Got Sick and Tracked: Privacy Analysis of Hospital Websites

Xiufen Yu, Nayanamana Samarasinghe, Mohammad Mannan, Amr Youssef
Concordia University, Montreal, Canada
{*y_xiufe,n_samara,mmannan,youssef*}*@ciise.concordia.ca*

***Abstract*—Many healthcare organizations, including hospitals, are moving some of their services to digital space. While web privacy in popular sites including commercial and social media sites has been widely studied, privacy and security issues of hospital websites have not been studied at a global scale, which we target for our analysis in this paper. We successfully crawl 19,483 hospital websites from 152 countries and provincial jurisdictions located in Asia, Europe, North America, Latin America, Africa and Oceania. We identify wide-spread use of trackers on hospital websites — 10,417 (53.5%) hospital websites included tracking scripts/cookies; Beside privacy issues, we also identify sites with potential security issues — 33 hospital sites were flagged as malicious by VirusTotal (by at least 3 security engines); and 699 hospital websites included *FullStory*, *Yandex*, *Hotjar* session replay services that sent sensitive information to external servers. We hope our findings will raise awareness in improving privacy and security posture of hospital websites, as the number of online hospital services is expected to increase in the future.**

***Index Terms*—hospital websites, tracking, privacy, security**

1. Introduction

Increased tracking of online user behaviours has become the norm for most commercial web services [1], although users can still choose a relatively less privacy invasive service, e.g., between search engines such as Google vs. DuckDuckGo. On the other hand, some websites (e.g., government health and hospital services) do not have any alternatives [2], should a user identifies potential tracking activities. With the COVID-19 pandemic, more health services are being offered online to limit the spreading of the virus—e.g., a general practitioner can be channeled in minutes, around the clock [3], without having to wait for an in-person meeting. As such, patients are able to consume health related services from online services with a few clicks — book appointments, health checkups, and view medical results. Unlike the interactions with other commercial websites, a variety of sensitive information items (e.g., identity information, health status, mental health, reproductive care including abortion, substance abuse) are exchanged with hospital sites. These sensitive information can be leaked to third-parties if trackers/session-replay scripts are deployed on hospital websites. Disclosure risks of such sensitive information may include discrimination, social stigma and physical harm.

Privacy and security of health care systems is paramount, and appropriate policies to safeguard its users needs to be enforced [4]. However, lapses in the deployment of such effective measures are common. For example, a German security firm (Greenbone Networks) found that medical files of 107 million medical images (e.g., X-rays, scans) of Indian patients were leaked and made available online [5]. These medical records happen to contain various sensitive information of patients (e.g., patient name, date of birth, medical institution name, ailment, physician name). In another incident, computer systems of a major hospital chain, with hospitals in more than 400 locations, failed when it was hit by a ransomware attack [6]. Stolen health records may have a higher demand (cf. credit card numbers) in the darkweb [7]. Similarly, the cost to remediate breaches in health care is also high [7].

There are several studies (e.g., [8]–[11]) relating to privacy of health services, but they target a specific geographical location. Robinson [11] analyzed 210 public hospital websites in Illinois, USA and found 94% of websites include trackers on them; most common trackers on these websites include Google Analytics (74%), Google (88%), and Facebook (26%). Niforatos et al. [9] analyzed 61 US hospital websites, and found they collect information relating to advertisements (61, 100%), third-party cookies (55, 90%) and session recording (14, 23%) services. Most of these trackers are from Facebook (40, 61%) and Google (54, 89%).

In this work, we perform a large scale web privacy measurement study of hospital websites, using 19,635 hospital websites from 152 countries. We collect hospital URLs from several sources (e.g., [11]–[13]) by scraping the source code of the corresponding web pages. Thereafter, we crawl the extracted hospital website URLs using the OpenWPM [14] web privacy measurement framework; 152 sites were unreachable. To the best of our knowledge, this is the first measurement study on the privacy/security of hospital websites, performed at a global scale. We analyze the instrumented tracking metrics (third-party scripts/cookies, fingerprinting APIs) using the OpenWPM database. We filtered the websites using session replay services , and we inspected the potential sites using session replay with *HTTP Toolkit* [15] to identify specific information leaked (e.g., date of birth). We also use VirusTotal [16] to identify hospital sites and domains hosting scripts/cookies that are malicious.

Contributions and notable findings.

1) We develop a framework to collect hospital websites from various external sources, and a test methodology to evaluate these sites for possible privacy exposures.

2) We found that 699/19,483 (3.6%) hospital websites included session replay services — e.g., *FullStory*,[1] *Yan-*

1. https://www.fullstory.com/session-replay/

dex,[2] *Hotjar*.[3] 91/699 (13.0%) of these websites belong to EU hospitals. We observed users' information was sent from these hospital sites to third-party servers (*FullStory*, *Yandex* and *Hotjar*). The information sent to these external servers (owned by session replay services) include sensitive information such as phone number, date of birth, user credentials, residential address, passport information, booked medical services.

3) We found widespread use of commercial trackers on hospital websites. Major known trackers[4] include Google, Addthis, Facebook and Baidu. We observed 10,417/19,483 (53.5%) hospital websites included tracking scripts/cookies. There were tracking cookies set to last for a very long time — 5.8% (1136/19,483) of sites included 1713 known tracking cookies expiring in the year 9999. These trackers are embedded in analytic services, and other third-party services (e.g., Google maps) on landing pages of hospital websites.

4) We observed hospital websites in Oceania (61.7%. 140/227) and North America (60.1%, 2805/4666) included a large proportion of known tracking scripts, compared to Asian hospital websites (39.6%, 2844/7183). Known tracking cookies were set in less than 15% of hospital websites except for North America (8186/28,960, 28.3%). Known trackers in China are location specific, perhaps due to the use of alternative local services, as foreign web services (e.g., Google, YouTube, Facebook) are mostly blocked in China.

