Rauti *et al. J. Surveill. Secur. Saf.* **2025**, *6*, 1-16
DOI: 10.20517/jsss.2024.15

Journal of Surveillance, Security and Safety

Research Article

Open Access

Check for updates

Third-party data leaks on websites of medical condition support associations

Sampsa Rauti, Robin Carlsson, Panu Puhtila, Timi Heino, Tuomas Mäkilä, Ville Leppänen

Department of Computing, University of Turku, Turku 20500, Finland.

Correspondence to: Sampsa Rauti, Department of Computing, University of Turku, Vesilinnantie 5, 20500 Turku, Finland. E-mail: sjprau@utu.fi

How to cite this article: Rauti, S.; Carlsson, R.; Puhtila, P.; Heino, T.; Leppänen, V. Third-party data leaks on websites of medical condition support associations. *J. Surveill. Secur. Saf.* **2025**, *6*, 1-16. http://dx.doi.org/10.20517/jsss.2024.15

Received: 29 Jun 2024 **First Decision:** 29 Oct 2024 **Revised:** 15 Dec 2024 **Accepted:** 6 Jan 2025 **Published:** 22 Jan 2025

Academic Editor: Qiong Huang **Copy Editor:** Ting-Ting Hu **Production Editor:** Ting-Ting Hu

Abstract

The internet has become a primary source of health information for many people. For example, the websites of many medical condition support associations, meant for people suffering from various medical conditions, contain information on different medical conditions, treatments, and general health advice. However, accessing such information can be a serious privacy threat for the end user. In this article, we study the privacy of the websites of 18 Finnish medical condition support associations. The websites were analyzed to find leakages of sensitive personal data to third parties. Our investigation concludes that 88.9% of the websites leaked potentially sensitive personal data to third parties, usually private corporations offering web analytics tools such as Google Analytics. Furthermore, we discovered that users are not adequately informed about these data processing activities. We suggest several measures to alleviate third-party data leaks on websites handling sensitive personal data.

Keywords: Medical conditions, online privacy, support associations, data leaks, third parties, web analytics

1. INTRODUCTION

Websites of medical condition support associations (MCSAs) give essential resources and community support, and make information accessible to people struggling with various health challenges. These websites are places for sharing experiences, advice and coping strategies, and they give a sense of belonging to a community to their members. Information about medical conditions and possible treatment options is also provided on MCSA

websites[1,2]. This helps users to make informed decisions about their healthcare and treatments. Although the users may be far away from each other physically, MCSA websites strengthen the voices of people with medical conditions. The associations concentrate on awareness, advocacy, and peer support online, and act as intermediaries between several actors in the health sector[3].

MCSA websites offer valuable support and resources for their users; however, they also raise serious privacy concerns. Various third-party services collecting personal data, such as web analytics tools, are often an integral part of modern websites. The browsing behavior of users, such as visited pages and interactions with different elements and sections of the website, is being tracked by these services[4,5]. Such surveillance undermines the privacy of vulnerable individuals who seek online support for their medical conditions. If sensitive health data is sent to third parties without explicit consent, users' confidentiality and anonymity are compromised[6,7]. There is a risk that the shared data may be used for targeted advertising or handed over or sold to some "fourth party". The worst case scenario is that users may end up being exploited or discriminated against because their state of health has leaked[8].

In this study, we assess the privacy of 18 Finnish MCSA websites by analyzing their network traffic. More specifically, we examine the HTTP requests from the website to see what kind of personal data is sent to third-party services such as web analytics providers. In particular, we analyze *whether the network traffic sent to third parties contains data that can disclose the health state of a website visitor*. We also study the privacy policy documents of the websites to evaluate whether users are informed adequately about data processing and third parties involved.

The contributions of the current study are as follows. Firstly, to the best of our knowledge, there has not been research on third-party data leaks on MCSA websites before. While there have been studies on health-related websites generally, this study presents an in-depth multiple case study on MCSA websites and the nature of their data leaks specifically. This is important because third-sector websites are likely to differ from public and private sector health websites, as they are maintained by non-profit associations, often with limited resources, and the associations behind them may not always possess adequate technical expertise. Second, we examine transparency of privacy policies and the prevalence of dark patterns. Third, we also discuss implications of such data leaks for users of MCSA websites. Fourth, we offer recommendations for web developers and data protection officers to mitigate or prevent third-party data leaks. Finally, this study also gives insights about Finnish MCSA websites specifically, which have not been studied before regarding privacy.

The rest of the paper is structured as follows. In Section 2, we cover the previous research carried out on the topic of health-related third-party data leaks in the web environment. In Section 3, we explain the study setting and methodology used to collect the data and analyze it. In Section 4, we present the results of the network traffic analysis to reveal possible data leaks, as well as the examination of privacy policies and dark patterns. In Section 5, we discuss the key findings and implications drawn from our results. Finally, in Section 6, we present the definitive conclusions of this study.

2. RELATED WORK

While there is very little research specifically on the subject of formal MCSAs and privacy, there is a relatively large body of research done on the privacy issues of both online health communities (OHCs from this point onward) and general health-related websites. As the risks for user privacy in these kinds of online environments are similar to what we are studying in this paper, they are surveyed in this section to contextualize our own research interests and position. We present a review of research on OHCs in Section 2.1, while the research on general health-related websites is discussed in Section 2.2.

2.1. Privacy issues in online health communities

In contrast to formal associations, OHCs are usually categorized as more informal internet-based social networks that largely fulfill the same societal role as associations, i.e., offer peer support and general information on condition in question. The main difference between associations is that OHCs are unofficial gatherings of people, meaning that they are not in any way under the same legislation that governs the actions of registered associations, which may reflect in both the factual quality of their online content and the way they handle their users' personal data.

However, users of these communities in many instances communicate directly with doctors and other medical professionals, placing them in a grey area within the online health landscape. This situation is further complicated by the inconsistent use of the term OHC across the research papers we have surveyed, and it is obvious that some of the studied online instances in these papers are more official medical service providers than others. For example, some websites considered as OHCs can be systems for contacting doctors, while others can be discussion forums or even Facebook groups.

Dang *et al.*[9] investigated the correlation between the willingness of doctors to share medical information in OHCs and the degree of privacy protections deployed by the OHC. Their results validated a direct relation between these two factors, and a negative correlation when privacy matters were not addressed. In other words, doctors were more ready to share their professional knowledge in OHCs if they believed their privacy was protected.

