [13]	Not specified
New Zealand	Privacy Act 1993 [3]	Not specified

[a] Specific definitions and rules around data concerning health, genetic data and biometric data

[b] Specific conditions on obtaining explicit consent, usage of health information and access to collected information

[c] Selected provinces have specialized health-related privacy legislation in place

Table 2 Definitions of health information

Framework / Regulation	Term used	Definition summary
HIPAA 1996 [6]	Protected health information (PHI)	Individually identifiable health information that is transmitted by electronic media, maintained in electronic media, or transmitted or maintained in any other form or medium [a].
Data Protection Bill 2017 [7]	Health record	Any record of information relating to someone's physical or mental health that has been made by (or on behalf of) a health professional.
GDPR 2018 [8]	Data concerning health	Personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status.
Privacy Act 1988 [9]	Health information	Any information about someone's health or a disability, as well as any other personal information collected while the person is receiving a health service.

[a] Excludes certain categories such as education and employment records

- Handled in any form, physical or digital.
- Associated with all aspects of personal health, including physical and mental health.
- Related to past, present and future encounters with medical practitioners
- Collected, transmitted, processed and stored by various types of organizations (not only healthcare providers).
- Usually linked to individually identifiable data.

Unlike other definitions of health information, which operate in broader terms, the HIPAA's protected health information (PHI) description also includes a list of 18 elements, which must be removed from the data set so that it is no longer identifiable health information. In addition to the more conventional attributes such as names and contact details, this list also includes endpoint identifiers such as device serial numbers, Internet Protocol (IP) addresses and Universal Resource Locators (URLs). Therefore, it is important to have full clarity on the particulars of the definition that is applicable to the legal context of any digital forensic investigation because identical sets of attributes may not necessarily be considered health information in different contexts.

Sensitivity and motives

It is also important to recognize the various sensitivity levels are associated with health information [15]. These levels, as shown in Table 3, are based upon the perceived impact of the breach on the individual's privacy and social wellbeing, as well as the potential types of crime that can be committed based on different categories of health information. For example, in certain contexts where healthcare systems do not include compulsory health insurance such as the US, identity theft using stolen health information can facilitate access to healthcare services by non-insured individuals and also allow them to obtain prescription drugs for financial gain [16]. Whilst financial gain is by far the most common motive, other reasons can include curiosity, grudge and espionage [5].

The attraction of individuals with nefarious objectives to health information is clearly motivated. Given the different types of crime are made possible using such data, the black market value of health information is at least 10 to 20 times more than the value of credit card data [17]. Depending on completeness, recency and accuracy, patient files can be sold starting from 10 USD per record up to 1000 USD per record [3]. For example, following a failed extortion attempt a database containing health information of 655,000 patients was made available for purchase through one of the dark web marketplaces [18]. In this case, the extortion demands were not targeted directly at the individuals whose information was exposed, but rather at the entities responsible for safeguarding their data. Scenarios where criminals become aware of sensitive diagnoses or clinical history via a data breach and then target the affected high-profile individuals are also theoretically possible.

Threat actors and actions

External actors are not necessarily always the biggest threat to health information. The fact that the sector is facing a unique threat pattern has been recognized by the latest special issue Protected Health Information Data Breach Report (PHIDBR) published by Verizon [5]. The 2018 report examines 1368 healthcare data breaches disclosed since the beginning of 2016 and in particular breaches that affect the healthcare industry or involve patients or their health information. Based on this report, the unique threats pertinent to the sector can be summarized as follows:

Table 3 Health information sensitivity levels based on [15]

Sensitivity	Data categories	Access control scope	Possible crime
Normal	Personal	Wide	Identity theft
Sensitive	Social Financial information	Based on medical staff role	Fraud
Highly sensitive	Health risks Clinical diagnoses	Treatment by nominated medical staff only	Extortion

- Internal actors (insiders) are associated with the majority (58%) of recent data breaches.
- Similar to electronic health information, breaches also affect paper-based records.
- Given the lack of fundamental, standard security controls across the affected entities, the sector remains highly susceptible to malware attacks and in particular ransomware.

In contrast to the previous special issue report by Verizon published in 2015 [23], which presented external actors as the most significant threat, we observe a slight pattern shift where internal actors have become the most significant concern. Actions performed by internal actors encompass both human error and misuse. Although the majority of insiders are not malicious, some form of misuse is still involved in almost a third (29.5%) of all analyzed data breach incidents.

Table 4 presents several published healthcare data breach reports and key associated threat pattern characteristics. We found minor disparities in the reported results leading to the potential lack of clarity around the associated landscape. Thus, we perform our own analysis in order to obtain a clearer understanding of the associated threats.

Based on the types of breaches and threat action groups shown in Fig. 1, when it comes to the "how" aspect – which represents the cause of the breach – internal actors would primarily be associated with unauthorized access, disclosure, loss and improper disposal of health information. In contrast, external actors would generally be associated with hacking, malware and physical incidents – primarily, theft. Thus, both internal and external actors need to be considered as similarly significant groups.

