了解医院网络存在的安全性和性能：一项测量研究

Mohammed Alkinoon计算机科学系 中佛罗里达大学奥兰多，佛罗里达州，美国
malkinoon@Knights.ucf.edu

Abdulrahman Alabduljabbar计算机科学系 中佛罗里达大学奥兰多，佛罗里达州，美国 jabbar@Knights.ucf.edu

Hattan Althebeiti计算机科学系 中佛罗里达大学奥兰多，佛罗里达州，美国
hattan@Knights.ucf.edu

Rhongho Jang计算机科学系 韦恩州立大学底特律，密歇根州，美国
r.jang@wayne.edu

DaeHun Nyang计算机科学系 梨花女子大学首尔，韩国
nyang@ewha.ac.kr

David Mohaisen计算机科学系 中佛罗里达大学美国佛罗里达州奥兰多
mohaisen@ucf.edu

arXiv:2304.13278v1 [cs.CR] 26 Apr 2023

摘要—**近期，医疗保健病历从纸质形式转变为数字化和联网形式，引发了人们对患者安全和在线隐私的担忧。例如，由于缺乏适当的安全和隐私机制，患者姓名、地址和社会保障号码等敏感个人信息可能成为攻击目标。**

本研究选取了4,774家医院，并将其分为政府医院、非营利医院和私立医院，通过相关性分析，对医院网站进行了首次基于测量的分析，并将研究结果与数据泄露事件联系起来。我们分别研究这三类医院的安全属性，并将其进行总体比较和对比，考察域名、内容和SSL证书层面的特征。我们发现，各类医院在域名注册商的使用、顶级域名分布和域名创建分布，以及内容类型和HTTP请求特征方面，都具有独特的特征。在安全方面，与网站总体情况一致，仅有1%的政府医院使用了DNSSEC，而私立医院的比例为6%。令人担忧的是，我们发现25%的医院使用明文HTTP，而整个网络中的比例为20%。同样令人担忧的是，我们发现根据医院类型的不同，有8%至84%的医院含有某些恶意内容，这些内容大多归因于缺乏维护。

最后，我们将研究结果与 414 起已确认且经过人工审查的医院数据泄露事件进行相关性分析。我们的研究还发现，医院网站分析中所强调的安全属性，是衡量其遭受入侵可能性的重要指标。我们的分析是理解患者在线隐私的第一步，揭示了许多医院网站缺乏基本安全性，并为未来研究开辟了多个方向。

索引词—**医疗保健、网络安全、SSL证书、测量**

I. 引言

电子健康记录（EHR）是纵向电子患者健康信息记录，其中包括患者人口统计信息、病程记录、健康问题、药物、生命体征、病史、免疫接种记录、实验室数据等。 [29]。电子健康记录的采用提高了患者和医疗保健提供商获取医疗保健信息的便利性，从而带来了更高质量的患者护理以及医院之间更高效的协调，然而，尽管这些益处显著，向电子健康记录的转型也引发了隐私和安全方面的担忧，尤其是在电子健康记录数据可通过网站系统检索时。电子健康记录集中存储敏感患者数据，这可能使其成为网络犯罪分子窃取或利用这些信息的主要目标。此外，电子健康记录可由多个医疗保健提供商访问和共享，从而增加数据泄露事件和未经授权的访问风险。此外，电子健康记录系统的实施可能引入新的漏洞，而这些漏洞可能被网络犯罪分子利用。

例如，医疗保健领域的漏洞和软件漏洞利用已成为定向网络攻击的主要目标 [24]，，并可能造成灾难性的数据泄露。鉴于私人医疗数据的敏感性，未经授权且非法披露此类信息可能造成灾难性后果 [4], [36]。

了解医疗保健数据泄露的影响至关重要，相关文献中的研究将此类事件分为内部数据泄露和外部数据泄露。内部数据泄露通常由人为错误造成，尤其是医疗保健从业人员的失误。相比之下，外部数据泄露更为严重，是由未经授权的第三方通过黑客攻击基于网络的用户系统和面向医疗服务提供者的系统，盗窃私人健康记录所致 [47]。网络犯罪分子通常会实施此类事件，其影响仍是一个悬而未决的问题，相关成本也尚无准确评估。例如，参与外部数据泄露的攻击者可能意在窃取敏感记录并索要赎金，或在暗网上以每位患者数百美元的价格出售这些记录 [38], [45]。

鉴于了解医疗保健领域数据泄露事件的重要性，以及网络技术在促成攻击面中相当大一部分方面所发挥的作用，本研究致力于分析三类医院之间的共性与差异：公立医院、非营利医院

和私立医院。具体而言，我们分析了网站和患者门户网站的安全配置及常见隐私实践。我们注意到，攻破患者门户网站会使攻击者获得有关患者记录的敏感信息，包括诊断、治疗记录、就医记录和未来预约，以及个人信息。据我们所知，本研究是该方向上的首项工作，将医院实际存在的潜在攻击和数据泄露事件与网站安全和隐私配置联系起来。

我们的分析基于总计4,774个医院网站，这些网站被划分为三大医院类别：政府公立医院、非营利医院和私营（私人）医院。在测量评估中，我们同时开展域层面和内容层面分析，以了解网站属性之间的相似之处和差异。

我们的分析涵盖多个方面，通过考察和比较网站的域名SSL证书、创建日期、HTTP 请求、页面大小、内容类型、平均加载时间以及与恶意活动的关联，研究一系列特征。本分析所探讨的特征都较为轻量化，不需要对内容进行深入分析，而是侧重于元属性，因此我们的分析技术更适用于大规模测量。我们进一步通过探究这些网站与恶意行为的关联，调查了其安全属性，包括评估这些网站基于域名和基于内容的恶意行为，以及相关趋势和特征。

为了解这些特征的影响，我们进一步研究了它们与美国卫生与公众服务部民权办公室（OCR） [21]提供的、经过人工审查的近期披露数据泄露事件数据集之间的相关性。借助有关网站及其相关数据泄露事件的信息，我们提取了遭受这些数据泄露攻击的医院网站之间的共同点，以对其进行建模和特征描述。我们认为，这项工作是理解可能导致数据泄露的网站属性的第一步，并将推动未来开展漏洞预测和检测研究。

研究问题。我们旨在回答一个总括性问题：**在内容、性能和安全方面，就所研究的特征而言，不同类别的医院网站之间是否存在差异?** 我们将这一问题分解为以下可量化的问题。

- **研究问题1（RQ1）**。不同医院在域名、内容和传输层特征的使用方面存在多大差异？我们通过比较探索域名层面的特征（第IV-A节），包括域名注册商、顶级域分布、域名创建分布，以及内容层面的特征，包括内容类型和HTTP请求特征（第IV-B节），来回答这一问题。
- **研究问题2**。医院网站的主要安全特征是什么？这些特征在不同类型的网站之间有何差异？我们通过探究医院的DNSSEC部署情况及其对比（第IV-C节）、SSL证书特征和属性（第IV-D节）、针对各种检测引擎的恶意性特征（第IV-E节）以及与数据泄露事件的关联（第V节）来回答这一研究问题。