Tseng *et al.*[10] studied various OHCs, and how their emphasis on privacy concerns affected user engagement. Their results showed that there was a direct correlation between user engagement with these platforms and how much they focused, or at least claimed to focus, on ensuring the user's privacy.

Avizohar *et al.*[11] conducted a survey on the users of Facebook medical support groups, in which they gauged what effect the group having a privacy policy had on the participants. Their results indicated that the users of these groups valued the group having a privacy policy highly, and that the existence of privacy policy made the users more open to sharing their experiences in these groups.

Yuchao *et al.*[12] conducted a study on what factors influence the perception of trust and thus users' willingness to share their personal data and information in OHCs. The chief observation among their findings was that the perceived trust for the particular websites and doctors as a profession was central to lessening privacy concerns. Other aspects that influenced the users' trust towards being more willing to share their health data online were the quality of received personalized healthcare services, as well as reciprocity norms that had a very strong impact on how much people cared about privacy matters.

Zhang *et al.*[13] studied the factors influencing user privacy concerns in OHCs and concluded that the efficacy of the responses and perceived self-efficacy in addressing the issue had a negative effect on privacy concerns; in other words, they lessened the concerns over privacy, while perception of potential threats and their severity positively influenced privacy concerns. Zhu *et al.*[14] conducted research on privacy leakages in the OHC forums for breast cancer patients, using AI text mining to identify the topics users were talking about and the personal information they disclosed. The largest category of leaked information was Emotional Feelings, followed by Detailed Medical Data. However, it should be understood that privacy leakages in this sense are not exactly the same as those studied in our paper, as the leakages Zhu *et al.* examined are literally things directly revealed by the website users in their written outputs, whereas the privacy leakages we are studying are data items transmitted by the web analytics tools.

Feng *et al.*[15] studied the effect of outcome and procedural fairness on the willingness of patients to disclose

their private information and otherwise engage with the online medical community platforms. Their results supported their hypothesis that the more fair both of these factors were perceived by the patients, the more willing the patients were to disclose private matters.

2.2. Privacy issues on medical websites

Masters [16] studied the phenomenon of web analytics tools used in medical websites in his pioneering research in 2012, in which he examined how many of these applications collect user data without informing them properly. Huesch conducted research in relation to searching for medical information [17], in which he arrived at similar conclusions. Brown and Levy [18] developed a tool to benchmark the actual data collection happening at the medical websites, as opposed to the documented data collection.

Burkell and Fortier [19,20] addressed in their two separate research papers the issue of medical websites incorrectly disclosing their data collection practices. Their results showed that this led both the website users to give consent to data collection against their better knowledge or interests, and that medical websites actively collected the user data to create distinct profiles of them.

Surani *et al.* [21] studied online mental health services and concluded that many had deficiencies when it came to user privacy. Zheutlin *et al.* [7] inspected how various health services operating in the USA, in both private and public sectors, tracked their users extensively with web analytics tools. Friedman *et al.* [22] conducted research on how the use of third-party analytics tools might put hospitals in legal jeopardy, by endangering user privacy. Yu *et al.* [23] conducted a large-scale automated survey of hospital websites across the world and concluded that as many as 53.5% of them used web analytics tools.

Friedman *et al.* [24] researched the use of analytics tools at abortion clinic websites, and how they collected user data. Their findings showed that 99.1% of these websites used web analytics tools that leaked user data to third parties. Huo *et al.* [25] studied the privacy violations of patient portal websites, revealing that 14% of them leaked data that directly identified the users of these services, including phone numbers and names. Schnell and Kaushik [26] wrote a paper on how the design of hospital websites made it hard to locate the privacy policy, and thus to legitimately consent to the data collection happening at these websites. Wesselkamp *et al.* [27] developed a browser extension that could detect the use of tracking cookies on websites. They used this tool to analyze 385 medical websites within the EU jurisdiction, and concluded that not only 62% of them used tracking cookies before the consent to cookies was even given, but that as many as 15% used them regardless of whether the user consented or not.

3. STUDY SETTING AND METHODOLOGY

3.1. Website selection

In the current study, we examined the websites of officially registered MCSAs operating in Finland. These associations are mainly volunteer-operated organizations, each of which represents the community of people who suffer from some specific long-term or chronic medical condition, such as diabetes, heart disease, obesity or cancer. MCSAs run by volunteers have an important role in the Finnish healthcare support system. They offer peer support, resources, and guidance for individuals with chronic or long-term conditions. Such organizations also bridge the gap between patients and healthcare providers, give patients personalized assistance and raise general awareness about medical conditions. The associations meet the unique needs of their patient groups, and have significant effects on patient empowerment, quality of care, and policy changes. This benefits the whole healthcare support system.

The associations to be included in the study were selected by performing extensive online searches with the Google search engine. To search for MSCAs, we used the following search queries:

- "(condition name) + association"
- "(condition name) + peer support"
- "medical conditions"

The searches were made in Finnish and the following medical conditions were included in the searches: neurological disorders, memory disorders, autism, ADHD, mental illnesses, psychotic disorders, intellectual disabilities, substance use disorders, cancer, heart diseases, diabetes, physical disabilities, movement disorders, and HIV. More associations were collected by generally searching for MCSAs. A website of an association was included if it was clearly identified as a support association for individuals with medical conditions, or if the website clearly offered support, guidance, and information for people with medical conditions or their families.

The list of associations was compiled in this way since no official listing for these kinds of associations exists in Finland. The number of the MCSAs selected for the study was 18. Because we primarily study data leaks as a phenomenon here rather than specific associations, we believe it is ethical not to refer to the chosen MCSAs by their actual names. Therefore, the analyzed websites are referred to using pseudonyms MCSA1–MCSA18.

3.2. Network traffic analysis

The third-party data leaks on the MCSA websites were studied in the following way: The researcher navigated to the website and opened Google Chrome Developer Tools (DevTools from this point onward). The cache was disabled to ensure that no cached data would interfere with our experiments. Then, the researcher proceeded to navigate through the website in a pattern that consisted of using the search functionality and clicking the link that leads to the "join association" page. The DevTools were set to record all network data traffic, which was saved in to .HAR[1] log files, which were analyzed to determine whether leaks of personal data occurred.

The stored .HAR files were further studied with DevTools. Each third-party HTTP request was reviewed using DevTools one by one. We carefully examined the header information and payload of every request. We conducted a thorough manual analysis to find all personal data from the request. For example, URL addresses, search terms, and events (e.g., the user presses a button) contained in the requests were examined carefully, as these are some usual ways of leaking personal data. We looked at both identifying personal data such as device identifiers and contextual personal data such as health-related tied to a specific individual. In this study, we were specifically interested in three types of data leaks:

1. **Information page URL:**
   Leaks of the website URL (the domain or the front page of the association) indicate interest in the association in question, which, combined with identifying factors, can jeopardize the privacy of the user and imply an interest in a specific medical condition.