5) We found 33/19,483 hospital websites were flagged as malicious by at least 3 security engines used by VirusTotal (e.g., *cliniqueelmenzah.com*, *mathahospital.org*). Additionally, 11 and 18 domains of known tracking scripts and cookies were flagged as malicious by at least 3 security engines, respectively. We have notified administrators for 18 of these websites about our findings; no contact information were available for the remaining 15 websites.

2. Related work

Web tracking measurements. There are many past studies that measured the privacy exposures from a variety of popular web applications and mobile apps. Englehardt et al. [18] implemented the OpenWPM web privacy measurement framework to identify online tracking behaviours of websites, and used their framework to measure tracking in top-1M sites. The authors found Google and Facebook trackers dominate in tracking websites. Samarasinghe et al. [19] measured web tracking in top-1K sites from 56 countries, and found Google trackers are highly prevalent on those sites (irrespective of the location), and many cookies were valid for more than 20 years. Acar et al. [20] extended OpenWPM to investigate attacks that exfiltrate data using third-party scripts (i.e., misuse of browsers' internal login managers, social data exfiltration, whole-DOM exfiltration), and found sites that leak sensitive user information (e.g., credit card information, medical details, passwords) to session replay services. Xuehui et al. [21] studied tracking in top country specific sites (in Alexa [22] list) from 4 countries (UK, China, Australia, US), and found tracking behaviours that are specific to those countries — e.g., users in China were tracked less than those in the UK. Google Analytic is the most common tracker in 74% of hospital websites. Papadogiannakis et al. [23] found more than 75% of tracking activities happened even before interacting with the cookie banners, or after users reject all possible cookies. We measure tracking in hospital sites from 152 countries around the world, and found level of tracking in countries located in different regions vary — e.g., proportion of third-party scripts in North America is relatively higher compared to that of other regions; i.e., percentage of hospital websites with third-party scripts and cookies in North America was 60% and 29%, respectively. In addition, we observe location specific trackers (e.g., *baidu.com* on Chinese hospital websites).

Privacy and security issues in health related websites. Past studies on privacy and security issues of hospitals targeted hospitals only from a specific or a few jurisdictions. Zheutlin et al. [8] performed a study of patient data tracking on 86 pharmacy websites. The authors found that 76.4% of these websites included ad trackers; other tracking methods used include third-party cookies, session monitoring[5] (using *Blacklight* [24]), keystroke capturing, sharing data with top tracking entities (e.g., Google, Facebook). Joshua et al. [9] studied tracking on 61 US hospital websites, and found among other forms of tracking, 14 (23%) websites used session recording services to track users. Celine et al. [10] studied how caregivers' access to patient portals may jeopardize user privacy and security. The authors found 69/102 (68%) hospitals provided proxy accounts to caregivers; 94/102 (92%) hospitals were asked about password sharing between patient and caregiver, and 42/92 (45%) endorsed such practice. Robinson et al. [11] studied 210 public hospital websites in Illinois, USA, and found 94% of hospital websites included an average of 3.5 trackers. Wesselkamp et al. [25] found advertising cookies performing cross-site tracking in health related websites (i.e., for booking appointments). We found 10,417/19,483 (53.5%) hospital websites included tracking scripts/cookies. Google dominates in tracking hospital websites. Third-party scripts included in 699/19,483 (3.6%) hospital websites sent user information to external session replay servers (*FullStory*, *Yandex*, *Hotjar*). In addition, we observed 33/19,483 hospital websites were flagged as malicious by VirusTotal [16].

3. Methodology

In this section, we detail our methodology for hospital website (URLs) collection, and privacy analysis and measurement techniques for the collected websites; see Fig. 1 for an overview.

3.1. Collecting hospital websites

To extract hospital websites, we use *webometric world hospital websites* as the primary source of information. We programmatically parse the content of each of the tabs appearing on the landing page of *webometric world*

2. https://yandex.com/support/metrica/general/counter-webvisor.html

3. https://www.hotjar.com/session-replay-software/

4. We define a *known tracker* as the third-party (e.g., script/cookie on a first-party website) blocklisted by *EasyPrivacy* [17] filtering rules.

5. Session monitoring reveals only the use of tracking technologies in a browsing session, but not able to replay a recorded session.

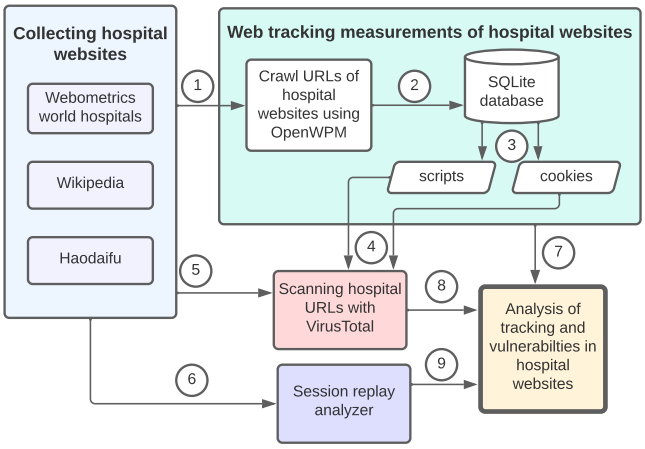

Figure 1: Overview of our methodology: hospital website collection and tracking measurement on websites — steps ①, ⑤, ⑥ represent hospital websites served as input to OpenWPM, VirusTotal scans, session replay analyzer, respectively; step ② is instrumented data saved to OpenWPM database; step ④ is third-party script/cookie domains fed to VirusTotal scans; steps ⑦ (OpenWPM measurement data), ⑧ (malicious domains detected), ⑨ (domains subjected to session replay) are the output for further analysis from OpenWPM, VirusTotal scan results, and websites subjected to session replay, respectively.

hospital websites [13] corresponding to different regions. For every hospital website URL, we extract corresponding meta data (e.g., hospital name, country, continent). Since *webometric world hospital websites* is not a complete list of hospital websites, we also complement other available hospital website lists — e.g., for China we use *Haodaifu* (see Appendix).