回答这一研究问题的方法包括探究医院的DNSSEC部署情况及其对比（第IV-C节）、SSL证书特征和属性（第IV-D节）、针对各种检测引擎的恶意性特征（第IV-E节）以及与数据泄露事件的关联（第V节）。

贡献与发现。鉴于目前缺乏系统性研究来理解医院在网络上的存在及其相关的安全和性能属性，本研究着手从一系列属性出发，探索这些医院的网络存在。此外，通过比较分析，本研究揭示了美国政府、非营利和私营医院之间的差异与相似之处。我们的分析涵盖三个维度：安全、内容和域名。为此，我们的贡献如下：

1) 域名级分析（§IV-A）。域名是网站的入口，对于理解这些域名及其背后实体的各种粗粒度且易于获取的特征至关重要。为此，我们开展了域名注册商和顶级域名分析，以揭示并对比网站/医院的特征。我们发现了网站与注册商之间选择的关联性、顶级域名选择以及域名创建日期方面的规律。在其他有趣的发现中，我们观察到，政府医院和非营利医院的网站数量近年来一直在下降，这暗示着私营医疗保健系统正在积极扩张。

2) 内容级分析（§IV-B）。 我们检查了医院网站的内容，以深入了解其所采用的内容类型、规模和安全特性。通过这项分析，我们发现这些网站在采用不同内容类型（如图像和脚本）方面存在差异，这会影响各种性能指标，包括加载时间。更有趣且颇令人意外的是，我们发现 6% 的私立医院使用域名系统安全扩展（DNSSEC），而政府医院和非营利医院中使用该扩展的比例不到 1%。

3) SSL证书级分析（§IV-D）。证书对于网站身份验证至关重要，并有助于在传输层实现网页内容加密，从而提供安全的应用媒介。我们研究了医院网站的HTTPS协议配置、相关SSL功能以及SSL证书有效性。我们根据证书颁发机构归属、所使用的算法和证书有效性对网站进行分类。在其他有趣的发现中，我们的调查揭示，超过25.25%的医院网站仍在使用不安全的HTTP协议。此外，在采用HTTPS的网站中，多达23%的SSL证书无效。

4) 恶意活动分析（§IV-E）。由于其复杂性，可能会出现意料之外的弱点

这是由于医院网站中第三方代码的聚集以及各种共享基础设施的使用所导致的。为了解这一维度，我们利用各种扫描工具，从域名和内容级别探索这些网站的恶意活动。在其他有趣的发现中，我们发现大量网站包含恶意内容，并与恶意行为相关联。

5) **数据泄露分析（§V）**。数据泄露是不可避免的。但从相关性的角度来看，是什么使网站容易发生数据泄露？我们通过将医院与近期报告的数据泄露事件进行关联和分析来探讨这一问题，并发现非营利医院更有可能涉及数据泄露事件。我们展示了导致数据泄露事件的最重要属性，包括托管恶意代码、大量图像等。

II. 相关工作

关于分析医疗保健提供商安全配置的研究较为有限。为了将我们的工作置于具体背景中，我们回顾了网站内容、安全分析以及医疗保健数据泄露分析方面的既有研究。

网站分析。过去几年中，在线服务和 Web 应用程序的开发与使用大幅增加。伴随这一增长，人们日益关注这些在线服务和应用程序的隐私与安全，例如，不同组件可能遭到入侵，使其用户面临风险。Chung等[12] 首次对在线公钥基础设施（PKI）中的错误证书进行了深入分析，表明大多数 PKI证书都是无效的。同一项研究审查了无效 SSL证书的来源，并总结指出，大多数无效证书由终端用户设备生成，同时还会定期更新新的自签名证书。研究人员已针对网站风险和漏洞分析对 SSL证书进行了研究 [3], [6], [8]–[10], [12], [14], [25]–[28], [31], [48]。例如，Meyer 等 [30] 检验了 SSL证书的内容和信息，以区分网络钓鱼网站与良性网站。Alabduljabbar 等 [1], [2] 探讨了免费与高级内容网站基于 SSL证书的结构差异，并指出免费网站的证书中有 35.85% 存在严重安全问题，其中 17% 无效、7% 已过期，12% 存在域名不匹配。

数据泄露分析。近期有多项研究调查了医疗保健行业中的数据泄露事件 [4], [5], [36]–[42]。例如，Seh 等 [36]对HIPPA数据泄露报告进行了全面分析。其研究指出，黑客攻击事件、未经授权的访问（内部）、盗窃或丢失，以及不必要数据的不当处置，是受保护的医疗保健信息泄露的主要类型。此外，作者对数据应用了简单移动平均（SMA）和简单指数平滑（SES）时间序列方法，以确定医疗保健数据泄露事件及其对医疗保健行业造成的成本。Choi 等 [11] 估计了数据泄露与医院广告支出之间的联系，研究了泄露事件发生后两年的时期，发现医院的广告支出大幅增加。Siddartha 等 [37]发现，医疗保健行业受到攻击主要有两个原因：它是有价值数据的丰富来源，而且防御能力薄弱。

Siddartha 和 Ravikumar [37] 指出，医疗保健行业采用的安全技术缺少数据分析改进，例如数据格式保留、数据大小保留以及其他因素。Luis 等人 [42] 利用两年收集的数据定义 DNS 查询和 TLS/SSL 连接，以识别医院环境中遇到的危险，同时不影响网络的正常运行。另一项研究《2022 年数据泄露调查报告》（DBIR） [19],调查了医疗保健行业及其他行业中的数据泄露事件。根据该报告，医疗保健行业在 2022 年发生了 849 起事件，其中 571 起确认涉及数据披露。报告总结了多项发现，并确定外部行为者导致了 61% 的数据泄露事件，而 39% 的数据泄露事件涉及内部行为者。此外，根据同一份报告，经济利益是攻击者最主要的动机，占 95%；其次是间谍活动，占 4%。

本文工作。 我们聚焦于医疗保健机构的在线存在，通过对三种主要医院类型进行对比，深入研究医院网站和患者门户的安全配置与功能，以及这些特征与已报告数据泄露事件之间的相关性。

III. 数据集、流程与研究问题

在本研究中，我们使用了从国土安全部基础设施基金会级数据（HIFLD） [33]获取的美国医院真实数据集。该数据集包含分布于美国50个州、华盛顿特区以及美国领地波多黎各、关岛、美属萨摩亚、北马里亚纳群岛、帕劳和维尔京群岛的医院。

我们将数据集中的医院分为三类：政府医院、非营利医院和私立医院。政府医院包括联邦、地区、地方和州级医院。根据美国国税局（IRS） [22]的定义，非营利医院是指由慈善机构运营的医院。私立医院则是由个人、合伙企业或在大多数情况下由公司[39]所有并以营利为目的运营的医院。总计包括1,034家政府医院、2,187家非营利医院和1,550家私立医院。