2. **"Join association" page URL:**
   Leaks of the "Join association" page URL indicate interest in joining the association in question, which is potentially even more hazardous from the perspective of privacy, as a connection to a specific medical condition may be revealed.

3. **Search term:**
   Data leaks containing search field inputs may also reveal detailed and sensitive personal data about the user (e.g., medical conditions or symptoms). This is especially problematic in terms of privacy because the user can freely decide the contents of the search terms, potentially disclosing sensitive information to third

[1]The HTTP Archive file logs a web browser's interaction with a website in JSON-format

parties.

Paying special attention to contextual data described above, such as sensitive URLs and search terms, we carefully listed and documented all instances of personal data. We especially concentrated on data concerning health, as this was the focus of our study. It is also mentioned as special category data requiring special protection in the GDPR (General Data Protection Regulation). Any piece of data that could reveal a connection between an individual and a medical condition was recorded. We also carefully listed all the third parties that received HTTP requests containing personal data.

As this work focuses on studying leaks of personal data, we must define what "personal data" means in this context. The definition in the GDPR of the European Union, also used by the Finnish Office of the Data Protection Ombudsman, fits for the purposes of the current study[2]. Here, "personal data" is given as "all data related to an identified or identifiable person". By this definition, all technical data such as device identifiers, IP addresses, location data or any data item that can be used in the identification of the website user is "personal data". It is also important to note that while the majority of such data items are not by themselves enough to uniquely identify someone, in combination with other such items they can be used to create a digital fingerprint of the user in question. For this reason, they are also treated as "personal data".

3.3. Privacy policies and dark patterns

The privacy policy documents of the websites were collected. The contents of these documents were analyzed to determine whether they correctly informed the user of the website about the data collection taking place. More specifically, we examined whether they mentioned all third parties to whom personal data was being leaked and all the categories of personal data items that were collected. We also investigated whether sharing information concerning health with third parties was mentioned in privacy policies. The studied documents included general privacy policies, cookie policies, and cookie banners. The claims of the privacy policies were then contrasted with the actual data collection taking place at the websites.

We also examined the use of dark patterns on these websites. This study was based on the European Data Protection Board's Cookie Banner Taskforce's "Report of the work done"[3], which defines a list of commonly occurring dark patterns in website designs. For our purposes, we chose four of these patterns to focus on:

- *Absence of rejection button from the first layer of the cookie consent banner:* This means whether the layer first accessible to the user of the cookie consent banner has a button to reject data collection or not. If the button exists in some subsequent layer or does not exist at all, it is considered to be a dark pattern.
- *Pre-ticked consent boxes:* The use of pre-ticked consent boxes means a situation where the cookie consent banner offers the user options for what kind of cookies they want to accept, and that some of these options are turned on by default.
- *Use of deceptive colors or contrasts:* In this context, the deceptive color means a color that is commonly used in visual communication to imply either positive or negative things, such as green or red, and deployment of such colors to mislead the user. A very common example is coloring the "consent" button green or blue on websites, or the "reject" button red. The deceptive contrast, on the other hand, means the use of two colors or two shades of the same color in the cookie consent banner to mislead the user. A common example of such is coloring the "consent" button bright color while leaving the "reject" button colorless.

This choice was informed by the observation that these four are very clearly defined in the aforementioned document and easy to detect on websites, while many other dark patterns described in the aforementioned

[2]https://gdpr-info.eu/

[3]https://edpb.europa.eu/our-work-tools/our-documents/other/report-work-undertaken-cookie-banner-taskforce_en