We collect 19,635 unique hospital websites from different sources (i.e., *webometrics world hospital websites* [13], *Wikipedia* [26] and *haodaifu* [12]) for our privacy measurements. The hospital websites that we collect are hosted in countries pertaining to different regions — Asia (7183), Europe (5936), North America (4666), Latin America (1362), Oceania (227), Africa (261).

We also identify hospital websites with login forms on landing pages by matching the corresponding source code with specific keywords (e.g., login, user id, password). This approach will work for any site irrespective of the language of page content, as the source code syntax of a web page is independent of the language of page content.

3.2. Web privacy measurements

We configure OpenWPM [27] web privacy measurement framework to run with 15 parallel browser instances in headless mode. We explicitly enable OpenWPM instrumentations for HTTP requests, JavaScript, cookies, DNS requests, callbacks and page navigations. Javascript instrumentation includes passive fingerprinting APIs used in the website. We clear the browser profile after each URL visit, to simulate the first visit to the browser instance, to avoid any influences from past browsing history. We use a physical machine (connected to our university network) running Ubuntu server 20.4 LTS, 32GB RAM, 1TB SSD, Intel Core i7-6700 CPU for our measurements between Sept. 1, 2021–Dec. 31, 2021. A total of 19,635 hospital websites from 152 countries were crawled using OpenWPM from a city in North America; 152 websites failed due to expired domain registrations and unreachable websites. The instrumented tracking metrics extracted from OpenWPM are saved to a SQLite database for further analysis. The saved information in the database contains both stateful (i.e., scripts/cookies) and stateless (fingerprinting) forms of tracking metrics. We then examine the saved tracking scripts/cookies for third-party domains (i.e., domains of scripts/cookies that do not match the domain of the hospital site that they are on).

Categorize third-party scripts and cookies. A third-party is a script/cookie included on a first party website (i.e., hospital website). We use filtering rules [17] that block third-parties on hospital sites to identify 3 categories of third-party domains: *EasyList* rules block ad-related third-parties; *EasyPrivacy* blocks known trackers; third-parties that are not blocked by EasyList/EasyPrivacy filtering rules are treated as unknown trackers.

Identify fingerprinting APIs. We use the instrumented JavaScript data to extract fingerprinting APIs included in hospital websites. Third-party domains hosting scripts that include these fingerprinting APIs are of different types — e.g., window.navigator, window.screen, window.document, HTMLCanvasElement, CanvasRenderingContext2D, AudioContext, RTC. These fingerprinting APIs are used to passively track users by leveraging various characteristics of a user's environment, including hardware, operating system and software characteristics.

3.3. Session replay scripts

We extract hospital websites that include scripts (e.g., *fs.js*, *tag.js*, *hotjar-HotjarID.js*) with known session replay functionality [20] from the *javascript* table of OpenWPM SQLite database. These scripts pertain to *Hotjar*, *Yandex* and *FullStory* session replay services. We observe websites with *Hotjar* (but not *FullStory*, *Yandex*) session replay scripts send data over websockets. Therefore, we use *selenium-wire* [28] to automate the crawling of the landing page of 469 hospital websites with *Hotjar* session replay scripts, to identify the sites sending data over web sockets directly from the landing pages.

While existence of the session replay scripts (and the use of websockets by *Hotjar*) can be easily enumerated, it requires some manual effort (e.g., filling out forms) to understand what is leaked to the session replay servers. Therefore, we limit our manual tests to a selected set of hospital websites (183, of which 101 sites with *Yandex* services across multiple continents, 78 EU sites with *Hotjar*, and 4 sites with *FullStory* scripts). We observe 40/183 hospital websites require to create an account prior to booking an online appointment; 74 hospital websites have online forms (without account registration) to book an online appointment; remaining websites (69) do not have functionality to book an online appointment. We created accounts in 40 of hospital websites that require an online registration. Then we use crafted (fake) data (e.g., user name, password, email address) to book online appointments with 114 (i.e., 40 sites with registration and 74 without registration) hospital websites. Thereafter, for

those 114 hospital websites, we use *Chrome DevTools* [29] and *HTTP Toolkit* [15] to identify sensitive information transmitted to remote servers during session replay.

3.4. Detecting malicious domains

Potential security issues in hospital websites can lead to privacy issues. Therefore, to determine hospital websites and included third-party script/cookie domains that are malicious, we scan all 19,483 hospital websites, and 3673 third-party domains hosting scripts/cookies using VirusTotal. We report only those domains that are flagged by at least 3 security engines as malicious.

3.5. Limitations

Our hospital website collection technique may not find all hospital websites in any given jurisdiction. Additionally, we use filtering rules [17] to identify known advertisers and trackers, which are not comprehensive enough to find all possible tracking domains (especially country specific trackers). Some known advertisers/trackers may operate in a dual role of advertising and tracking. We also involved manual steps in verifying false positives/negatives of hospital websites including scripts pertaining to session replay services, which is non-trivial to automate.

4. Results

In this section, we report our findings on privacy issues of hospital websites. We also report additional result on hospital sites using HTTP in the Appendix.

4.1. Session replay

Session replay services are used to replay a visitor's session through the browser, to get a deeper understanding of a user's browsing experience; information replayed include user interactions on a website such as typed inputs, mouse movements, clicks, page visits, tapping and scrolling events. During this process, users' sensitive information can be exposed to third-party servers that host session replay scripts. We identified three session replay services in the analyzed hospital websites (19,483): *Hotjar* (469, 2.4%), *Yandex* (226), *FullStory* (4); see Table 2 for examples of hospital websites with session replay services. The regions that have a heavy presence of session replay services on their hospital websites include North America (291/4666, 6.2%) and Europe (299/5936, 5.0%); see Table 1. In total, we found session replay scripts on 699 hospital websites; 91/699 (13.0%) of sites were from EU countries.