Although physical crimes such as device and document theft have been a major issue in the past, the rate of theft has dropped significantly in the last three years. Stolen devices such as laptops and external storage media may contain digital forensic artifacts that could be used to determine whether health information was actually accessed by an unauthorized individual. However, the recovery rate of stolen office equipment in the US is reported to be under 6% [24]. Thus, medical information present on stolen devices is assumed to have been compromised by default and we do not consider the digital forensic implications associated with physical threats further in this work. Similarly, breaches involving human error are excluded from the scope of the subsequent discussion because document mishandling, improper disposal, loss and publishing errors are unlikely to be a common focus of digital investigations.

Digital forensic implications

Digital forensics

The discipline of digital forensics aims to extract court-admissible evidence by using scientifically designed and validated methods applied to data on digital devices [25]. The common digital forensic process involves evidence identification, collection, examination, analyses and reporting. The integrity of the evidence must be verifiably preserved throughout all stages of the process. There are several digital forensic

Table 4 Notable Data Breach Reports

Publisher	Coverage	Sample size	Context	Key threat pattern characteristic
Verizon [5]	2016–2017	1368	Global	Prevalence of internal actors (57.5%) closely followed by external actors (42%)
Bitglass [19]	2014–2017	1179	US	Increasing pervasiveness of external actors (70.9% in 2017)
Maryland Health Care Commission [20]	2010–2016	1780	US [a]	Growing concern over external actors in 2014–2016
Office of the Australian Information Commissioner (OAIC) [21]	Q1 2018	63	Australia	Most breaches (24%) reported by health service providers, internal actors (50%), external actors (44%)

[a] Maryland, US compared to the entire US

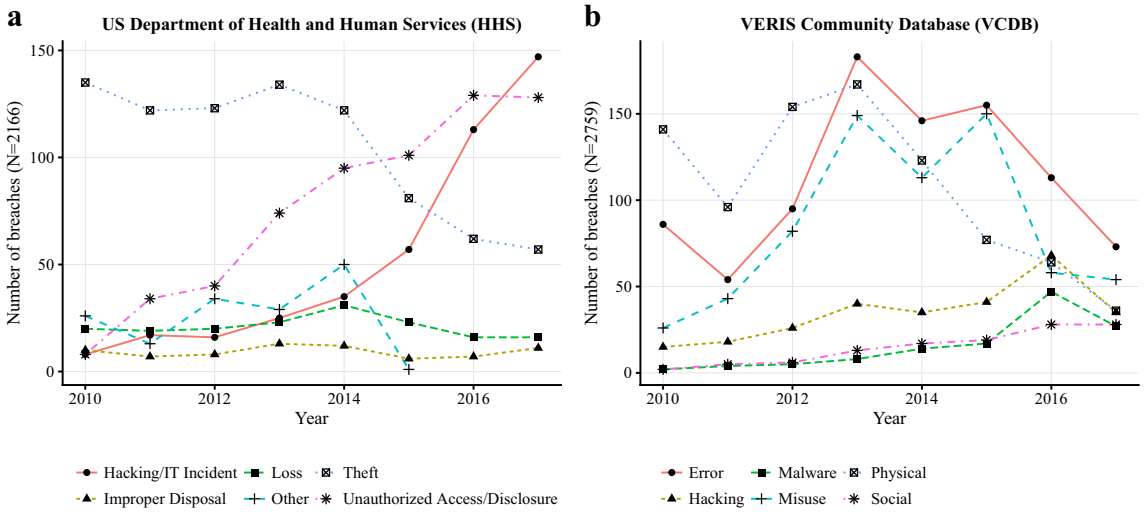

Figure 1 Breakdown of healthcare breach types by year based on data provided by the US Department of Health and Human Services (HHS) including archived breaches and breaches under investigation (1A) [14]. Breakdown of threat actor actions by year based on data available in the VERIS Community Database (VCDB) (1B) [22]

process models (DFPM) that are used to guide digital investigations in a structured and documented manner [26–28].

In particular, The integrated digital forensic process model (IDFPM) was proposed to standardize the forensic process terminology and eliminate the differences that exist among the various other models [29]. One of the key features of IDFPM is the introduction of the preparation phase to satisfy the need to establish operational and infrastructure readiness as a critical component of the process model. Subsequently, we argue that the healthcare industry can facilitate improved outcomes in digital investigations by assisting with the establishment of the necessary infrastructure readiness aspects. Such readiness would lead to the increased usefulness of evidence as well as the decrease of the associated investigation costs [30]. In the next sections, we focus on the identification of the relevant data sources associated with the most common threat actions and varieties. Knowing how to facilitate digital artifact collection and what sources to target is one of the key steps in the forensic readiness implementation process [31].

Threat actions

Given the motives and various types of actors involved in incidents leading to healthcare breaches, we also examine the different types of threat actions (such as hacking, ransomware, phishing and privilege abuse) and the specific types of assets where digital artifacts containing potential evidence may be located. From Fig. 2, we observe that:

- Specific actions taken as part of hacking are generally not known. This is likely because cyber criminals can rely on tailored tactics and zero-day exploits. As the alleged representative of the "TheDarkOverlord" hacking collective stated, "I keep all my exploits private for my own use. Never publish it…" [18]. The tracks associated with these exploitations, if left by the attackers, may be difficult to uncover without expensive specialist forensic assistance.
- Ransomware is by far the most common type of malware used.