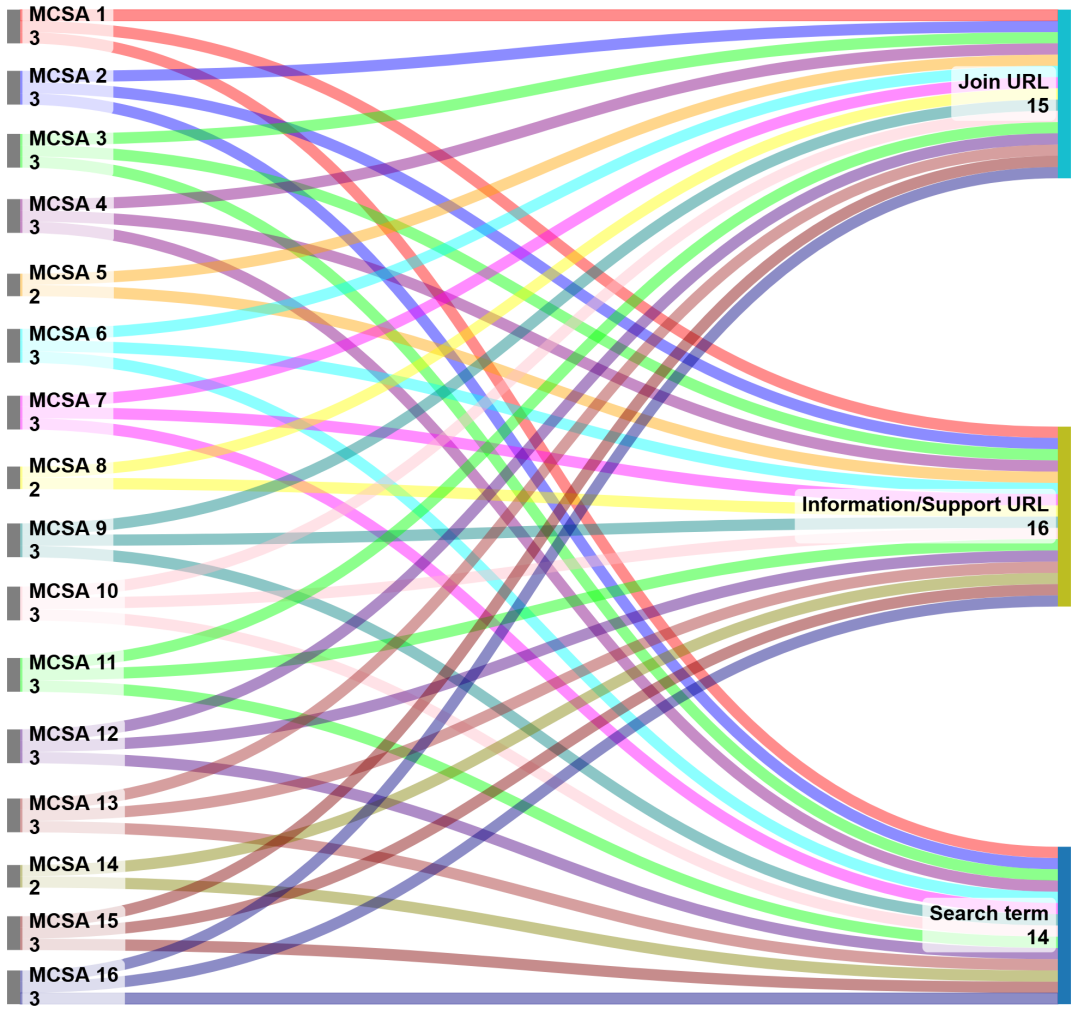

Figure 1. Sensitive data categories leaked by the MCSA websites. MCSA: medical condition support association.

report are more ambiguous and thus harder to detect and definitely prove.

4. RESULTS

4.1. Network traffic analysis

As can be seen in Figure 1, 16/18 (88.9%) of the studied MCSA websites leaked potentially sensitive data to third parties. The two websites that did not leak the user data did not also have any third-party analytics deployed at the site.

Of the specific leak categories we concentrated on, URL addresses of the visited web pages were the most leaked data item. All 16 websites that had any third-party web analytics in use leaked the URL address, in addition to other leaks. On MCSA websites, visited pages can imply, for example, that the user is looking for information or support for a specific condition or related topic. Therefore, the leaks of visited pages are labeled as "Information/Support URL" leaks in Figure 1. It is also worth noting that even if the user does not navigate to a specific subpage, their visit to the front page is recorded, which already implies some kind of interest in the association.

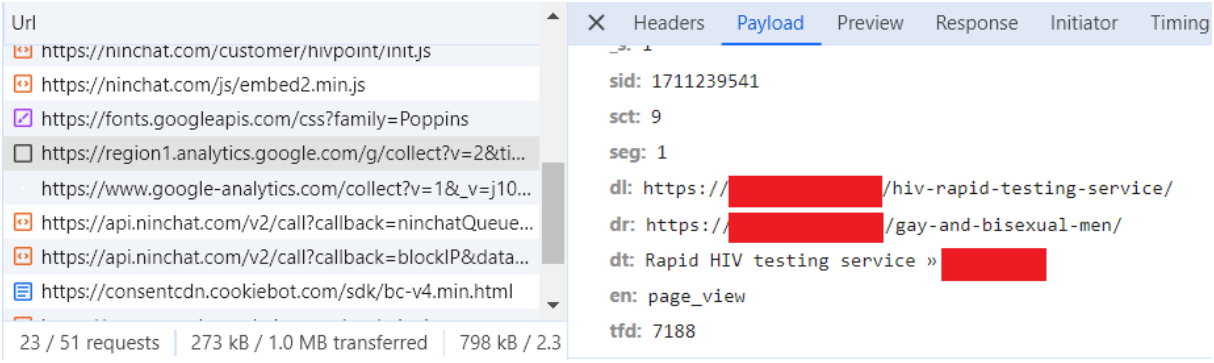

Figure 2. An example of an HTTP request payload sent to Google on a website of a support association for HIV patients.

On 14/18 (77.8%) websites, the search terms user input also leaked as a part of a visited URL. This is significant because the user freely decides the contents of these searches, and the information they contain can be highly personal and confidential, such as names of diseases or symptoms. Search terms can also give clues about the user participating in the association's activities. Leaking this kind of information can thus be highly injurious to the social standing of the website end user.

Because joining an association is also very critical in the sense that it implies that the user somehow has a strong connection to a specific medical condition, we also examined "Join association" pages. If a "Join association" page was unavailable, we examined another page implying a strong connection to the association (e.g., pages for signing up for peer support groups or specific courses). Information about navigating to these "Join association" pages, that is, the page URL, was leaked in 15/18 (83.3%) cases. Information of this kind obviously can be used to profile the website user as someone who likely suffers from the disease in question.

In addition to the leak categories in Figure 1, there were individual cases in which the data transmitted to third parties was found to be highly sensitive. Figure 2 shows one such example in network traffic of a website of a support association for HIV patients. We can see that the current page being tracked (field `dl`) is about HIV rapid testing service. What is more, the user is arriving at this page from one targeted at gay and bisexual men (field `dr`). Therefore, there is a high probability that the user is a man belonging to a sexual minority and interested in HIV testing. This information is collected by Google Analytics and sent to Google. Furthermore, we found out that information of the user pressing a link leading to HIV test appointment booking was also leaked to Google (this leak is not shown in the figure). Information concerning such medical treatments is generally considered to be data concerning health in the sense of the GDPR. This health data is considered a special category of personal data, which requires higher protection due to its sensitive nature. In the example, highly sensitive information about the user's medical procedure and sexual orientation can be shared with Google without the user realizing this.

Finally, Figure 3 shows the third parties that received contextual personal data from the studied websites. It is not surprising that Google's and Meta's third-party tools are the most popular and widely used services. Google receives sensitive personal data on 83.3% (15/18) and Meta on 33.3% (6/18) of the studied websites. Other third parties to which data leaked included React & Share, a real-time content analytics service; ShareThis, a content sharing tool; ActiveCampaign, a tool for customer experience automation; LinkedIn, a professional networking platform; Pingdom, a web performance monitoring solution; Serviceform, a service providing online forms and bookings; and SiteImprove, a digital optimization platform.