Yandex. The session replay scripts hosted by *Yandex* were included in 153/226 (67.7%) hospital websites in Russia. These *Yandex* session replay scripts collect sensitive medical information of users and send to remote servers (over HTTPS). In addition, sensitive information is exposed while performing common interactions with hospital websites, including booking online appointments, contacting hospital by entering sensitive information (e.g., medical description); see Table 6 (in Appendix). There were 24 (out of 101 — see Sec. 3.3) Russian hospital websites that leak sensitive information with *Yandex* session replay services — user name, password, phone number, date of birth, address (street, city, country), passport information collected from *lk.baltclinic.ru*; requested medical service and login information collected from *medvedev.ru*, *zdordet.ru*, *vizus1.ru*, *gutaclinic.ru*, *presidentclinic.ru*; user comments/messages collected from *alfa-med.ru*, *benefacta.ru*, *gkb12.ru*, *glazalazer.ru*, *presidentclinic.ru*, *onclinic.md*, *vizus1.ru*. We found 13 hospital websites in EU countries include *Yandex* session replay scripts, and 3 of these EU hospital websites (in Greece, Portgual and Czech Republic) apparently violate GDPR [30] privacy regulation. These 3 EU hospitals leak information to *Yandex* remote servers as follows: *multiscan.cz* (in Czech Republic) leaked search information from the search functionality; *lifeclinic.gr* (in *Greece*) leaked user name, phone number, email, subject and message sent; and *chpvvc.pt* (in Portgual) leaked name, email, service rendered and message sent.

Region	FullStory	Hotjar	Yandex
Europe	1	108	190
NorthAmerica	2	282	7
LatinAmerica	-	37	1
Asia	-	20	28
Africa	-	3	-
Oceania	1	19	-

TABLE 1: Session replay services on hospital websites.

FullStory. We observed *FullStory* session replay scripts included in *www.mater.org.au, www.ramsayhealth.co.uk*, sent visited page, and screen width and height of the user's display to a remote server.

Hotjar. Session replay code from *Hotjar* is included within the *head* tags in the hospital website page source as a JavaScript snippet [31]. The session replay data captured from *Hotjar* scripts is sent to a remote server using websocket connections. From our automation with *selenium-wire*, we found 27/469 (5.8%) hospital websites that include *Hotjar* session replay scripts, sent data over websockets to remote servers — e.g., 3 EU hospital websites and 18 US hospital websites enable such data transmission (apparently, violating GDPR and HIPPA privacy regulations, respectively); the remaining 6 hospital websites are in non-EU countries. In addition, by manually inspecting 78 (see Sec. 3.3) EU hospital websites with *HTTP Toolkit*/*Chrome DevTools*, we found 4 of the inner URLs from those sites, leaked sensitive information through websockets — e.g., user name, email, phone number, medical service are sent from *www.bilicvision.hr* (in Croatia); user name, email, phone number, message and country are sent from *www.reprofit.cz* (in Croatia); see Table 7 (in Appendix) for information leaked by hospital websites with session replay services in the EU countries.

4.2. Domains flagged as malicious

With VirusTotal, we found 33/19,483 websites were flagged as *phishing*, *malicious* or *malware* by at least 3 VirusTotal engines;[6] 26 of the flagged sites were part of

6. https://support.virustotal.com/hc/en-us/articles/115002146809-Contributors

SRS	Hospital domain names	Leaked data
Hotjar	bilicvision.hr, multi-scan.cz	name, email, password, phone, chat
Yandex	alfa-med.ru, bakulev.ru	name, email, password, phone, speciality
FullStory	ramsayhealth.co.uk	URL, screen width, screen height

TABLE 2: Examples of hospital websites with session replay services — SRS = Session replay service.

more than one VirusTotal category; 27 sites were flagged as *phishing*. We did not consider scan results from some VirusTotal engines (e.g., *CRDF*, *Quttera*) as the results from those engines were unreliable. Most hospital websites flagged by VirusTotal were in China (10/33, 30.3%) and India (3/33, 9.1%). We also looked into malicious JavaScript files that were included in the 33 flagged hospital websites; *ultramed.pl* (in Poland) and *bcm.es* (in Spain) included 10 and 2 unique malicious JavaScript files, respectively. The common malicious JavaScript files contained *jQuery* keyword in its file name (e.g., *jquery.min.js*, *jquery.themepunch.tools.min.js*), or were part of *WordPress* web applications (e.g., *wp-embed.min.js*, *wp-emoji-release.min.js*). *jQuery* is a commonly used JavaScript library, and it is the base for many add-on scripts/plugins that are also included in platforms such as *WordPress* [32], [33]. Fake *jQuery* scripts with malicious source code [34] can be dangerous for users.

The following 6 hospital websites (in 4 countries) were flagged as malicious by more than 5 security engines: a Tunisian hospital website (*cliniqueelmenzah.com*) was flagged as malicious by 9 security engines); sites from China (*jrszyy.com*, *zyxyfy.com*, *ahzxy.com*), India (*mathahospital.org*) and Brazil (*hsja.com.br*) were flagged as malicious by 6 security engines. The malicious categories of these flagged websites include *known infection source*, *media sharing*, *compromised websites*, *malicious*, *malware* and *spyware*.

We also scanned all 3673 third-party domains (of scripts/cookies) using VirusTotal, and found 27 of them (e.g., *iclickcdn.com*) were flagged by at least 3 VirusTotal engines. For the domains hosting third-party scripts/cookies, 11 and 18 were flagged as malicious and malware, respectively. In Table 4 and Table 5 (in Appendix), we list examples of potentially malicious domains hosting tracking scripts and cookies (including the presence of such domains on hospital sites), respectively.