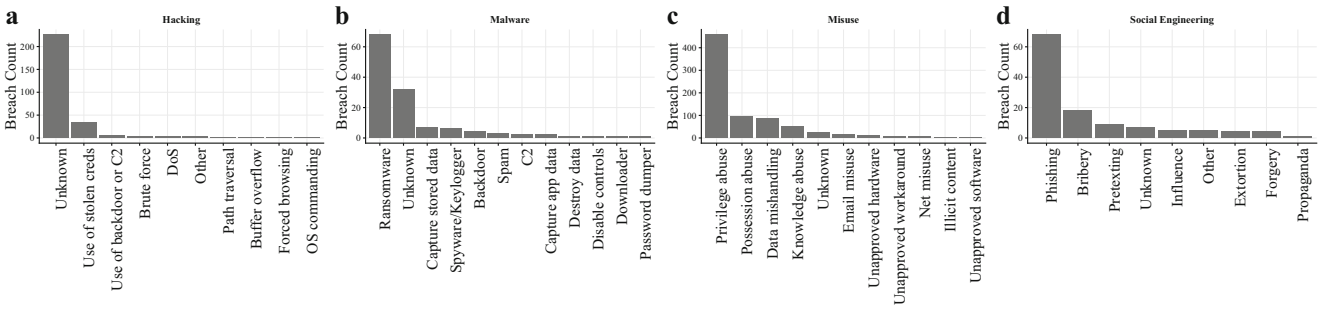

Figure 2 Breakdown of healthcare breach threat action varieties for hacking (2A), malware (2B), misuse (2C) and social engineering (2D). As explained in section 2B, physical actions and human error are not included in the analysis [22]

- Privilege abuse is the most significant concern when it comes to misuse.
- Phishing is the most popular tactic used as part of social engineering attacks.

Subsequently, we also examine the types of assets most commonly involved in these incidents. As shown in Fig. 3, we identify the following sources where digital artifacts may be located:

- Servers (such as web and email servers).
- Databases.
- End-user computers.
- Removable media.

The relatively high prevalence of paper and films in the HHS data set (Fig. 3A) is explained by the fact that breach causes that include both improper disposal, human error and misuse are grouped under the same category (unauthorized access and disclosure).

Forensic readiness implications

Hacking

In the context of hacking, achieving forensic readiness is considerably difficult and perhaps is no different to any other industry. The attack surface is vast and can encompass all elements that comprise the infrastructure of the business as well as any of its third-party services including desktop and mobile endpoints, local wired and wireless networks, border routers, as well as the Internet service provider (ISP) and the cloud infrastructure. To keep up with the constantly changing landscape infrastructure, the science of digital forensics must undergo rapid, continuous evolution [32].

Therefore, digital forensic readiness for investigations involving targeted and random hacking is largely dependent on the security architecture of the HIT ecosystem elements and its ability to capture and preserve vast amounts of digital artifacts that are associated with network-borne threats. These artifacts can be collected via:

1) Traditional sources such as routers, firewalls, security information and event management (SIEM) solutions, intrusion detection systems (IDS) (host and network), honeypots, data loss prevention (DLP) solutions and others.
2) Dynamically evolving components such as software defined networks (SDN), contemporary cloud architectures such as serverless computing and microservices, application programming interfaces (APIs) and container-based virtual environments.

Given that breach discovery can take months and even years [5], the key aspect of readiness in this context is associated with the ability to retain and efficiently analyze vast amounts of digital artifacts. Forensic readiness can be facilitated by using a cloud forensic logging-as-a-service solution to enable artifact capture and retention [33]. Subsequently, intelligent and highly scalable visualization schemes (such as those using frequent item mining and hypergraphs) can be used to assist with streamlining the artifact analysis to pinpoint the potential evidence [34].

Ransomware

In the context of ransomware, which prevents resource access by authorized users until a ransom is paid, achieving digital forensic readiness is also very challenging. Modern ransomware variants called “cryptolockers” encrypt user data and any mounted backup locations upon infection and subsequently demand ransom payment in cryptocurrency such as

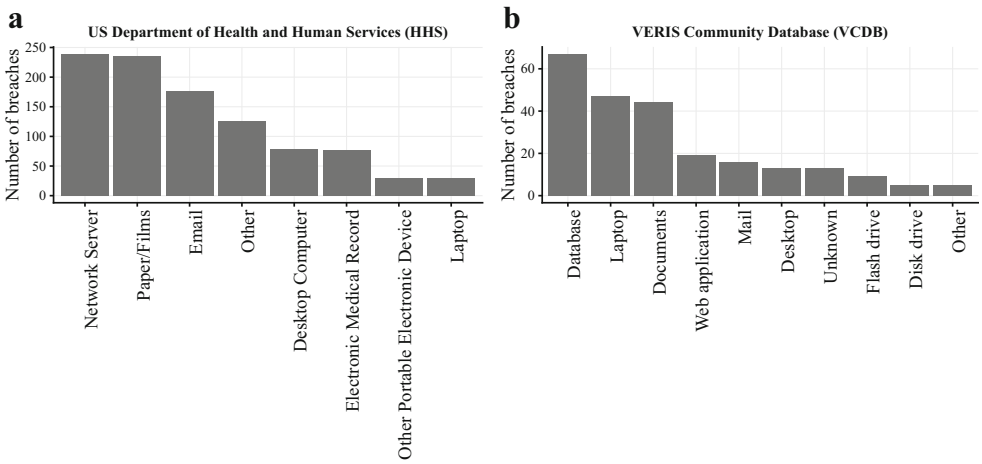

Fig. 3 Locations of breached health information based on data breaches not caused by theft or loss (3A) [14]. Top ten asset varieties compromised in healthcare data breaches (3B) [22]

Bitcoin. To identify the perpetrators, digital investigations handling ransomware infections usually adopt the "follow the money" strategy [35].