One reason the data leaks we have discovered are serious is the ability of big tech companies to identify and profile individual users very effectively. It is also worth noting that by employing tracking cookies, big tech

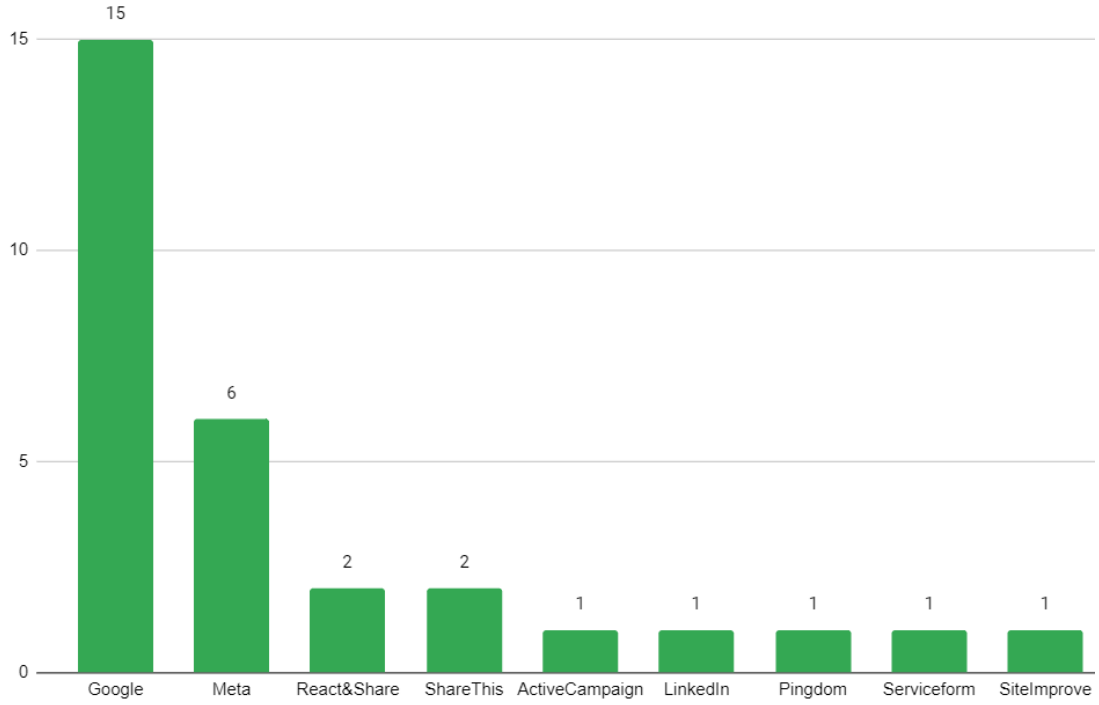

Figure 3. Third parties receiving data leaks on the studied MCSA websites. MCSA: medical condition support association.

companies such as Google and Meta can link the actions the user has taken on various websites to their Google or Facebook accounts. Therefore, the user's behavior online can often be connected to their real name. For instance, Google Analytics makes use of a tracking cookie containing a unique identifier, client ID (cid), which is assigned to each individual browser-device pair. The tracking cookies used by Google Analytics enable Google to distinguish users from one another and are stored on users' devices for two years[4].

While MCSA websites have not been studied in terms of data leaks, our results align with earlier research, revealing extensive data leaks. For example, Yu *et al.* [23] found in their study that 53.5% of the hospital websites included web analytics, and Huo *et al.* [25] reported that 14% of patient portals included Google Analytics. Although the numbers in these studies are lower than in our research, all findings show the problem of third-party data leaks is widespread. In associations often run by volunteers, website maintainers are likely to pay less attention to privacy issues caused by third parties than hospitals and official health portals. The earlier studies also use automatic analysis which may not always recognize all personal data leaks and third parties which our manual method does.

4.2. Privacy policies and dark patterns

Table 1 shows the collected privacy policies. None of the MCSAs had every kind of privacy document available. While some had a privacy policy, a cookie banner, and a more accurate cookie policy, others lacked one or two of these documents. In such cases, the disclosures of used third-party services had to be evaluated only by analyzing the available document(s). On some occasions, a privacy policy covered both general data protection within the organization and its cookie policy on the web. Such sites are considered to have both policies.

Table 2 displays the third-party services receiving URL addresses observed in our network traffic analysis and

[4]https://policies.google.com/technologies/cookies?hl=en-US

Table 1. The analyzed privacy policies, cookie policies, and cookie banners

Website	Privacy policy	Cookie policy	Cookie banner
MCSA 1		X	X
MCSA 2	X	X	X
MCSA 3			X
MCSA 4	X		
MCSA 5	X	X	
MCSA 6	X	X	X
MCSA 7			X
MCSA 8			
MCSA 9	X	X	X
MCSA 10	X	X	X
MCSA 11	X	X	X
MCSA 12	X	X	X
MCSA 13			X
MCSA 14	X		X
MCSA 15	X		X
MCSA 16	X	X	X

Table 2. The observed third parties receiving URL leaks and their categorizations in the cookie policies. The cells highlighted with bold and italic indicate that the tracker was observed in our analysis, but not mentioned in the policies. The highlighted GA (Google Analytics) within MCSA 7 means the MCSA informed that their non-necessary cookies are used for analytics, but they did not disclose the use of GA. In the cases of the highlighted cells in MCSAs 13 and 14, the privacy policies reported the use of Google Tag Manager (GTM) for analytics and functionalities instead of GA. While they did not disclose the use of GA specifically, it is common to use GTM to manage GA

Website	Necessary	Functional	Analytics	Marketing	Unclassified or vague
MCSA 1			GA	Meta Pixel	
MCSA 2			GA		
MCSA 3			GA	GA	RUM Collector
MCSA 4					***GA, Meta Pixel, Appspot***
MCSA 5			GA		
MCSA 6	Siteimprove-analytics		GA, Siteimprove-analytics	GA, YouTube, Meta Pixel	
MCSA 7			***GA***		
MCSA 8					***Sharethis***
MCSA 9		React & Share	GA, YouTube, Serviceform	Meta Pixel, Active-Campaign	
MCSA 10					***GA, YouTube, Meta Pixel, Sharethis***
MCSA 11			GA, YouTube	GA, Meta Pixel, LinkedIn Ads	
MCSA 12	YouTube		GA, React & Share	Google Ads, Doubleclick	React & Share
MCSA 13	YouTube	***GA***	***GA***		
MCSA 14		***GA***	***GA***		
MCSA 15			GA	GA	
MCSA 16			GA		***YouTube***

the justifications for their use in the collected policy documents. The cells highlighted in bold and italics mean a third party was observed in our analysis but not mentioned in the policy. The justifications can be divided into five categories. First, there are necessary cookies that the website allegedly needs to function. Usually, these cookies are related to authentication or session management, for example. Second, functional cookies are used to improve the functionality of a website and provide features that improve user experience, such as remembering user preferences. Third, analytics cookies collect data about website usage and visitor interactions. Fourth, marketing cookies are used for targeted advertising based on the browsing behavior and interests of the website visitors. Fifth, the justifications for the use of many third-party services were not mentioned or they were stated very vaguely.