4.3. Third-party tracking scripts

We found 9443/19,483 (48.5%) of hospital websites included at least one known tracking script. Hospital websites in Oceania (140/227, 61.7%) and North America (2805/4666, 60.1%) had a high percentage of websites with known tracking scripts. Hospital sites in Asia (2844/7183, 39.6%) had a relatively lower proportion of sites with known tracking scripts; see Fig. 2. Top known trackers included on hospital websites (19,483) are: *googleanlaytics* (6607, 33.9%), *googletagmanager* (4816, 24.7%), *facebook* (2552, 13.1%) and *cloudflare* (564, 2.9%); see Fig. 3. Both *googletagmanager* and *googleanalytics* are used to collect tracking/marketing data on hospital websites; *gtag.js* sent event data to Google Analytics,

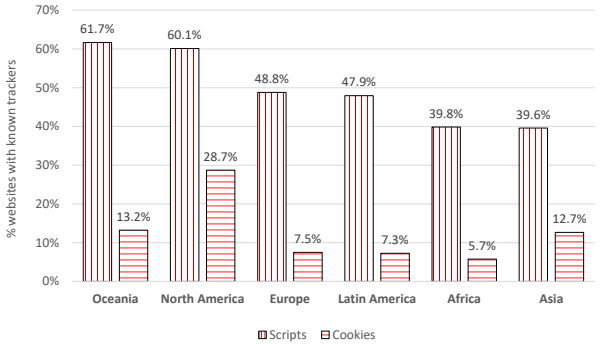

Figure 2: Percentage of hospital websites with known tracking scripts/cookies.

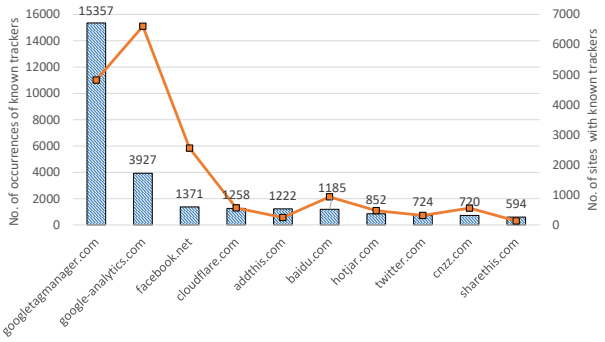

Figure 3: Top-10 known tracking scripts on hospital sites - the bars show the number of occurrences of known tracking scripts (vertical axis to the left), while the line chart shows the number of hospital websites with known tracking scripts.

Google Ads and Google marketing platforms. Google Maps was included in 1591 hospital websites; YouTube videos were embedded in 1372 hospital websites; Addthis (*s7.addthis.com*) contained adware that redirected users to promotional websites (246/19,483, 1.3%).

There were no significant differences relating to the proportion of hospital websites with various categories of third-parties (i.e., ads, known trackers, unknown trackers) between different geographical regions; see Fig. 4. However, some countries (with more than 9 hospital websites in our dataset) in different regions had known tracking scripts in most of its hospital websites — Finland (18/23, 78.3%); Belarus (10/13, 76.9%); Norway (28/38, 73.7%); Latvia (18/25, 72.0%); Kuwait (7/9, 77.8%); Japan (702/1012, 69.4%). We also found known tracking scripts that are region specific; *bdstatic.com*, *qq.com* and *50bang.org* only tracked websites in Asia; *adsrvr.org*, *rtrk.com*, *btttag.com* and *cloudfront.net* were only found on North American hospital websites; Oceania had only one regional script domain (*turbolion.io*); Africa had no regional tracking script.

4.4. Third-party tracking cookies

We found 2839/19,483 (14.6%) hospital websites from 85 countries set known tracking cookies; see Fig. 5. The top-3 regions with the highest proportion of known tracking cookies set on hospital sites were Asia (3086/8689, 35.5%), North America (8186/28,960, 28.7%) and Oceania (141/594, 23.7%); see Fig. 6. Taobao (Alibaba) that

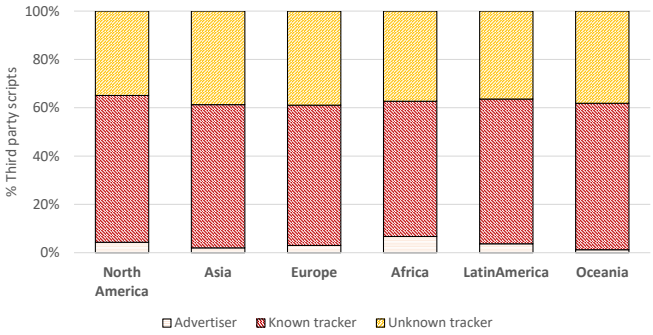

Figure 4: Proportions of third-party scripts in different categories (tracking, advertising and unknown) included on hospital websites by region.

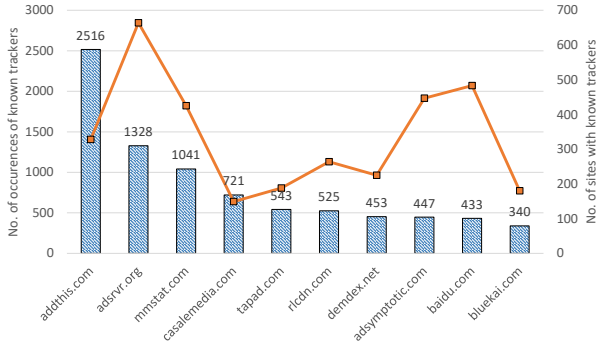

Figure 5: Top-10 known tracking cookies on hospital sites - the bars show the number of occurrences of known tracking cookies (vertical axis to the left), while the line chart shows the number of websites with such cookies.

collects user behaviours for targeted advertising [35], *mmstat.com* sets third-party cookies on a large proportion of hospital websites in China (425/4324, 9.8%). Similarly, a large proportion of known tracking cookies (483/4324, 11.2%) were set by *baidu.com* on Chinese hospital sites.