Identification of payment recipients requires the ability to apply the relevant Bitcoin and blockchain de-anonymization tools and techniques. For example, the freely available open-source tool called BitCluster [36] can be used to analyze Bitcoin transactions and group them by participating entities based on public key hashes, which can be effective when tracing ransom payments associated with the same public key as long the recipients maintain this key for the duration of the campaign.

Several commercial services such as Elliptic[1] state that they offer the capabilities necessary to detect and investigate criminal activity that involve cryptocurrency leading to criminal convictions [37]. The need to focus on these currencies as part of investigations has facilitated the emergence of the "crypto forensics" concept. In addition to the availability of validated data mining-based approaches that aim to identify payment recipients based on transaction patterns, readiness can be facilitated by cross-jurisdictional sharing of information by cryptocurrency payment processors to enable monitoring the financial criminal activities such as tax evasion and money laundering [35].

Phishing

Combating phishing is challenging and resource-intensive [38]. Phishers have the advantage of being able to easily spawn a new infrastructure and leverage infected web servers and botnets to support their campaigns. The mechanisms that trick the targets into following links and opening malicious attachments are also evolving constantly. There are many approaches aimed at detecting phishing emails including those based on email content, structural characteristics, behavior and hybrid algorithms [39]. To enable the collection of relevant digital artifacts, a traceback framework that includes a forensic backend to capture the relevant attributes of phishing emails to support the subsequent investigations would need to be established. However, as phishing attacks are constantly advancing, there are still many unresolved challenges to overcome [40].

Privilege abuse

Health information handled by a service resides in a data store. Although emerging architectures to be adopted as part of the digital transformation of healthcare will feature a higher degree of decentralization, most health service providers still rely on systems implemented using the traditional client-server architectures.

[1] https://www.elliptic.co/

Specifically, Electronic Medical Record (EMR) systems usually use a centralized data store supported by a relational database management system (RDBMS) which can be accessed by a local or remote client. Whilst access to health information is generally controlled via role-based access control (RBAC) policies, not all EMR systems share the same degree of granularity when it comes to their support for policy definitions and the effectiveness of RBAC controls often depends on the correctness of these definitions. In addition, some systems require specific considerations to support open access policies to enable streamlined handling of emergencies [41].

Subsequently, users of EMR systems often have the ability to access more health information than necessary based on their role and patient context, as evident by the number of breaches caused by misuse or unauthorized internal access. Several examples of internal actors who have accessed many patient records via an EMR system can be found in the descriptions of data breaches published via the HHS portal [14]. Although it is not clearly evident how such cases of privilege misuse have been uncovered, it reasonable to assume that discovery would require either random or targeted access log audits to take place or be prompted by external discoveries. Such audits can be manual, heuristics-based or powered by machine learning algorithms [42]. Therefore, from a digital forensic infrastructure readiness perspective, investigation of privilege abuse would be supported by the ability to identify and produce forensically sound digital evidence based on the digital artifacts contained in EMR audit logs.

Table 5 presents a summary of the various types of threats discussed and the associated forensic readiness challenges from an infrastructure perspective. These threats are applicable to the entire industry landscape. In the context of EMR systems, however, the issue of privilege abuse could be considered a high-impact area for achieving infrastructure forensic readiness due to high prevalence.

Misuse of electronic medical record (EMR) systems

Traces of actions performed by EMR users (or other connected services such as programmatic access clients) are expected to be captured in audit logs [41]. In fact, technical safeguards contained in the HIPAA security standards include specific requirements around establishing "mechanisms that record and examine activity in information systems that contain or use electronic protected health information" [6]. However, given that the standard has no associated implementation specifications, specific implementations normally differ making it difficult to compare their logging actions. First, this could be because a large portion of the smaller healthcare service providers do not have an established information security capability. There is also lack of incentive to fund and implement specific information security initiatives that focus on key risks associated with the lack of auditable activity

Table 5 Digital forensic readiness complexity levels for top healthcare threat types

Threat type	Prevalence [a]	Data source diversity	Data source examples	Infrastructure readiness challenges
Hacking	Medium	Extreme	Firewall Intrusion Detection System (IDS) Web server Hypervisor host Container orchestration service	Limited ability to process and visualize vast amounts of digital artifacts
Ransomware	Low	High	Local computer Cryptocurrency blockchain	Lack of cross-jurisdictional cooperation to support the "follow the money" strategies
Phishing	Low	Medium	Email content and metadata Landing server	Highly transient nature of linkable artifacts
Privilege abuse	High	Low	Database server Application access logs	Varying degree of logging built into Electronic Medical Record (EMR) systems

[a] Based on data breach count per threat variety overall as presented in the VERIS Community Database (VCDB) [22]

trails. Such entities also possibly run outdated or freely available EMR systems that cannot be patched or upgraded due to business and technological constraints. In other words, integration of additional components than can facilitate forensically sound audit logging may not always be possible, especially if not natively supported by the EMR system already in place. Second, in absence of a formal specification that defines exactly what events and attributes must be logged, the way to address the requirement remains open to interpretation thus representing one of the other key challenges. Therefore, system architects must consider the concept of mandatory log events (MLEs) (forensics-enabling activities and actions) as part of the solution design and implementation. Identification of MLEs can be achieved using standards, resource and heuristics-driven methods as described in [43]. However, as the study suggests, identifying the best way to specify mandatory logging requirements for ease of comprehension and adoption by software engineers remains on open research challenge.