Examining the third parties in Table 2 by category, the necessity of some is very questionable. Siteimprove

Table 3. Dark patterns on cookie banners of the studied medical condition support association websites

Website	No consent asked	No reject button on the first layer	Pre-ticked consent boxes	Deceptive colors/contrast
MCSA 1				X
MCSA 2				
MCSA 3				X
MCSA 4	X	N/A	N/A	N/A
MCSA 5	X	N/A	N/A	N/A
MCSA 6				X
MCSA 7		X		X
MCSA 8	X	N/A	N/A	N/A
MCSA 9				X
MCSA 10			N/A	X
MCSA 11				
MCSA 12				X
MCSA 13				
MCSA 14				
MCSA 15		X		X
MCSA 16				X

Analytics is a service used to monitor performance and improve user experience, but it is hardly strictly necessary. Similarly, YouTube is not required for any website to function, despite its justification in some cases. In the functional category, the inclusion of Google Analytics is debatable, because it serves more as a usage measurement tool rather than a service directly enhancing user experience. In the analytics category, Google is very prevalent with 13 occurrences. It is generally categorized correctly as an analytics service, with only one omission.

The marketing category for services geared towards targeted advertising is dominated by Google and Facebook, which is not surprising. What is surprising, however, is that MCSA websites have third parties and cookies meant for marketing in the first place. These are websites for non-profit third-sector associations in a highly sensitive field, after all. Finally, we can see that several third parties on many MCSA websites have been left unclassified or are justified very vaguely. This includes Google and Facebook on two occasions. It is obviously always bad for the user if they are not being informed adequately. Vague classifications also likely reflect the confusion of website maintainers about the real purpose of third-party services on their websites.

Overall, these findings on third-party and cookie classification indicate that one service can be interpreted to belong to very different categories, confusing users. Although one third-party service sometimes naturally has many purposes, some of the justifications used in privacy policy documents are still very misleading.

Privacy policies were also reviewed to determine whether they mentioned the collection of data that may reveal details on a user's health. It is not very surprising that there were no mentions of sensitive health data. This is because the developers and data protection officers most likely have not thought about the possibility of such data leaks and the personal data revealed through URLs and HTTP payloads sent to third parties in general. Still, the fact remains that the privacy policies were not transparent, and by reading these documents, the user has no possibility of finding out that their health-related data can leak.

Finally, Table 3 shows the dark patterns observed on cookie consent banners of those websites that leaked personal data to third parties. It also shows whether the website asked for consent for cookies and data collection. The empty cells indicate a positive outcome; that is, consent was asked or a dark pattern was not present on a cookie banner. The X marking indicates a negative outcome; that is, consent was not asked or a dark pattern was present. The "N/A" marking means that the dark pattern is not applicable, because there is no cookie banner or the cookie banner contains no checkboxes.

Three websites did not ask for consent at all. The most common dark pattern was the use of deceptive colors or

contrasts with nine occurrences, which is 50% of the websites. This practice tricks users into accepting cookies and data collection by making the "accept" button more prominent than the "reject" button, undermining informed consent. On two occasions, there was no reject button on the first layer of the cookies consent banner, making it harder for the user to reject cookies. On a positive note, there were no pre-ticked boxes, which would violate GDPR's requirement for explicit consent, as they assume consent without an active decision from the user.

Our findings on informing users and consent practices align with many previous studies. Previous studies also show that the fact health data is shared with third parties is often not properly divulged to the user[17,21] and the language in privacy policies is difficult to understand[20]. Krisam *et al.*[28] found that over 85% of the top German websites used visually distracting factors to manipulate the user into consenting to cookies. Simply collecting data without proper consent is not rare either[29,30].

5. DISCUSSION

5.1. Key findings

The key findings of the current study can be summarized as follows:

- Almost 90 % (16/18) of the studied websites leaked potentially sensitive personal data such as visited pages or health-related search terms to third parties. Not surprisingly, Google was the most prevalent third party with (13/18) occurrences.
- On over 80 % (15/18) of the analyzed websites, the data on accessing the "join association" page or a similar page, which strongly implied the user's intended involvement with the association, was leaked to at least one third party.
- Some very sensitive personal data, such as booking an appointment for an HIV test, also leaked to third parties. This highlights that the privacy threat posed by third-party analytics has not been considered by MCSA websites.
- In many cases, the privacy policies failed to sufficiently list third parties and categorize them correctly, causing the user not to be informed adequately of the data processing activities.
- Several dark patterns were found on the cookie consent banners, making users more susceptible to inadvertently agreeing to data collection without fully understanding the implications. Particularly, deceptive colors of the "accept cookies" button were present on 50 % (9/18) of the studied websites.

5.2. Implications for users

A leak of sensitive personal data always violates the user's privacy. Such an incident can cause a sense of betrayal, if the user's sensitive health-related data is shared without consent. It can also result in a loss of trust in the service in question[31,32]. Users may also feel discouraged from using online health resources, seeking health information online or contacting MCSAs. A negative experience with one online service can also reduce users' trust in digital health services in general, especially when these services process sensitive personal data.

It is possible for third-party companies to build accurate profiles on users based on their medical conditions, their treatments and their potential participation in a support association's activities. The health data a third party has collected can then be used to expose users to targeted advertising[33]. In the worst case, users could receive advertisements based on their medical conditions, which may cause distress or discomfort.

If the user's sensitive data concerning health is revealed to wrong third parties and people, users can become stigmatized or discriminated against because of their medical conditions and state of health[34]. The data on the user's state of health may be misused by various parties such as employers, peers, or insurance companies, leading to unfair assessments or decisions.

Sharing users' sensitive health data with third parties without consent, especially when discovered by users, can significantly affect their mental health. Vulnerable users may become even more at risk due to non-transparent data-sharing practices.

It is important for providers of web-based health services to remember that health data is highly sensitive and extremely valuable to many third parties. Data concerning health differs from many other sensitive data in the sense that many parts of it never change – its exposure can have negative consequences for users when shared without explicit consent [35].

5.3. Implications for web developers and data protection officers

The problem with modern web analytics is that they usually collect data on their own servers. Based on our findings, it is evident that web developers and data protection officers need to be more vigilant about the personal data transmitted to third parties from their websites. Modeling outgoing data flows for critical parts of the website and using a network traffic analysis approach similar to the one employed in this study should be integral parts of the website development process.

To effectively manage risks, a deep understanding of the third-party services present on the website is necessary. It is important to understand the potential risks that come with using these third parties, and to make sure they are transparent and comply with privacy regulations such as GDPR. When third-party services are selected, assessing them in terms of privacy and thoroughly reviewing them is important. The number of third parties should be minimized based on this careful review. Carefully justifying the inclusion of every third party and documenting this design decision is also essential. Different third-party alternatives should be explored carefully, and third-party services with a high level of data protection should be chosen. For some critical purposes, even a code review of third-party services may be necessary to get a good understanding of its functionality.