网站预处理和爬取。我们的研究采用网站爬取方法，系统地从网站收集信息，并将其用于进一步分析。图 1 展示了我们遵循的总体流程：首先通过网站枚举确定需要爬取的网站。接下来，我们进行网站预处理，包括移除不相关的网站（即包含与研究范围无关内容的网站）以及无法正常运行的网站。最后，我们进行特征提取

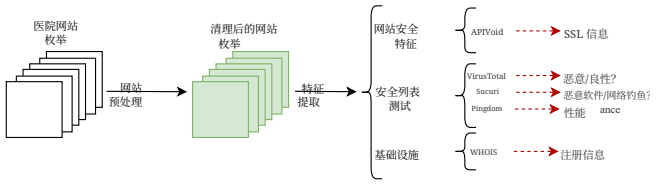

图 1：我们在网站爬取和数据增强过程中所采用的流程，涵盖 SSL、恶意性、漏洞、性能和域名属性等不同维度。

以提取我们用于进一步分析的网站属性。

为开展我们的分析，我们引入了数据增强步骤。在抓取与每家医院相关的网站后，我们使用其他属性丰富了所收集的数据，包括以下内容。

- **SSL 属性**。为提取 SSL证书信息，例如域名不匹配、SSL 到期日期和证书有效性，我们使用了 APIVoid [7], ，这是一个提供网络威胁分析和检测功能的框架。
- **恶意性属性**。 为分析医院在线内容的恶意性，我们采用了 VirusTotal API [43],，这是一项汇总了 70 多个检测引擎结果的在线服务。
- **漏洞属性**。 为检查每个网站是否存在漏洞并识别任何恶意代码，我们使用了 Sucuri [40], ，这项服务会根据多个已知恶意软件和黑名单列表对网站进行检测。
- **性能属性**。 为评估网站的性能和可用性，我们使用了 Pingdom [34],，这是一款面向网站的全球监控软件。
- **域名属性**。 为确定每个网站的所有权和 DNSSEC 信息，我们使用了 WHOIS [46],，这是一个互联网资源所有权数据库，并查询了每个网站的创建日期。

总体而言，网站爬取和数据增强的步骤使我们能够提取两类信息： ① 网站内容数据，例如图像、字体、HTML、CSS、脚本、XHR 和重定向；以及 ② 性能指标，例如页面大小、加载时间和请求数。

IV. 网站分析

为了了解不同类别医院的在线存在和结构差异，我们对其网站开展了一系列分析：域名级、内容级、基于 SSL 证书的分析，以及基于恶意活动的分析。下面我们将介绍这些分析的结果。

A. 域名级分析

域名是任何组织的一项关键资产，在品牌建设中发挥着重要作用，为其提供强大的在线存在和搜索引擎优化（SEO）优势。因此，为了启动我们的网站分析，我们首先考察域名

表 I：医院的网站 URL 与域名注册商组织相对应。请注意，*NetworkSolutions* 和*GoDaddy* 在列表中最为突出，相关网址占比最高达 67.65%。

域名注册商	政府		非营利		私营	
	#	%	#	%	#	%
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
其他	188	20.06	325	15.27	252	16.73

详细信息，包括域名注册商、顶级域名和域名创建日期。

域名注册商。域名注册商是负责管理互联网域名注册以及为这些域名分配 IP 地址的组织 [15], ，某些注册商在安全措施和政策方面往往较为宽松 [17], [18], [41]。分析域名注册商对于评估网站的整体安全性和可靠性至关重要。这是因为注册商提供了有关组织在线存在和安全措施的重要信息。注册商的安全措施和政策水平各不相同，这可能会影响网站的安全性和可信度。考察域名注册商可以深入了解组织的安全策略，并帮助评估与该域名相关的潜在风险。此外，了解注册商还能揭示组织的在线策略和网络托管安排，从而进一步为分析网站的结构和性能提供信息。表 I 展示了不同类型医院所使用的域名注册商组织的分布情况。值得注意的是，*NetworkSolutions* 和*GoDaddy* 是最主要的注册商，所占域名比例最高达 67.65%。此外，我们发现，尽管*Mark Monitor* 和*eNomLLC* 在政府和非营利网站中相对少见，但它们服务于 20.92% 的私营医院网站。

顶级域名。 顶级域（TLD）是域名的“后缀”。除品牌标识外，顶级域名在域名系统（DNS）查询中也发挥着重要作用，有助于对域名进行分类并说明其用途。顶级域名的示例包括“.com”、“.org”、“.net”和“.edu”。顶级域名提供有关网站用途、组织类型或地理位置的信息。例如，“.com”通常用于商业网站，“.org”用于非营利组织，“.edu”用于教育机构，而“.gov”用于政府网站。了解顶级域名有助于洞悉网站的目标受众，以及其提供的内容或服务类型。此外，分析顶级域名还有助于识别与网站相关的潜在安全风险。例如，已知某些特定国家/地区的顶级域名与恶意活动有关，使用此类顶级域名的网站可能更容易构成安全威胁。此外，一些网站可能会使用拼写错误或与知名顶级域名相似的顶级域名，试图欺骗访问者并实施欺诈活动。通过分析顶级域名，我们可以识别此类风险并采取适当措施加以缓解。因此，分析顶级域名是评估网站整体安全性和可靠性的关键步骤。互联网名称与数字地址分配机构（ICANN）负责管理互联网上的所有顶级域名，并将这些顶级域名的管理职责委托给各个组织

表 II：顶级　-级域名比较：政府、非营利和私立医院之间的比较。

Type	.org	.com	.gov	.net	.mil	.edu	.us
政府	48.45%	36.56%	4.35%	4.35%	2.90%	2.42%	0.48%
非营利	67.49%	28.81%	0.05%	1.92%	0.00%	1.37%	0.27%
私营	7.81%	87.35%	0.00%	3.48%	0.00%	0.13%	0.45%

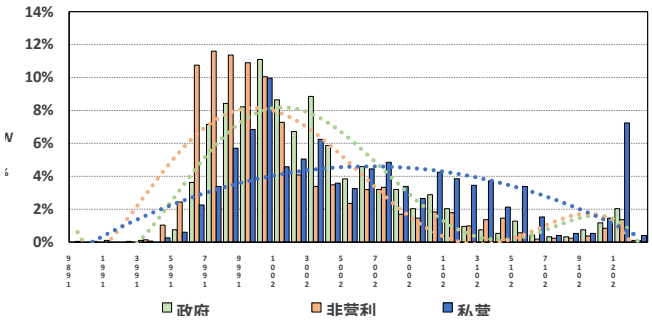

图 2：三类医院之间的域名创建日期时间分析。点线表示移动平均值。

表 II 展示了数据集中不同医院类别的顶级域名对比。我们发现，政府医院（48.45%）和非营利医院（67.49%）最常使用“*.org*”，而私立医院中这一顶级域名相对少见（仅占 7.81%）。另一方面，与政府医院（36.56%）和非营利医院（28.81%）相比，私立医院“*.com*”占主导地位（87.35%）。我们还注意到，总体而言，92.15% 的医院网站使用“*.com*”或“*.org*”。出乎意料的是，与普遍认知不同，我们发现只有 4.35% 的政府网站使用“*.gov*”顶级域名。