Availability of auditable action trails is crucial in safeguarding patient privacy and assisting with investigation of incidents involving privacy violations via means of privilege abuse. Once available, the data can be used to facilitate auditing activities at various sophistication levels such as from random audits, regular algorithmic audits, rule-based alerting, machine learning-powered behavior analysis, and intelligent proactive analytics. In the latter category, we can observe vendors introducing the concepts of Proactive Patient Privacy Analytics (P3A) [44] and Ambient Cognitive Cyber Surveillance [45] for healthcare specifically aimed at addressing the issue of activity log capture and streamlining of automated analysis to detect violations. The specific analysis use cases are also unique to the healthcare sector and reflects the previously discussed motives such as personal gain, snooping or curiosity. For example, such use cases focus on uncovering specific patterns involving access to health information of co-workers, neighbors, family members and high-profile individuals [44].

Despite many privacy-enhancing and digital forensic benefits, the EMR system landscape faces the challenging task of integrating forensically sound audit logging capabilities. These challenges include both the requirement specification and implementation issues, and there is an opportunity to introduce intelligent solutions that are able to identify MLEs based on operations of health information and can preserve the associated attributes in a forensically sound manner.

Table 6 shows a comparison of several open-source EMR systems. There are significant differences across their available audit logging capabilities. Some of these systems have been adopted worldwide and are used to handle millions of patient records. However, despite most of these systems having some form of audit logging focusing on patient record creation and modification, the majority of them do not have provisions for tracking viewing activity. Unfortunately, the lack of coverage of these activities means that there are likely no artifacts captured by these systems that could become

Table 6 Comparison of built-in audit logging capabilities of selected open source electronic medical record (EMR) solutions

Name	Version	Last updated	Access model	Audit logging capabilities
OpenMRS	2.3.1	2018	Web-based	Separate audit module that tracks create, update, and delete (CRUD) operations on database objects (not production-ready, not built-in)
OpenEMR	5.0.1	2018	Web-based	Database query-based logging of all operations on patient records with log record integrity checks and optional encryption (HIPAA-friendly)
FreeMED	n/a	2017	Web-based	Built-in provisions for logging patient record operations exist, but no specific logging method calls could be located in source code upon manual inspection by the authors
NOSH	2.0	2018	Web-based	Automated logging of create, update, and delete (CRUD) operations on data (not context-specific)
Solismed	2.3	2018	Web-based	Built-in audit log module that tracks access to all system activity areas that contain patient data
HospitalRun	1.0.0-beta	2018	Web-based Offline	Tracking of creation and modification of patient records

useful in privilege abuse investigations. Integration of audit logging capabilities, where not already present or inadequate, remains a challenging task. Thus, we propose a different conceptual solution.

Proposed audit logging architecture for EMR systems

The proliferation of cloud services and the variety of the "X as a service" type offerings did not bypass the field of digital forensics. Numerous solutions and architectures have been proposed to leverage this service delivery model to enable secure cloud-based aggregation and analysis of digital artifacts to improve forensic infrastructure readiness [46–49]. Subsequently, we leverage the established concept of forensic logging as a service (FaaS) to propose an audit logging architecture for EMR systems that is generic and EMR implementation-agnostic as shown in Fig. 4. To implement this architecture, an EMR deployment would require the integration of two additional elements namely, payload analyzers and identification module. These elements are not required to be part of the EMR system itself. Rather, they can be deployed alongside at the technical infrastructure level. Specifically, payload analyzers are expected to reside inline transparently and perform asynchronous lightweight operations such as identifying payloads carrying health information to flag MLE activities.

At the technical level, payload interception can be achieved via multiple options which include 1) placing a scriptable reverse proxy server in front of the back-end web server, 2) deploying a packet analyzer, 3) replaying binary database logs, 4) installing a custom browser extension (for limited access terminals) and others. Payload analyzers can be protocol-aware and can rely on simple heuristics (such as keyword, expression and medical term matching) to detect payloads of interest that need to be directed to the identification module for analysis. This module utilizes a multi-step process for health information identification, such as the PHI data identification chain based on natural languages processing (NLP) presented in [50]. Given the user and device context, the time of activity and the set of identified health information categories, the identification module uses the FaaS application programming interface (API) to enable the collection and preservation of the identified artifacts in the FaaS platform.