Getting rid of third-party services altogether is the best option on critical websites. Third-party tools should be avoided especially on pages that handle sensitive personal data concerning health. If web analytics are deemed strictly necessary, the data they have collected can be stored locally on secure servers (hosted either by the association or some trusted party). The idea is to ensure that the association retains control over data and no third party is able to access it. There are several free open-source solutions that can be used to achieve this, such as Matomo [36,37].

The configurations of third-party analytics often pose a problem. For example, Google Analytics records events, such as pressing buttons, and sensitive user actions such as booking an appointment. Additionally, if the web analytics tool is not properly configured to anonymize the user, their actions can often be connected to their identity. However, on many occasions, such privacy-friendly configurations are not enabled by default and must be implemented programmatically, which website maintainers may find difficult [37].

Nowadays, the use of third-party analytics services has become widespread in electronic commerce. These services track and measure conversions, such as completed sign-ups and purchases. This information gives valuable insights for marketing strategies and website optimization. However, it is hard to understand why developers of MCSA websites would want to include these features. Non-profit associations should not need marketing insights and usability and performance can be monitored with local analytics solutions without involving third parties. It is likely that MCSA website maintainers include analytics without fully understanding the potential privacy risks associated with these services and the processing of sensitive data.

Often, third-party services may be added by external developers who do not fully understand the application area and do not realize the risks of highly sensitive personal data leaking on MCSA websites. Respectively,

most MCSAs likely lack the resources or technical expertise to evaluate the potential downsides of third-party web analytics. For association websites handling sensitive data that may reveal a user's health status, an external privacy audit would be advisable. Web developers should also familiarize themselves with the domain they are working with in order to understand the basic risks associated with processing sensitive health data. Successfully communicating with the stakeholders and domain experts from the association is also very important. This will also help assess the privacy requirements of the specific parts of the service, evaluate the risks involved and take the necessary precautions.

5.4. Limitations

The current study focuses specifically on Finnish MCSA websites, which can limit the generalizability of our findings. The level of privacy and the prevalence of data leaks may vary on similar websites in different jurisdictions. Most countries outside the European Union do not have privacy regulations as strict as the GDPR. Weaker privacy requirements may lead to higher numbers of third-party data leaks. However, privacy regulation varies even within the EU, as the GDPR only sets minimal requirements, and regulations may be stricter due to more rigorous enforcement or heightened cultural sensitivity to privacy matters. Germany is one well-known example of a country with strict privacy laws[38], and would most likely fare well in comparison with other EU countries when it comes to data leaks and privacy flaws on websites. Further research could study these variations to better understand privacy practices in various regulatory and cultural contexts.

Our analysis also has some technical limitations. The navigation was limited to the cases of using the search functionality, accessing information pages and visiting the "join association" page. Therefore, the analysis of a specific website is not exhaustive. Also, some network traffic may be encoded or encrypted, possibly leaving some data leaks out of our analysis. Additionally, our analysis is done on the client side. We cannot observe how the third parties use the data they have collected or whether they use it at all. Lastly, we analyzed a limited set of websites, and the findings can not be fully generalized.

6. CONCLUSIONS

As an important source of health information, MCSAs have an obligation to protect the privacy of the user of their web service. This issue has not been adequately addressed on the MCSA websites we studied. Our findings can hopefully remind web developers and data protection officers responsible for web services processing sensitive data about the importance of protecting users' privacy. In association websites and peer support groups created with limited resources and expertise, the privacy-by-design approach can be easily forgotten.

Many people using these kinds of services are already in a vulnerable position to begin with, which makes privacy even more essential. On MCSA websites associated with serious health conditions, it is difficult to justify the use of any third-party service that may collect sensitive personal data and potentially even send it outside Europe without the user's consent. The goal should be to build web services to be as trustworthy and confidential as other offline processes and activities of the associations. Future work could involve studying user-centered design approaches to strengthen privacy on MCSA websites, particularly exploring solutions for improving the user's control over their personal data and informing them on the data processing activities in a transparent manner.

DECLARATIONS

Authors' contributions

Designed the study idea and methodology, also contributed to data analysis and interpretation, and wrote the majority of the manuscript: Rauti, S.
Collected, analyzed and interpreted the data: Carlsson, R.; Heino, T.

Contributed to data analysis and interpretation, conducted the literature review, and wrote the article: Puhtila, P.
Involved in interpreting the results and providing critical revisions: Mäkilä, T.
Interpreted the results, critically revised the article, and acquired the funding: Leppänen, V.
All authors reviewed and approved the final version of the manuscript.

Availability of data and materials
Data is available upon request.

Financial support and sponsorship
This research has been funded by the Academy of Finland project 327397, IDA – Intimacy in Data-Driven Culture.