域名创建。域名创建日期是指特定域名最初向域名注册商注册的日期。这是一项至关重要的信息，因为它可以帮助了解网站的历史和在线存在。注册时间较长的网站可能比新网站更加成熟，并且具有更显著的在线存在。分析域名创建日期有助于识别潜在的欺诈性或恶意网站。例如，最近注册的网站更有可能涉及网络钓鱼诈骗或欺诈计划。因此，分析

域名创建日期可能是评估

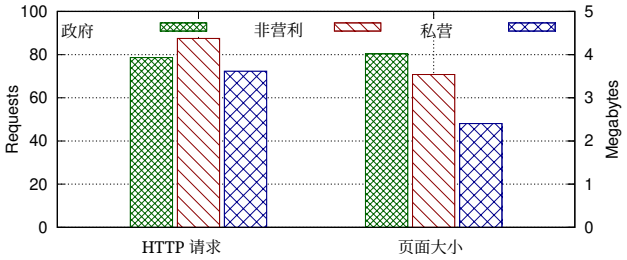

图3：请求与响应大小对比。

网站整体可信度和潜在安全风险的重要步骤。图2展示了不同类别医院的域名创建日期。如图所示，政府医院和非营利医院的网站出现在相近时期（1995年至2009年），2009年后呈下降趋势。然而，私立医院网站的出现较为稳定，其数量在2021年快速增加。

> **要点：（研究问题1。）**近年来，政府医院和非营利医院的网站数量一直在减少，而私立医院的数量则显著增长，尤其是在2021年。此外，尽管政府医院获得政府支持，但大多数政府医院并未使用“*.org*”顶级域名。

B. 内容层面分析

为了分析不同类别医院之间的内容差异，我们使用 Pingdom [34],抓取医院网站，获取 HTTP请求信息以及所有相关文件：脚本、图像、HTML 和 CSS 文件。

内容类型。在结构层面上，表III展示了三类医院中文件类型的分布。XHR 是一种 API，可作为对象与服务器交互，并在服务器和网页浏览器之间交换数据。含有“*XHR*”的内容在政府医院和私营医院网站中较为突出，合计占10.08%。总体而言，除“图像”和“脚本”外，各类别的文件类型分布相似。“脚本”内容被定义为一种用于为网站添加动态功能的计算机程序，在私立医院中使用最多，占33.12%。另一方面，“重定向”内容是一种网站功能，可将用户从当前 URL 引导至另一台服务器；该功能在政府医院（17.37%）和非营利医院（16.11%）中使用较多，在私立医院中使用最少（13.76%）。

HTTP 请求。HTTP（超文本传输协议）请求是一条从客户端（例如

表 I　　II：政府、非营利和私立医院之间的内容类型比较　　.

类别	CSS	Font	HTML	图像	重定向	脚本	XHR
政府	6.05%	8.18%	2.71%	40.55%	17.37%	20.34%	4.79%
非营利	5.30%	7.80%	2.86%	38.01%	16.11%	27.01%	2.91%
私营	6.63%	9.61%	3.53%	28.05%	13.76%	33.12%	5.29%

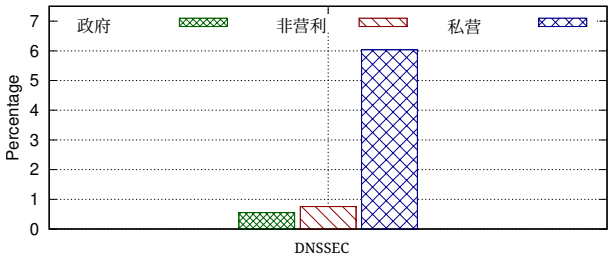

图 4：域名系统安全扩展。

网页浏览器）发送到服务器的消息，用于请求特定资源或执行特定操作。请求通常包含一个 URL（统一资源定位符），用于指定所请求的资源或操作，以及服务器处理该请求所需的其他信息，例如请求头和 Cookie。

通过分析HTTP请求，企业和组织可以深入了解其网站的性能，例如页面加载时间、每个页面的请求数以及所请求文件的大小。这些信息有助于识别可以进一步优化网站性能的方面。

图 3 展示了所分析的三类医院中每个网站的平均 HTTP 请求数。我们发现，大多数网站每次访问会产生 65 到 90 个 HTTP 请求，其中非营利医院的数量最高。尽管 HTTP 请求数相近，私立医院的平均页面大小比政府医院小 45%。进一步分析发现，与政府医院和非营利医院相比，私立医院网站中图像所占比例最低，这有助于解释这一趋势。

要点：(RQ1。) 从结构上看，各类医院的内容类型分布相似，唯有 “图像”和“脚本”内容类型例外。尽管各类医院的 HTTP 请求数相近，私立医院的平均页面大小仍比政府医院小 45%。

C. DNSSEC 的普及情况

DNSSEC（域名系统安全扩展）是一种安全协议，用于防范针对域名系统（DNS）的某些攻击，以保护其完整性和来源真实性。

DNSSEC 通过向 DNS 记录添加数字签名来工作，DNS 解析器可以验证这些签名，以确保记录未遭篡改或伪造。这有助于防止攻击者将用户重定向到虚假网站，

表 IV：医院网站对应的证书颁发机构。请注意，*Let's Encrypt* 是最主要的证书颁发机构，关联网址最多占 25.21%。

颁发机构	Paid	政府		非营利		私营	
		#	%	#	%	#	%
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
其他	-	148	14.31	385	17.60	149	9.61
未找到 SSL 证书	-	250	24.17	506	23.13	441	28.45

拦截电子邮件，或实施其他依赖篡改 DNS 记录的攻击。DNSSEC 旨在提供端到端安全，这意味着可以从根服务器一直到各个域名服务器验证 DNS 记录的完整性。大多数现代网页浏览器和操作系统都支持 DNSSEC，而且许多顶级域名已经采用 DNSSEC，为用户提供额外一层安全保护。

为了评估 DNSSEC 在这一相当关键的网站类别中的普及程度，我们调查了 DNSSEC 部署情况，发现如图 4 所示，6 % 的私营医院网站启用了 DNSSEC，这一比例显著高于政府医院网站（0.55%）和非营利医院网站（0.75%）。

要点：(研究问题2。) 我们注意到，6% 的私立医院使用了 DNSSEC，而政府和非营利医院中这一比例不到 1%。虽然 1% 被视为较低比例，但这与 .com 和 .org 普通网站的 DNSSEC 部署情况一致（即 0.75%–1%） [13],；相比之下，私营医院组的比例是这一水平的 6 至 8 倍，凸显了 DNSSEC 在该类别中的普及程度更高。