The FaaS component in the proposed architecture is based on the architectures described in [47, 48]. In particular, it describes a secure collection and preservation module, an examination and analysis module and a presentation module to support the remaining phases of the digital forensic process. In addition, the component features a provenance module which is responsible for guaranteeing the chain of custody and maintaining the record of artifact access and usage. The FaaS API can be leveraged by external digital forensic tools and privacy analytics solutions in read-only fashion to support additional use cases and analysis types. Such an architecture still requires practical validation using real-life EMR systems and a FaaS component implementation.

Despite it only being presented at the conceptual level, we believe it has practical applicability given its key benefits. First, it does not require modification of the EMR system itself because it only requires the ability to inject payload analyzers across the various EMR system component communication channels which are often based on standard technological components. Second, given the intelligent health information detection capability (as long as the payload itself is readable by the analyzer), such an architecture could be generically

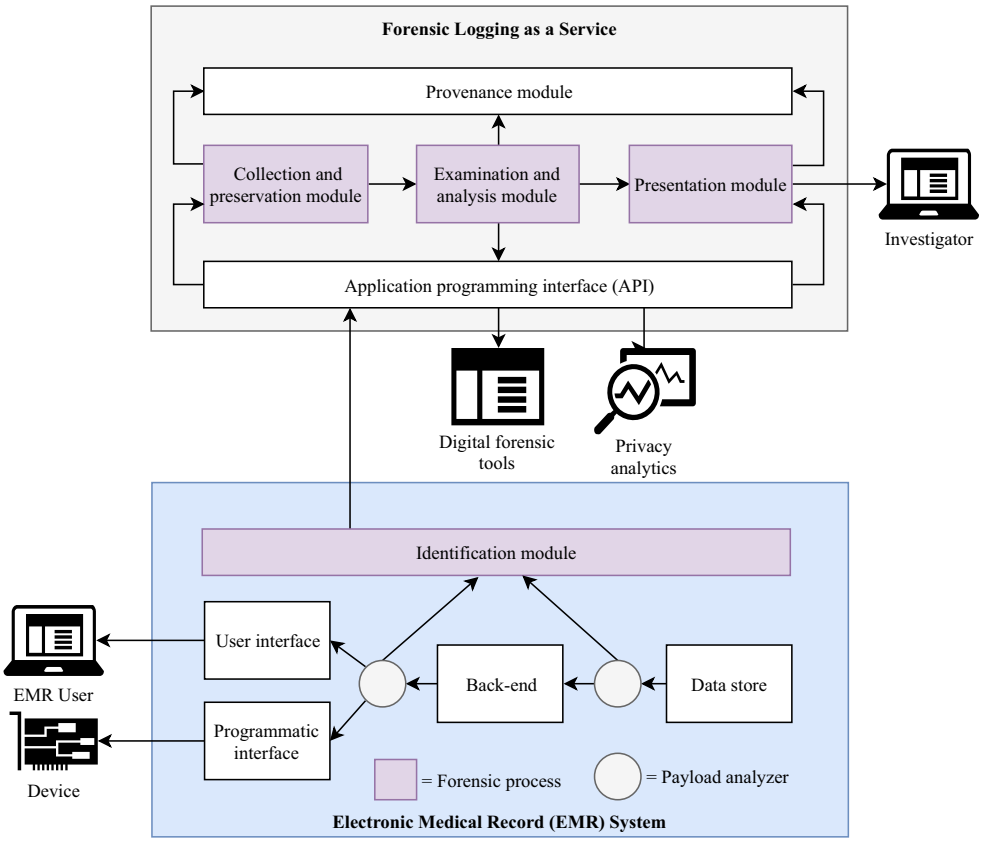

Fig. 4 Conceptual architecture describing the implementation of the digital forensic process in the context of forensic logging for an Electronic Medical Record (EMR) system. The architecture of the forensic logging as a service component is based on [47, 48]

applicable to systems deployed on premise, in the cloud and also future service deployments that support device communications of healthcare IoT services. Finally, the FaaS platform required is also a generic service that could be part of a larger system collecting artifacts from a diverse set of sources, enabling further context augmentation and additional artifact diversity.

We do acknowledge that not all of the EMR systems such as legacy systems already deployed today would be compatible with the proposed architecture. In cases where system updates that require mandatory audit logging are not feasible, database forensics-based approaches [51] could be considered as an alternative artifact identification and extraction avenue. Specifically, these tasks could be achieved based on the use of query logs natively supported by RDMBS to identify operations involving health information.

Another potential limitation of the proposed architecture is its explicit focus on assisting the enabling of MLE logging to support future digital investigations potentially involving privilege abuse. However, we believe that given the perceived prevalence of this threat in the industry, addressing it requires specialized solutions such as the architecture proposed in this paper. Collection of artifacts associated with threats applicable to EMR systems such as hacking can readily be achieved from traditional sources as we have described in Section C. At the same time, the proposed architecture can be utilized for other types of systems that involve access to and operations on potentially sensitive personal information. For example, a human resource application could leverage the same architecture following a recalibration of the associated payload analyzers and the identification module to recognize the data of interest, such as personnel data, for the associated operating domain.

Conclusion

Despite several transformational technological changes, the healthcare industry still remains highly susceptible to compromises of valuable health information. The threats behind numerous healthcare data breaches represent a unique pattern with the majority of incidents attributed to internal actors. Although human error is widespread, the rate of privilege abuse associated with unauthorized access to health information is alarming.