We also examined the cookie validity duration by regions, and found that 1017/3264 (31.2%) known tracking cookies set on hospital websites in Asia, were valid for more than 1000 years. Known tracking cookies that expire after 5 years include *mmstat.com* (1039) and *baidu.com* (431); see Table 3.

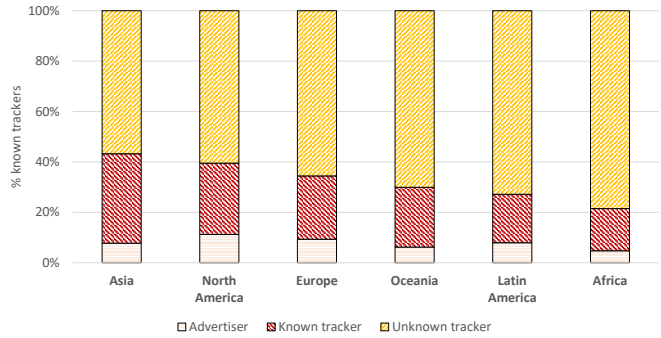

Figure 6: Proportions of third-party cookies in different categories (tracking, advertising and unknown) set on hospital websites by region.

		Cookie Expiry Duration			
Tracker	**#Sites**	**1m-1y**	**1y-5y**	**5y-100y**	**> 1000y**
addthis.com	2516	110	2345	-	-
adsrvr.org	1328	1328	-	-	-
mmstat.com	1041	2	-	237	802
casalemedia.com	721	572	-	-	-
tapad.com	543	543	-	-	-
rlcdn.com	525	525	-	-	-
demdex.net	453	453	-	-	-
adsymptotic.com	447	447	-	-	-
baidu.com	433	-	-	418	13
bluekai.com	340	340	-	-	-

TABLE 3: The top-10 known tracking cookies and their expiry periods (m=month, y=year).

4.5. Fingerprinting APIs

We found a large number of fingerprinting APIs (total: 3,082,179, unique: 222) included in the JavaScript source files used in hospital websites. Most common fingerprinting APIs include: *window.navigator* (1,146,303), *Storage* (407,847), *CanvasRenderingContext2D* (340,164), *HTMLCanvasElement* (133,657), hardware related APIs (32,394), *window.screen* (18,888), *RTCPeerConnection* (747), *window.navigator.geolocation* (722) and *AudioContext* (291). We also found several fingerprinting APIs with acoustically relevant characteristics of the audio signal — *GainNode* (49), *AnalyserNode*(110), *OscillatorNode*(374) and *ScriptProcessorNode* (38). Combinations of multiple fingerprinting APIs can be used to identify a user with a high precision [14].

5. Recommendations

Based on our analysis, we suggest a few possible mitigation strategies to reduce privacy exposures to third-parties from the perspective of site developers and regulators. Developers should analyze scripts used for tracking/fingerprinting, and use only those scripts that are required for the proper functioning of the sites. Similarly, the use of session replay scripts should be avoided, or at least configured properly to reduce the risk of data exposures. Since software packages and applications are becoming a target for malware and supply chain attacks (cf. SolarWinds [36]), developers should always scan the dependent software packages/libraries to ensure that hospital websites do not inherit such vulnerabilities.

From our manual analysis, we observed that while the privacy policies of some hospitals explicitly mention that they do not share any information with third-parties, several sites still send personal information to session replay services such as *Yandex* and *Hotjar*. For example, sanfil.pt (in Portugal) explicitly states in its privacy policy[7] that the information collected from users will not be shared with third-parties, while in reality, when a user uses the online chat function available on the website, all the chat messages are sent to *Hotjar*. In addition, despite the privacy policy[8] of *lifeclinic* (in Greece), and user agreement[9] of rami-spb.ru claim that a user's personal data

7. https://www.sanfil.pt/cookies/
8. https://www.lifeclinic.gr/privacy-policy/
9. https://www.rami-spb.ru/Content/poljzovateljskoe-soglashenie-ob-ispoljzovanii-sajta/4091

will not be disclosed to third-parties, personal information (e.g., username, phone, email and doctor's speciality) is leaked to *Yandex*. Therefore, regulators should invest into developing tools to detect such contradictory statements and violations to improve data privacy in the long run.

6. Conclusion

Similar to other popular commercial sites, hospital sites include commercial trackers hosted by top tech giants. We found that 10,417 (53.5%) hospital websites included such tracking scripts/cookies. We also observed that 33 of hospital websites are flagged as malicious by VirusTotal, possibly due to the use of malicious third-party resources (e.g., the use of fake and malicious *jQuery* libraries) in those sites. Therefore, developers need to be vigilant in including third-party libraries in hospital websites, and should do proper scanning before using such dependencies. Furthermore, sensitive user information is relayed to remote servers by including session replay scripts in hospital websites. Hospital websites continue to expand its services in digital space; the COVID-19 pandemic also contributed to the recent rapid increase of online hospital services. Given such growth, and the use of sensitive information at hospital services, proper safeguards should be implemented to prevent potential privacy/security exposures. Furthermore, governments should introduce and periodically review existing privacy regulations (e.g., the US HIPPA [37]) to protect sensitive information pertaining to patient identity and health records.

References

[1] A. Lerner, A. K. Simpson, T. Kohno, and F. Roesner, "Internet Jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016," in *USENIX Security Symposium (USENIX Security'16)*, Austin, TX, USA, Aug. 2016.

[2] N. Samarasinghe, A. Adhikari, M. Mannan, and A. Youssef, "Et tu, brute? Privacy analysis of government websites and mobile apps," in *TheWebConf'22*, Online, Apr. 2022.

[3] Maple, "Online doctors, virtual health & prescriptions in Canada," 2022, Online article (2022). https://www.getmaple.ca/.