D. HTTPS 和 *SSL* 证书分析

HTTP 协议负责将网站内容从网页服务器传输到终端浏览器。然而，该协议并不安全，会使内容面临未经授权的访问。因此，大多数网站已转而在安全套接层（SSL）之上使用 HTTPS，即 HTTP 的安全版本。SSL 除了其他功能外，还实现了加密机制，用于保护网页服务器与终端浏览器之间传输的内容。医疗保健网站通常要求用户输入敏感个人信息，例如健康相关数据、保险信息和病史。HTTPS 有助于保护这些信息免遭

表 V：SSL 签名算法比较。

	政府		非营利		私营	
算法	#	%	#	%	#	%
SHA256与RSA	751	95.00	1,535	91.31	891	80.34
SHA256 与 ECDSA	26	3.30	108	6.42	179	16.14
SHA384 与 ECDSA	10	1.27	35	2.08	37	3.34
SHA1 与 RSA	1	0.13	2	0.12	1	0.09
SHA384 与 RSA	1	0.13	-	-	1	0.09
SHA512 与 RSA	-	-	1	0.06	-	-

攻击者拦截，从而确保患者数据的机密性。

为此，接下来我们将研究与 SSL 相关的配置：证书颁发机构、签名算法和证书验证。在所研究的医院中，我们注意到，仍有 25.25% 的网站使用 HTTP，而整个网络中的这一比例仅约为 20% [44]。尽管网站数量相差 5%，但考虑到医院相关数据的性质，这一比例仍高得令人担忧（即至少每四家医院中就有一家使用不安全协议）。

证书颁发机构。 证书颁发机构（CA）是通过签发电子文档，将实体与密码学密钥绑定，从而验证包括网站、电子邮件地址等实体身份的组织。在调查证书颁发机构组织（即证书的签发者）时，表 IV 显示，大多数医院都在使用免费的*Let's Encrypt* 服务 [23]（即免费的SSL证书），其中政府医院组的比例最高，达到 31.04%。我们还注意到，医院网站广泛使用免费的 SSL 证书。令人意外的是，我们发现政府、非营利和私营网站中分别有 24.17%、23.13% 和 28.45% 未使用 SSL 证书。

算法。 表 V 展示了政府、私营和非营利网站所使用的 SSL 签名算法。如表所示，*SHA256与RSA* 是使用最广泛的方案，在政府、非营利和私营医院网站中的占比分别为 95.00% 、 91.31% 和 80.34%。这主要是因为医院倾向于使用服务提供商所采用的传统通用算法。另一方面，我们注意到，较少医院网站使用*SHA256与ECDSA*（椭圆曲线数字签名算法）算法；在相同安全级别下，与使用更长密钥的 RSA 相比，该算法使用更短的密钥[2]。在政府、非营利和私营医院网站中，其占比分别为 3.30%、6.42% 和 16.14%。我们注意到，ECDSA 是一种更新且效率更高的算法，主要用于较新的网站 [2]。不过，ECDSA 更

根据近期研究，在后量子对手面前，其相较于较旧的 RSA 更容易受到攻击 [35]。

证书有效性。我们进一步调查了 SSL证书的有效性及潜在问题。下文将讨论与 SSL证书失效相关的问题（见图 5）。

*SSL*域名不匹配。域名不匹配可能表明存在网站冒充或不一致的网站迁移，而这两种情况都凸显出安全实践不够严格。我们发现，18.45%的私立医院

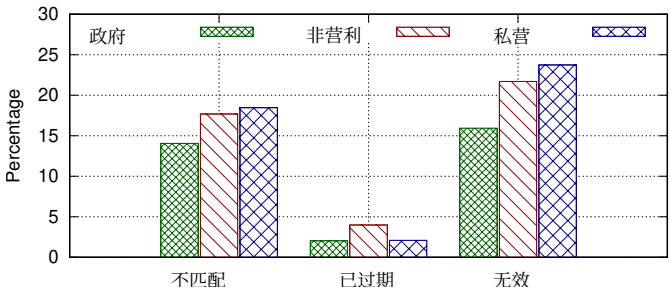

图 5：政府、非营利和私营医院网站的 SSL 有效性比较。

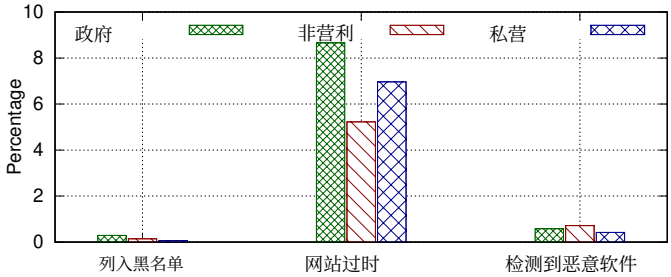

图 6：政府、非营利和私立医院网站恶意性的比较。

的SSL证书存在域名不匹配，相比之下，政府医院为 14.05%，非营利医院为17.67%。尽管比例各不相同，但所有医院网站的域名不匹配率都令人担忧。

*SSL*证书过期。 我们的分析发现，约有 3.97% 的非营利医院拥有过期证书，而政府医院和私立医院的比例分别为 2.20% 和 2.09%。与我们此前对过时网站的分析类似，这可能会带来潜在的用户信息和数据隐私风险。

*SSL*证书无效。 SSL证书无效意味着证书中的某些字段不正确。令人惊讶的是，所有医院类别的无效SSL证书比例都令人担忧：政府医院为 15.94%，非营利医院为 21.70 %，私营医院网站为 23.72%。

要点：(研究问题2。) 超过25%的医院网站使用明文HTTP协议，这一比例明显高于普通网站中的 ≈20% [44]。在使用HTTPS的网站中，88.77%使用了*SHA256*与*RSA*。在 ≈75%拥有SSL证书的医院中，我们发现20.45%的SSL证书无效，另有16.72%的证书存在域名不匹配；在这两种情况下，问题主要出现在私立医院中。

E. 恶意活动分析

除了结构差异和 SSL证书分析之外，我们还研究了与医院网站相关的恶意活动。本研究考察的恶意活动包括提供恶意或网络钓鱼内容，或网站资源与恶意攻击相关联。

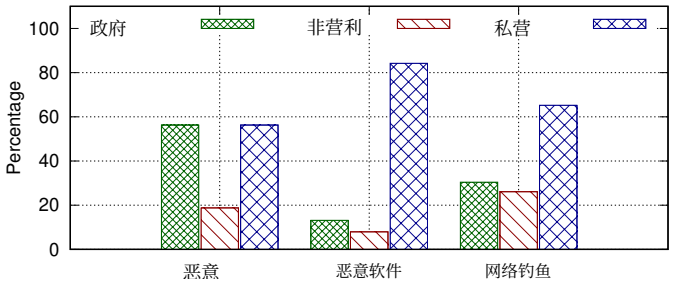

图7：政府、非营利和私立医院的潜在恶意性。

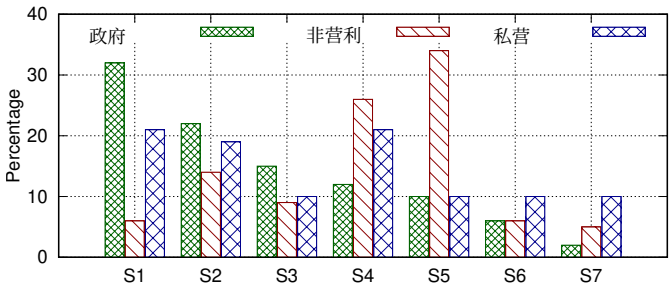

图 8：比较政府、非营利和私立医院的数据泄露面。S1=纸张/胶片，S2=网络服务器，S3=电子病历，S4=其他，S5=电子邮件，S6=笔记本电脑，S7=台式计算机。