Focusing on privilege abuse, we have discussed the implications for infrastructure forensic readiness in the context of EMR systems used by many healthcare services providers to handle health information. We have examined several long-standing and freely available open-source EMR systems that are used to handle millions of patient records worldwide. We found that these systems do not always incorporate a sufficient level of forensic logging required to assist investigations

focusing on privilege abuse. Subsequently, without the necessary audit trails it may be challenging to support or refute hypotheses that aim to identify and prove the fact that these incidents occurred. To address this issue, we have proposed an architecture that incorporates an intelligent real-time artifact identification module which can be deployed alongside the EMS and be integrated into cloud forensic logging service. In future work, we plan to validate the conceptual architecture and further assess its practical applicability and the ease of implementation and EMS integration.

Acknowledgements We thank the anonymous reviewers for their valuable comments which helped us to improve the organization and content of this paper.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval.

This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Cresswell, K. M., and Sheikh, A., Health information technology in hospitals: current issues and future trends. Future Hospital Journal 2(1):50–56, 2015.
2. Bhavnani, S. P., Parakh, K., Atreja, A., Druz, R., Graham, G. N., Hayek, S. S., Krumholz, H. M., Maddox, T. M., Majmudar, M. D., Rumsfeld, J. S., and Shah, B. R., 2017 Roadmap for Innovation—ACC Health Policy Statement on Healthcare Transformation in the Era of Digital Health, Big Data, and Precision Health: A Report of the American College of Cardiology Task Force on Health Policy Statements and Systems of Care. Journal of the American College of Cardiology 70(21):2696–2718, 2017. https://doi.org/10.1016/j.jacc.2017.10.018.
3. Trustwave, The value of data: a cheap commodity or a priceless asset, 2017.
4. Islam, S. R., Kwak, D., Kabir, M. H., Hossain, M., and Kwak, K.-S., The internet of things for health care: a comprehensive survey. IEEE Access 3:678–708, 2015.
5. Verizon, Protected health information data breach report, 2018.
6. U.S. Department of Health & Human Services (HHS), The HIPAA privacy rule. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. Accessed 8 April 2018.
7. Information Commissioner's Office (ICO), Data Protection Bill 2017. https://ico.org.uk/for-organisations/data-protection-bill/. Accessed 8 April 2018.
8. European Union (EU), Home Page of EU GDPR. https://www.eugdpr.org/. Accessed 8 April 2018.
9. Office of the Australian Information Commissioner (OAIC), Privacy Act. https://www.oaic.gov.au/privacy-law/privacy-act/. Accessed 8 April 2018.
10. Office of the Australian Information Commissioner (OAIC), Notifiable data breaches scheme. https://www.oaic.gov.au/privacy-law/privacy-act/notifiable-data-breaches-scheme. Accessed 8 April 2018.
11. Singapore Personal Data Protection Commission, Personal data protection act overview. https://www.pdpc.gov.sg/Legislation-and-Guidelines/Personal-Data-Protection-Act-Overview. Accessed 8 April 2018.
12. Office of the Privacy Commissioner of Canada, The Personal information protection and electronic documents act (PIPEDA). https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/. Accessed 8 April 2018.
13. Japan Personal Information Protection Commission. Act on the Protection of Personal Information Act No. 57 of (2003), 2005.
14. U.S. Department of Health & Human Services (HHS). Breach Portal: notice to the secretary of HHS breach of unsecured protected health information. https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf. Accessed 7 April 2018.
15. Blum, B. I., Orthner, H. F., Implementing health care information systems. In: Implementing Health Care Information Systems. Springer, pp 3–21, 1989.
16. Medical Identity Fraud Alliance (MIFA), The growing threat of medical identity fraud: a call to action, 2013.
17. Czeschik C (2018) Black Market Value of Patient Data. In: Claudia Linnhoff-Popien RS, Michael Zaddach (ed) Digital Marketplaces Unleashed. Springer-Verlag. 10.1007/978-3-662-49275-8_78
18. Dissent, D., 655,000 patient records for sale on the dark net after hacking victims refuse extortion demands. The Daily Dot. https://www.dailydot.com/layer8/655000-patient-records-dark-net/. Accessed 21 April 2018.
19. Bitglass, Healthcare breach report 2018: Security Procedures Thwart Attacks, 2018.
20. Moffit, R. E., Health care data breaches: a changing landscape, 2017.
21. Office of the Australian Information Commissioner (OAIC), Notifiable Data Breaches - Quarterly Statistics Report: January 2018–March 2018., 2018.
22. VERIS Community Database (VCDB) Project. The VERIS Community Database (VCDB). http://veriscommunity.net/vcdb.html, 2018.
23. Verizon. Protected health information data breach report, 2015.
24. Federal Bureau of Investigation (FBI). Table 16 property stolen and recovered. https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/tables/table-16. Accessed 22 April 2018.
25. Palmer, G., A road map for digital forensic research. In: First Digital Forensic Research Workshop, Utica, pp 27–30, 2001.
26. Baryamureeba, V., and Tushabe F., The enhanced digital investigation process model. In, 2004.
27. Carrier, B., Spafford EH An event-based digital forensic investigation framework. In: Digital forensic research workshop, 2004.
28. Cohen, F., Toward a Science of Digital Forensic Evidence Examination. In Advances in Digital Forensics VI. Springer Berlin Heidelberg, pp 17–35, 2010.
29. Kohn, M. D., Eloff, M. M., and Eloff, J. H. P., Integrated digital forensic process model. Comput Secur 38:103–115, 2013. https://doi.org/10.1016/j.cose.2013.05.001.
30. Tan, J., Forensic readiness. Cambridge: @ Stake, 2001, 1–23.
31. Sachowski, J., Implementing Digital Forensic Readiness: From Reactive to Proactive Process. 1st edn. Syngress, 2016.
32. Hunt, R., and Zeadally, S., Network Forensics: An Analysis of Techniques, Tools, and Trends. Computer 45(12):36–43, 2012. https://doi.org/10.1109/MC.2012.252.
33. Khan, S., Gani, A., Wahab, A. W. A., Bagiwa, M. A., Shiraz, M., Khan, S. U., Buyya, R., and Zomaya, A. Y., Cloud log forensics: Foundations, state of the art, and future directions. ACM Computing Surveys (CSUR) 49(1):7, 2016.
34. Jiang, J., Chen, J., Choo, K.-K. R., Liu, C., Liu, K., Yu, M., A Visualization Scheme for Network Forensics Based on Attribute Oriented Induction Based Frequent Item Mining and Hyper Graph. In Digital Forensics and Cyber Crime. Cham: Springer International Publishing, pp 130–143, 2018.