[4] S. Braghin, A. Coen-Porisini, P. Colombo, S. Sicari, and A. Trombetta, "Introducing privacy in a hospital information system," in *Software engineering for secure systems (SESS'08)*, Leipzig, Germany, May 2008.

[5] The Economic Times, "Data Breach of Indian Patients," Online article. https://economictimes.indiatimes.com/tech/internet/german-firm-finds-one-million-files-of-indian-patients-leaked/articleshow/73921423.cms?from=mdr.

[6] NBC news, "Major hospital system hit with cyberattack, potentially largest in U.S. history," 2020, Online article (2020). https://www.nbcnews.com/tech/security/cyberattack-hits-major-u-s-hospital-system-n1241254.

[7] Aha.org, "A high-level guide for hospital and health system senior leaders," 2022, Online article (2022). https://www.aha.org/center/cybersecurity-and-risk-advisory-services/importance-cybersecurity-protecting-patient-safety.

[8] A. R. Zheutlin, J. D. Niforatos, and J. B. Sussman, "Data-tracking among digital pharmacies," *Annals of Pharmacotherapy*, 2022.

[9] J. D. Niforatos, A. R. Zheutlin, and J. B. Sussman, "Prevalence of third-party data tracking by us hospital websites," *JAMA Network Open*, vol. 4, no. 9, pp. e2 126 121–e2 126 121, 2021.

[10] C. Latulipe, S. F. Mazumder, R. K. Wilson, J. W. Talton, A. G. Bertoni, S. A. Quandt, T. A. Arcury, and D. P. Miller, "Security and privacy risks associated with adult patient portal accounts in us hospitals," *JAMA internal medicine*, vol. 180, no. 6, pp. 845–849, 2020.

[11] R. Robinson, "Prevalence of web trackers on hospital websites in Illinois," *arXiv preprint arXiv:1805.01392*, 2018.

[12] Haodf.com, "Chinese official names of hospitals in China," 2021, Online article (2021). https://www.haodf.com/hospital/list-11.html/.

[13] Cybermetrics lab, "Ranking web of hospitals," 2015, Online article (2015). https://hospitals.webometrics.info/.

[14] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *ACM CCS'16*, Vienna, Austria, Oct. 2016.

[15] Httptoolkit.tech, "Http toolkit," 2022, online article (2022). https://httptoolkit.tech/.

[16] VirusTotal, "VirusTotal," 2021, Online article (2021). https://www.virustotal.com.

[17] EasyList, "EasyList," 2022, online article (2022). https://easylist.to/.

[18] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *ACM CCS'16*, Vienna, Austria, Oct. 2016.

[19] N. Samarasinghe and M. Mannan, "Towards a global perspective on web tracking," *Computers & Security*, vol. 87, p. 101569, 2019.

[20] G. Acar, S. Englehardt, and A. Narayanan, "No boundaries: data exfiltration by third parties embedded on web pages," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 4, pp. 220–238, 2020.

[21] X. Hu, G. S. de Tangil, and N. Sastry, "Multi-country study of third party trackers from real browser histories," in *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2020, pp. 70–86.

[22] Alexa, "The top 500 sites on the web," 2022, online article (2022). https://www.alexa.com/topsites/countries.

[23] E. Papadogiannakis, P. Papadopoulos, N. Kourtellis, and E. P. Markatos, "User tracking in the post-cookie era: How websites bypass GDPR consent to track users," in *TheWebConf'21*, Ljubljana, Slovenia, Apr. 2021.

[24] Blacklight, "Blacklight," 2020, online article (2020). https://themarkup.org/blacklight.

[25] V. Wesselkamp, I. Fouad, C. Santos, Y. Boussad, N. Bielova, and A. Legout, "In-depth technical and legal analysis of tracking on health related websites with ernie extension," in *WPES'21*, Online, Nov. 2021.

[26] Wikipedia, "List of hospitals in Canada," 2022, Online article (2012). https://en.wikipedia.org/wiki/List_of_hospitals_in_Canada.

[27] OpenWPM, "OpenWPM," 2022, Online article (2022). https://github.com/openwpm/OpenWPM.

[28] Pypi.org, "selenium-wire," 2022, online article (2022). https://pypi.org/project/selenium-wire/.

[29] Chrome DevTools, "Chrome devtools," 2022, online article (2022). https://developer.chrome.com/docs/devtools/.

[30] Europa.eu, "General data protection regulation," 2016, online article (1996). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[31] Hotjar, "How to install your hotjar tracking code," 2022, online article (2022). https://help.hotjar.com/hc/en-us/articles/115009336727-How-to-Install-your-Hotjar-Tracking-Code.

[32] Csoonline.com, "Researchers notice massive increase in malicious jquery libraries," 2014, online article (2014). https://www.csoonline.com/article/2136992/researchers-notice-massive-increase-in-malicious-jquery-libraries.html.

[33] Luke Leal, "Malicious javascript used in wp site/home url redirects," 2020, online article (2020). https://securityboulevard.com/2020/01/malicious-javascript-used-in-wp-site-home-url-redirects/.

[34] Sucuri blog, "jquery.min.php malware affects thousands of websites," 2015, online article (2015). https://blog.sucuri.net/2015/11/jquery-min-php-malware-affects-thousands-of-websites.html.

[35] Alibaba, "The best ecommerce advertising strategies for 2021," 2021, online article (2021). https://seller.alibaba.com/businessblogs/px001sb26-the-best-ecommerce-advertising-strategies-for-2021.

[36] S. Peisert, B. Schneier, H. Okhravi, F. Massacci, T. Benzel, C. Landwehr, M. Mannan, J. Mirkovic, A. Prakash, and J. Michael, "Perspectives on the SolarWinds incident," *IEEE Security & Privacy*, vol. 19, no. 02, pp. 7–13, mar 2021.