基于域名的恶意活动。 我们借助 Sucuri[40]探索基于域名的恶意活动。图 6 显示，尽管只有少数医院的网址被列入黑名单或被标记为恶意软件，但政府医院网站、非营利医院网站和私营医院网站中分别约有 8.66%、5.21% 和 6.96% 已过时，这引发了人们对数据泄露的担忧。

基于内容的恶意活动。接下来，我们使用 VirusTotal API [43] 分析了网站内容。图7显示，在三类医院中，84.21%的私立医院含有恶意软件，而政府医院和非营利医院中这一比例分别仅为13.15%和7.89%。此外，我们观察到，65.21%的私营医院网站疑似存在类似网络钓鱼的行为。需要指出的是，政府医院（30.43%）和非营利医院（26.08%）的比例明显低于私营医院，但仍然相当高。此外，我们观察到，56.25%的政府医院和私营医院以及18.75%的非营利医院与恶意活动有关联。

要点：（研究问题2。）大多数医院具有恶意性特征（域名或内容），并且由于缺乏维护（即网站过时），许多医院容易发生数据泄露。在汇总的医院网站中，8.66%的网站已经过时。此外，有相当一部分网站含有恶意内容，并与网络钓鱼和恶意行为有关联。

V. 数据泄露分析

分析数据泄露事件有助于理解网络存在安全性与事件之间的相关性。根据《健康保险携带和责任法案》（HIPPA），数据泄露可定义为违反《隐私规则》的不可许可使用或披露行为，该行为损害了受保护的健康信息的安全性或隐私性 [20]。在医疗保健领域，数据泄露事件具有毁灭性，因为它们会同时对患者和医疗保健机构造成损害。近期研究表明，医疗保健是网络犯罪分子最常攻击的行业，因为攻击者意图将患者记录出售到暗网，以获取经济利益。为了调查医院中的历史数据泄露事件，我们从美国卫生与公众服务部民权办公室（OCR）获取了医疗保健数据泄露数据集。OCR 门户列出了所有影响 500 名或以上患者的未受保护健康信息数据泄露事件。OCR 门户将数据泄露事件分为两类：（i）过去 24 个月内报告且目前仍在调查中的事件；（ii）已完成的数据泄露事件，即已解决且发生时间超过 24 个月的数据泄露报告。我们注意到，由于医院名称与数据泄露事件中所列实体的组织方式并不一致（例如混合或截断），因此很难将两者对应起来。为解决这一问题，我们首先使用医院名称作为锚点，然后利用自然语言工具包（NLTK） [32]对医院名称进行标点符号移除、大小写规范化、停用词移除、词形还原以及词干提取。对于数据泄露数据集中的实体名称，我们也采用了类似的处理流程。最后，将仅包含两个常用词的医院名称和实体筛选出来，以便进行人工分析和审查。总体而言，我们人工检查了 1,253 起事件，最终准确标注了 414 起数据泄露事件。

下文将分析数据泄露事件，并深入了解发生数据泄露的医院所具有的常见在线属性。

相关医院和个人。在数据集中记录的 414 起数据泄露事件中，49 起与政府医院相关，156 起与非营利医院相关，34 起与私立医院相关。值得一提的是，一家医院可能涉及多起事件。我们的分析表明，受影响个人的平均数量总体为 58,750 人，其中政府医院为 60,458 人，非营利医院为 64,977 人，私立医院为 50,815 人。值得注意的是，私立医院涉及的事件数量和受影响个人数量最少。

数据泄露面。 如图 8 所示，“纸张/胶片”是政府医院（32%）和私立医院（21%）最常被攻击的对象，而非营利医院仅为 6%。其次，“电子邮件”主要针对非营利医院（34%），而政府医院和私立医院并未成为其主要攻击目标（1%）。总体而言，“网络服务器”是仅次于“纸张/胶片”的第二常见攻击目标，这表明医院在线安全的重要性。

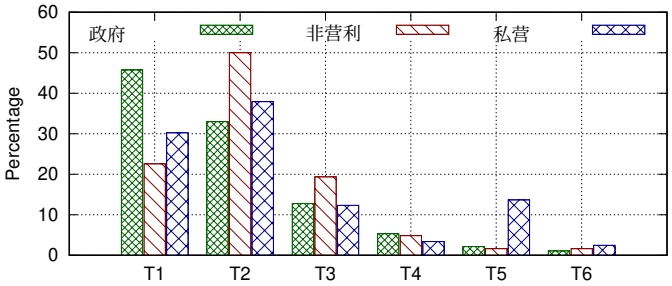

图 9：比较政府、非营利和私立医院的数据泄露类型。T1=黑客攻击/信息技术、T2=未经授权的访问、T3=盗窃、T4=丢失、T5=不当处置、T6=其他。

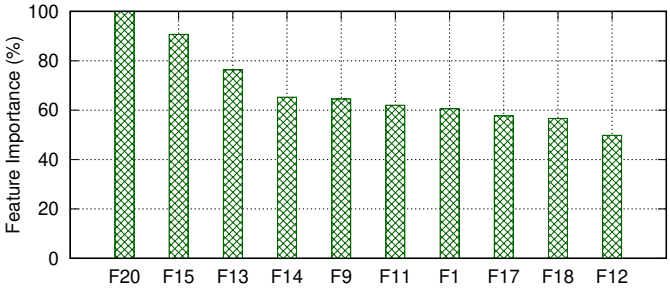

图 10：用于区分与数据泄露事件相关的医院网站的域名级和内容级属性重要性（%）。所有特征的标题和描述均见表 VI

表 VI：为数据泄露分析提取的属性。

	标题	描述
F1	证书无效	浏览器无法验证网站证书
F2	证书不匹配	网站名称与 SSL 证书不匹配
F3	证书已过期	网站证书变为无效
F4	剩余有效天数	网站证书的剩余有效天数
F5	检测为恶意	VirusTotal API 检测到网站域名
F6	恶意网站_	被 VirusTotal API 检测为恶意的网站
F7	恶意软件网站	被 VirusTotal API 检测为含有恶意软件的网站
F8	钓鱼网站	被 VirusTotal API 检测为钓鱼网站的网站
F9	页面大小（MB）	网站平均页面大小（兆字节）
F10	加载时间（秒）	网站平均页面加载时间（秒）
F11	请求数	网站平均请求数
F12	CSS	通过 Pingdom API 获取的 CSS 百分比
F13	Font	Pingdom API 检索到的字体百分比
F14	HTML	Pingdom API 检索到的 HTML 百分比
F15	图像	Pingdom API 检索到的图像百分比
F16	重定向	Pingdom API 检索到的重定向百分比
F17	脚本	通过 Pingdom API 获取的脚本百分比
F18	XHR	通过 Pingdom API 获取的 XHR 百分比
F19	列入黑名单标志	被 Sucuri API 检测为列入黑名单的网站
F20	恶意软件标志	Sucuri API 检测为含有恶意软件的网站
F21	DNSSEC 标志	使用 DNSSEC 的网站