35. MacRae, J., and Franqueira V. N., On Locky Ransomware, Al Capone and Brexit. In: International Conference on Digital Forensics and Cyber Crime, Springer, pp 33–45, 2017.
36. BitCluster, BitCluster. https://www.bit-cluster.com. Accessed 28 April 2018.
37. Elliptic, Elliptic. https://www.elliptic.co/what-we-do/bitcoin-forensics. Accessed 28 April 2018.
38. Vargas, J., Bahnsen, A. C., and Villegas, S., Ingevaldson D Knowing your enemies: Leveraging data analysis to expose phishing patterns against a major US financial institution. In: Electronic Crime Research (eCrime), 2016 APWG Symposium on. IEEE, pp 1–10, 2016.
39. Hamid, I. R. A., Samsudin, N. A., Mustapha, A., and Arbaiy, N., Dynamic Trackback Strategy for Email-Born Phishing Using Maximum Dependency Algorithm (MDA). In Recent Advances on Soft Computing and Data Mining. Cham: Springer International Publishing, pp 263–273, 2017.
40. Gupta, B. B., Tewari, A., Jain, A. K., and Agrawal, D. P., Fighting against phishing attacks: state of the art and future challenges. Neural Computing and Applications 28(12):3629–3654, 2017. https://doi.org/10.1007/s00521-016-2275-y.
41. Jayabalan, M., and Daniel T., Continuous and Transparent Access Control Framework for Electronic Health Records: A Preliminary Study. In: International Conference on Information Technology on Information Technology, Information Systems, and Electrical Engineering (ICITISEE 2017), 2017.
42. Kose, I., Gokturk, M., and Kilic, K., An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Applied Soft Computing 36:283–299, 2015.
43. King, J., Stallings, J., Riaz, M., and Williams, L., To log, or not to log: using heuristics to identify mandatory log events – a controlled experiment. Empirical Software Engineering 22(5):2684–2717, 2017. https://doi.org/10.1007/s10664-016-9449-1.
44. Protenus, Getting Schooled on Patient Privacy Analytics. https://blog.protenus.com/getting-schooled-on-patient-privacy-analytics. Accessed 3 May 2018.
45. Cognetyx, The inconvenient truth about patient data security and privacy in healthcare. https://www.cognetyx.com/the-inconvenient-truth-about-patient-data-security-and-privacy-in-healthcare-cognetyxs-new-ambient-cognitive-cyber-surveillance-solution-is-addressing-this-proble/. Accessed 3 May 2018.
46. Zawoad, S., Dutta, A. K., and Hasan R., SecLaaS: secure logging-as-a-service for cloud forensics. In: Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security. ACM, pp 219–230, 2013.
47. Nanda, S., Hansen, R. A., Forensics as a Service: Three-tier Architecture for Cloud based Forensic Analysis. In: Parallel and Distributed Computing (ISPDC), 2016 15th International Symposium on, 2016. IEEE, pp 178–183
48. Zawoad, S., and Hasan, R., Faiot: Towards building a forensics aware eco system for the internet of things. In: Services Computing (SCC), 2015 IEEE International Conference on. IEEE, pp 279–284, 2015.
49. Raju, B. K., Moharil, B., Geethakumari G FaaSeC: Enabling Forensics-as-a-Service for Cloud Computing Systems. In: 2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC). pp 220–227, 2016.
50. Yang, H., and Garibaldi, J. M., Automatic detection of protected health information from clinic narratives. Journal of Biomedical Informatics 58:S30–S38, 2015. https://doi.org/10.1016/j.jbi.2015.06.015.
51. Frühwirt, P., Kieseberg, P., Schrittwieser, S., Huber, M., and Weippl, E., InnoDB database forensics: Enhanced reconstruction of data manipulation queries from redo logs. Information Security Technical Report 17(4):227–238, 2013.