[37] U.S. Government, "Health insurance profitability and accountability act (HIPPA)," 1996, online article (1996). https://www.govinfo.gov/content/pkg/PLAW-104publ191/html/PLAW-104publ191.htm.

[38] SerpApi, "Baidu Organic Results API," 2022, Online article (2022). https://serpapi.com/baidu-organic-results.

Appendix

Malicious tracker domains. We list examples of malicious domains flagged by VirusTotal that are used to host known tracking scripts (see Table 4), and set known tracking cookies (see Table 5).

Sensitive information captured from session replay services. Table 6 lists the sensitive information captured from the Russian hospital websites using *Yandex* session replay service. Table 7 shows sensitive information captured from hospital websites in EU by *Yandex* and *Hotjar*.

Collecting hospital websites from China. We collect Chinese hospital websites from *Haodaifu* [12]. First, we crawl the list of hospital names from each of the 31 provinces in mainland China using [12]. Then we extract official names of these Chinese hospitals. These hospitals belong to different tiers (e.g., primary, secondary, tertiary). In order to determine the URL from the official names of these Chinese hospitals, we search each official name using the Baidu search engine. We observe that Baidu search results labels the official name of a hospital website (if exists) with two special Chinese characters — i.e., if a particular hospital does not have a website, Baidu search results will not label the official name of the hospital with the two special Chinese characters. Since the response from Baidu search results is not structured, it is not possible to mechanically parse the output. Therefore, we use the *Baidu Organic Results API* [38] to transform the search results to JSON format, and consider only the top 10 results to collect the hospital websites in mainland China.

Websites using HTTP and login forms. Hospital websites served over HTTP may allow an adversary to allow intercept sensitive information sent over the network traffic. We found 4062/19483 (20.8%) of hospital websites use HTTP. Some sites perform sensitive operations on these HTTP pages. For example, http://www.bfh.com.cn/Account/Register allows user registration functionality using HTTP. During user registration, the user is required to enter account information (user name, password) and other sensitive information (official name, national ID, mobile phone number, email, telephone number, province, city, marriage, home address, job, work address, MSN, QQ). Similarly, user registration information (user name, official name, password, national ID, mobile phone number and medical card ID) entered through http://www.zbdyyy.com/usersys/regist.aspx, is sent over HTTP, and can be intercepted by an adversary. We also found that the use of login forms in the landing page of hospital websites is mostly available in China (596/4324, 13.8%) and Australia (38/160, 23.7%), and some of these forms are submitted via HTTP. For example, 346/596 (58.1%) Chinese hospital websites with login forms sent login credentials in the clear — e.g., after clicking the top right button of hospital site http://www.ahs2y.com/, a login form is opened (http://111.39.250.98:7001/defaultroot/login.jsp); once the account name and password is entered and submitted, the credentials are sent over plain HTTP.

Category	Tracking domains	# hospital sites
Malware, malicious	iclickcdn.com, newrrb.bid, do-hero.com, wek7ipqx359.ru, 51.la, sc-static.net	120 (China, USA)
Malicious, phishing	ignorelist.com, popupsmart.com, leostop.com, popcash.net, secureservercdn.net	23 (USA, Chile, Malaysia)
Malware	che0.com, xc7789.top	4 (China, Spain)
Malicious	d10lpsik1i8c69.cloudfront.net, fontawesome.com, bitrix.info	138 (USA, Russia, France, Japan)

TABLE 4: Known tracking scripts hosted on potentially malicious domains that are flagged by VirusTotal. The countries within parenthesis in the 3rd column of the table are example location(s) of the hospital website(s).

Category	Tracking domains	# hospital sites
Malware, malicious, phishing	cnzz.space, crzenith.com,	2 (China, Sandi Arabia)
Malware, malicious	bedrapiona.com, medreviews.ru, informnikolase.live, 04zl.cn, greenklick.biz	85 (China, Mexico, Spain)
Malicious, phishing	onmarshtompor.com, clickmatters.biz	2 (Bulgaria, Spain)
Phishing	junmediadirect.com, 123formbuilder.com, app-us1.com	33 (USA, Australia, Belgium)
Malware	fontawesome.com, clarity.ms	124 (USA, Canada, Japan, United Kingdom, Portugal)
Malicious	sc-static.net	32 (USA, Saudi Arabia)

TABLE 5: Known tracking cookies set by potentially malicious domains that are flagged by VirusTotal. The countries within parenthesis in the 3rd column of the table are example location(s) of hospital website(s).

	Sensitive Information							Medical service information			
Site	**Name**	**Phone**	**Email**	**Password**	**DOB**	**Address**	**Passport**	**Specialist**	**Message**	**Clinic**	**Service Date**
alfa-med.ru	✓		✓			✓			✓		
bakulev.ru	✓	✓	✓	✓	✓	✓					
lk.baltclinic.ru	✓	✓	✓	✓		✓	✓				
solovevka.ru	✓		✓						✓		
smclinic.ru	✓	✓			✓			✓		✓	✓
rami-spb.ru	✓	✓	✓					✓			

TABLE 6: Examples of private/sensitive information collected by *Yandex* session replay service — DOB = Date of birth

SRS	Country	Site	Sensitive Information					Medical service information		
			Name	**Password**	**Email**	**Phone**	**Country**	**Specialist**	**Message**	**Chat/Search**
Hotjar	Croatia	bilicvision.hr	✓		✓	✓		✓		
	Czech Republic	reprofit.cz	✓		✓	✓	✓		✓	
	Italy	e-medical.it	✓	✓	✓					
	Portugal	sanfil.pt								✓
Yandex	Greece	lifeclinic.gr	✓		✓	✓			✓	
	Potgual	chpvvc.pt	✓		✓			✓	✓	
	Czech Republic	multiscan.cz								✓

TABLE 7: Examples of private/sensitive information collected by session replay service in EU Countries — SRS = Session replay service