数据泄露类型。 如图 9 所示，我们发现大多数事件属于“黑客攻击/信息技术”，在政府医院中占 45.75%，其次是私立医院和非营利医院，分别占 30.22% 和 22.58%。此外，我们观察到，“未经授权的访问/披露”是非营利医院中最常见的数据泄露类型，占 50%；私立医院和政府医院分别为 37.91% 和 32.98%。

数据泄露在线存在归因。 为了了解在线存在、安全属性与数据泄露事件之间的关系，我们使用带有非负约束（即单调约束）的梯度提升模型，来学习数据泄露事件的重要属性。

表 VI 展示了模型中使用的 21 个属性，图 10 展示了其中十个直接（且主要）
与发生数据泄露的网站相关。如图所示，当网站包含恶意软件（F20：Sucuri API 检测到含有恶意软件的网站）时，其（自然）更有可能涉及数据泄露事件。与网站数据泄露高度相关的其他特征包括（F15：Pingdom API 检索到的图像比例，以及 F13：Pingdom API 检索到的字体比例）。

要点：（**研究问题2**。）有 156 家非营利医院涉及已报告的数据泄露事件，显著多于另外两类医院。此外，根据最能表明医院网站发生数据泄露事件的前 10 项属性，医院的在线形象安全特征显然会影响其遭受攻击的可能性。我们还注意到，私立医院最不容易遭遇数据泄露。

VI. 结论

近期报告显示，针对医院网络发起攻击以窃取敏感患者数据的趋势日益上升。本文通过分析医院网站，研究了医院的在线存在。借助将医院分为政府、非营利和私营医院三类，我们开展了比较研究，揭示了各种结构和安全特征。值得特别注意的是，我们研究了SSL证书有效性、医院网站中的相关问题以及恶意关联行为。我们将收集到的属性作为特征，展示了能够指示与数据泄露事件相关的网站、并有助于理解其安全性的最重要属性。我们的研究结果是实现患者安全的最初步骤之一，同时令人警醒地凸显了许多医院网站中存在的安全漏洞。

参考文献

[1] Abdulrahman Alabduljabbar、Runyu Ma、Sultan Alshamrani、Rhongho Jang、Songqing Chen 和 David Mohaisen。海报：测量和评估免费内容网站的风险。载于 网络与分布式系统安全研讨会（*Network and Distributed System Security Symposium，NDSS'22*），美国加利福尼亚州圣迭戈，2022年。[2] Abdulrahman Alabduljabbar、Runyu Ma、Soohyeon Choi、Rhongho Jang、Songqing Chen 和 David Mohaisen。通过分析免费内容网站的SSL证书了解其安全性：一项比较研究。第19–25页。ACM，2022年。 [3] Abdulrahman Alabduljabbar 和 David Mohaisen。通过自动化隐私政策分析与注释测量免费内容网站的隐私维度。载于 *2022*年世界互联网大会伴随会议，线上活动／法国里昂，*2022*年*4*月*25*日至*29*日，第860–867页。ACM，2022年。 [4] Mohammed Alkinoon、Sung J. Choi 和 David Mohaisen。测量医疗保健数据泄露事件。载于 信息安全应用——第*22*届国际会议，*WISA 2021*，韩国济州岛，*2021*年*8*月*11*日至*13*日，修订精选论文，《计算机科学讲义》第13009卷，第265–277页。Springer，2021年。

[5] Mohammed Alkinoon、Marwan Omar、Manar Mohaisen 和 David Mohaisen。医疗保健领域的安全泄露事件：时空分析。载于 计算数据与社交网络——第*10*届国际会议，*CSoNet 2021*，线上活动，*2021*年*11*月*15*日至*17*日，会议论文集，《计算机科学讲义》第13116卷，第171–183页。Springer，2021年。[6] Omar Alrawi 和 Aziz Mohaisen。不信任链：理解用于签署恶意应用程序的证书。载于第*25*届万维网国际会议论文集，*WWW 2016*，加拿大蒙特利尔，*2016*年*4*月*11*日至*15*日，伴随卷，第451–456页。ACM，2016年。[7] APIVoid。提供用于网络威胁分析、威胁检测和防范的实用JSON API的框架，2022年。[8] Adam Bates、Joe Pletcher、Tyler Nichols、Braden Hollembaek、Dave Tian、Kevin R. B. Butler 和 Abdulrahman Alkhelaifi。通过动态链接确保SSL证书验证安全。载于*2014*年*ACM SIGSAC*计算机与通信安全会议论文集，美国亚利桑那州斯科茨代尔，*2014*年*11*月*3*日至*7*日，第394–405页。ACM，2014年。[9] Jake Berkowsky 和 Thaier Hayajneh。证书颁发机构的安全问题。载于第*8*届*IEEE*年度普适计算、电子与移动通信会议，*UEMCON 2017*，美国纽约市，*2017*年*10*月*19*日至*21*日，第449–455页。IEEE，2017年。[10] Yuting Chen 和 Zhendong Su。SSL/TLS实现中证书验证的引导式差分测试。载于 *2015*年第*10*届软件工程基础联合会议论文集，*ESEC/FSE 2015*，意大利贝加莫，*2015*年*8*月*30*日至*9*月*4*日，第793–804页。ACM，2015年。[11] Sung J. Choi 和 M. Eric Johnson。了解数据泄露事件与医院广告支出之间的关系。《美国管理式医疗杂志》，25(5)，2019年1月。[12] Taejoong Chung、Yabing Liu、David R. Choffnes、Dave Levin、Bruce MacDowell Maggs、Alan Mislove 和 Christo Wilson。测量并应用无效 SSL 证书：沉默的大多数。载于*2016*年*ACM*互联网测量会议论文集，*IMC 2016*，美国加利福尼亚州圣莫尼卡，*2016*年*11*月*14*日至*16*日，第527–541页。ACM，2016年。[13] Taejoong Chung、Roland van Rijswijk-Deij、Balakrishnan Chandrasekaran、David R. Choffnes、Dave Levin、Bruce M. Maggs、Alan Mislove 和 Christo Wilson。DNSSEC生态系统的纵向端到端视图。载于 第*26*届*USENIX*安全研讨会，*USENIX Security 2017*，加拿大不列颠哥伦比亚省温哥华，*2017*年*8*月*16*日至*18*日，第1307–1322页。USENIX协会，2017年。[14] Jeremy Clark 和 Paul C. van Oorschot。知识体系：SSL和HTTPS：重新审视过去的挑战并评估证书信任模型增强方案。载于 *2013*年*IEEE*安全与隐私研讨会，*SP 2013*，美国加利福尼亚州伯克利，*2013*年*5*月*19*日至*22*日，第511–525页。IEEE计算机学会，2013年。[15] CloudFlare。什么是域名注册商？，年份为 = 2022。[16] CloudFlare。什么是顶级域名（TLD）？，2022 年。[17] Scott E. Coull、Andrew M. White、Ting-Fang Yen、Fabian Monrose 和 Michael K. Reiter。了解域名注册滥用。计算机安全，31(7)：806–815，2012年。[18] Marco Cova、Corrado Leita、Olivier Thonnard、Angelos D. Keromytis 和 Marc Dacier。流氓杀毒软件攻击活动分析。载于 入侵检测最新进展，第*13*届国际研讨会，*RAID 2010*，加拿大安大略省渥太华，*2010*年*9*月*15*日至*17*日。会议论文集，《计算机科学讲义》第6307卷，第442–463页。Springer，2010年。[19] Verizon Enterprise。Verizon数据泄露调查报告，2022年。[20] 民权办公室（OCR）。数据泄露通知规则，2013年。[21] 民权办公室（OCR）。美国卫生与公众服务部民权办公室数据泄露门户网站，2023年。[22] Definitive Healthcare。非营利医院与营利性医院有何区别？，2022年。[23] 互联网安全研究组（ISRG）。Let’s Encrypt。[24] Faouzi Kamoun 和 Mathew Nicho。医疗保健数据泄露事件中的人为与组织因素：数据泄露的成因与防范的瑞士奶酪模型。国际医疗健康信息系统与信息学杂志，9(1)：42–60，2014年。[25] Doowon Kim、Haehyun Cho、Yonghwi Kwon、Adam Doup´e、Sooel Son、Gail-Joon Ahn 和 Tudor Dumitras。HTTPS网络钓鱼生态系统中证书颁发机构实践的安全分析。载于 *ASIA CCS*’21：ACM亚洲计算机与通信安全会议，线上活动，中国香港，2021年6月7日至11日，第407–420页。ACM，2021年。[26] Doowon Kim、Bum Jun Kwon和Tudor Dumitras。经认证的恶意软件：衡量Windows代码签名公钥基础设施中的信任破坏。载于2017年ACM SIGSAC计算机与通信安全会议论文集，CCS 2017，美国得克萨斯州达拉斯，2017年10月30日至11月3日，第1435–1448页。ACM，2017年。[27] Doowon Kim、Bum Jun Kwon、Kristi´an Koz´ak、Christopher Gates和Tudor Dumitras。破碎的盾牌：衡量Windows代码签名公钥基础设施中的撤销有效性。载于第27届USENIX安全研讨会，USENIX Security 2018，2018年17日，美国马里兰州巴尔的摩，2018年8月，第15–851–868页。USENIX协会。[28] Bum Jun Kwon、Sanghyun Hong、Yuseok Jeon和Doowon Kim。韩国的经认证恶意软件：代码签名公钥基础设施生态系统中信任破坏的本地化研究。载于信息与通信安全：第23届国际会议，ICICS 2021，中国重庆，2021年11月19日至21日，论文集，第一部分，《计算机科学讲义》第12918卷，第59–77页。Springer，2021年。[29] Nir Menachemi1和 Taleah H Collum。电子健康记录系统的优势与不足，2011年。[30] Ulrike Meyer和Vincent Drury。经认证的网络钓鱼：了解网络钓鱼网站的公钥证书。载于第十五届可用隐私与安全研讨会，SOUPS 2019，美国加利福尼亚州圣克拉拉，2019年8月11日至13日。USENIX协会，2019年。[31] Mishari Al Mishari、Emiliano De Cristofaro、Karim M. El Defrawy和 Gene Tsudik。收集SSL证书数据以识别网络欺诈。《国际网络安全杂志》，14(6)：324–338，2012年。[32] NLTK。自然语言工具包，2022年。[33] 美国国土安全部。国土基础设施基础级数据（HIFLD），2019年。[34] Pingdom。网站性能与可用性监测，2022年。[35] Martin Roetteler、Michael Naehrig、Krysta M. Svore和Kristin E. Lauter。计算椭圆曲线离散对数的量子资源估算。载于密码学进展：ASIACRYPT 2017：密码学与信息安全理论及应用第23届国际会议，中国香港，2017年12月3日至7日，论文集，第二部分，《计算机科学讲义》第10625卷，第241–270页。Springer，2017年。[36] Adil Hussain Seh、Mohammad Zarour、Mamdouh Alenezi、Amal Sarkar、Alka Agrawal、Rajeev Kumar和 Prof. Raees Khan。医疗保健数据泄露：见解与影响。《医疗保健》，8：133，2020年5月。[37] B. K. Siddartha和G. K. Ravikumar。分析掩码技术以发现医疗保健领域的安全性及其他效率问题。载于第三届I-SMAC国际会议，第660–666页，2019年。[38] Tanshanika Smith。审视医疗保健领域的数据隐私泄露。技术报告，Walden University，2016年。[39] Bruce Steinwald、和Duncan Neuhauser。私立医院的作用，1970年。[40] Sucuri。网站安全检查和恶意软件扫描器，2023年。[41] Michel JG Van Eeten和Johannes M Bauer。恶意软件经济学：安全决策、激励因素与外部性。2008年。[42] Luis Vargas、Logan Blue、Vanessa Frost、Christopher Patton、Nolen Scaife、Kevin R. B. Butler和Patrick Traynor。数字医疗相关感染：大型多院区医院系统安全性的案例研究。载于第26届年度网络与分布式系统安全研讨会，NDSS 2019，美国加利福尼亚州圣迭戈，2019年2月24日至27日。互联网协会，2019年。[43] VirusTotal。分析可疑文件和网址以自动检测恶意软件类型，2023年。[44] W3Techs。网站默认协议https的使用统计，2022年。[45] Steven Walker-Roberts、Mohammad Hammoudeh和Ali Dehghan-tanha。医疗保健关键基础设施中内部威胁应对措施的可用性和有效性的系统综述。IEEE Access，6：25167–25177，2018年。[46] WHOIS。注册数据查询工具，2023年。[47] Suanu Bliss Wikina。是什么导致了泄露？对信息技术使用与健康数据泄露的考察。《健康信息管理视角》，11（秋季），2014年。[48] Liang Zhang、David R. Choffnes、Tudor Dumitras、Dave Levin、Alan Mislove、Aaron Schulman和Christo Wilson。Heartbleed事件后SSL证书重新签发与撤销的分析。《ACM通讯》，61(3)：109–116，2018年。