Conflicts of interest
All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

REFERENCES

1. Soobrah, R.; Clark, S. Your patient information website: how good is it? *Colorectal. Dis.* **2012**, *14*, e90–94. DOI
2. Jadad, A. R.; Gagliardi A. Rating health information on the Internet: navigating to knowledge or to Babel? *JAMA.* **1998**, *279*, 611–14. DOI
3. Pais, S. C.; Guedes, M.; Menezes, I. The values of empowerment and citizenship and the experience of children and adolescents with a chronic disease. *Citiz. Soc. Econ. Educ.* **2012**, *11*, 133–44. DOI
4. Carlsson, R.; Rauti, S.; Heino, T. Data leaks to third parties in web services for vulnerable groups. In: 2023 46th MIPRO ICT and Electronics Convention (MIPRO). IEEE; 2023. pp. 1208–12. DOI
5. Huidobro, A.; Monroy, R.; Cervantes, B. A High-level representation of the navigation behavior of website visitors. *Appl. Sci.* **2022**, *12*, 6711. DOI
6. Samarasinghe, N.; Adhikari, A.; Mannan, M.; Youssef, A. Et tu, brute? privacy analysis of government websites and mobile apps. In: Proceedings of the ACM Web Conference 2022. ACM; 2022. pp. 564–75. DOI
7. Zheutlin, A. R.; Niforatos, J. D.; Sussman, J. B. Data-tracking on government, non-profit, and commercial health-related websites. *J. Gen. Intern. Med.* **2022**, *37*, 1315-7. DOI
8. Libert, T. Privacy implications of health information seeking on the web. *Commun. ACM.* **2015**, *58*, 68–77. DOI
9. Dang, Y.; Guo, S.; Guo, X.; Vogel, D. Privacy protection in online health communities: natural experimental empirical study. *J. Med. Internet Res.* **2020**, *22*, e16246. DOI
10. Tseng, H. T.; Ibrahim, F.; Hajli, N.; Nisar, T. M.; Shabbir, H. Effect of privacy concerns and engagement on social support behaviour in online health community platforms. *Technol. Forecast. Soc. Change* **2022**, *178*, 121592. DOI
11. Avizohar, C.; Gazit, T.; Aharony, N. Facebook medical support groups: the communication privacy management perspective. *Aslib J. Inf. Manag.* **2023**, *75*, 664–84. DOI
12. Yuchao, W.; Ying, Z.; Liao, Z. Health privacy information self-disclosure in online health community. *Front. Public Health* **2020**, *8*, 602792. Available from: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2020.602792. DOI
13. Zhang. X.; Liu, S.; Chen, X.; et al. Health information privacy concerns, antecedents, and information disclosure intention in online health communities. *Inf. Manag.* **2018**, *55*, 482–93. DOI
14. Zhu, Y.; Tong, X.; Wang, X. Identifying privacy leakage from user-generated content in an online health community-a deep learning approach. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2019. pp. 1–2. DOI
15. Feng, C. L.; Cheng, Z. C.; Huang, L. J. An Investigation into patient privacy disclosure in online medical platforms. *IEEE Access* **2019**, *7*, 29085–95. DOI
16. Masters, K. The gathering of user data by national Medical Association websites. *Internet J. Med. Informat.* **2012**, 6. Available from:

https://ispub.com/IJMI/6/2/14386.
17. Huesch, M. D. Privacy threats when seeking online health information. *JAMA Intern. Med.* **2013**, *173*, 1838–40. DOI
18. Brown, S. D.; Levy, Y. Towards a development of an index to measure pharmaceutical companies' online privacy practices. *Online J. Appl. Knowledge Manag.* **2013**, *1*, 93–108. Available from: https://www.iiakm.org/ojakm/articles/2013/OJAKM_Volume1_1pp93-108.php.
19. Burkell, J.; Fortier, A. Consumer health websites and behavioural tracking. In: Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI; 2012. . DOI
20. Burkell, J.; Fortier, A. Privacy policy disclosures of behavioural tracking on consumer health websites. In: Proceedings of the American Society for Information Science and Technology. Wiley Online Library; 2013. pp. 1–9. DOI
21. Surani, A.; Bawaked, A.; Wheeler, M.; et al. Security and Privacy of Digital Mental Health: An Analysis of Web Services and Mobile Applications. In: IFIP Annual Conference on Data and Applications Security and Privacy. Springer; 2023. pp. 319–38. DOI
22. Friedman, A. B.; Merchant, R. M.; Maley, A.; et al. Widespread Third-Party Tracking On Hospital Websites Poses Privacy Risks For Patients And Legal Liability For Hospitals. *Health Aff.* **2023**, *42*, 508–15. DOI
23. Yu, X.; Samarasinghe, N.; Mannan, M.; Youssef, A. Got sick and tracked: privacy analysis of hospital websites. In: 2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). IEEE; 2022. pp. 278–86. DOI
24. Friedman, A. B.; Bauer, L.; Gonzales, R.; McCoy, M. S. Prevalence of third-party tracking on abortion clinic web pages. *JAMA Intern. Med.* **2022**, *182*, 1221–22. DOI
25. Huo, M.; Bland, M.; Levchenko, K. All eyes on me: inside third party trackers' exfiltration of PHI from healthcare providers' online systems. In: Proceedings of the 21st Workshop on Privacy in the Electronic Society. WPES'22. New York, NY, USA: Association for Computing Machinery; 2022. p. 197–211. DOI
26. Schnell, K.; Kaushik, R. Hunting for the privacy policy - hospital website design; 2022. Available at SSRN: https://ssrn.com/abstract=406844. DOI
27. Wesselkamp, V.; Fouad, I.; Santos, C.; et al. In-depth technical and legal analysis of tracking on health related websites with ERNIE extension. In: Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society. WPES '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 151–166. DOI
28. Krisam, C.; Dietmann, H.; Volkamer, M.; Kulyk O. Dark patterns in the wild: Review of cookie disclaimer designs on top 500 German websites. In: Proceedings of the 2021 European Symposium on Usable Security; 2021. pp. 1–8. DOI
29. Hils, M.; Woods, D. W.; Böhme, R. Measuring the emergence of consent management on the Web. In: Proceedings of the ACM Internet Measurement Conference. Association for Computing Machinery; 2020. . DOI
30. Santos, C.; Nouwens, M.; Toth, M.; Bielova, N.; Roca, V. Consent management platforms under the GDPR: processors and/or controllers? In: Gruschka N, Antunes LFC, Rannenberg K, Drogkaris P, editors. Privacy Technologies and Policy. Springer International Publishing; 2021. pp. 47–69. DOI
31. Strzelecki, A.; Rizun, M. Consumers' change in trust and security after a personal data breach in online shopping. *Sustainability* **2022**, *14*, 5866. DOI
32. Raman, P.; Kayacık, H. G.; Somayaji, A. Understanding data leak prevention. In: 6th Annual Symposium on Information Assurance (ASIA'11). Citeseer; 2011. pp. 27–31. Available from: https://people.scs.carleton.ca/~soma/pubs/raman-asia2011.pdf.
33. Schoenebeck, S.; Goray, C.; Vadapalli, A.; Andalibi, N. Sensitive inferences in targeted advertising. *Northwestern J. Technol. Intell. Prop.* **2024**, *21*, 1. Available from: https://scholarlycommons.law.northwestern.edu/njtip/vol21/iss2/1/.
34. Preston, R. Stifling innovation: how global data protection regulation trends inhibit the growth of healthcare research and start-ups. *Emory Int'l L Rev* **2022**, *37*, 135. Available from: https://scholarlycommons.law.emory.edu/eilr/vol37/iss1/4.
35. van Zeeland, I.; Pierson, J. Data protection rsks in transitional times: the case of European Retail Banks. In: CPDP 2022: Data Protection & Privacy in Transitional Times. Bloomsbury Publishing; 2023. pp. 1–26. DOI
36. Gamalielsson, J.; Lundell, B.; Butler, S.; et al. Towards open government through open source software for web analytics: the case of Matomo. *JeDEM eJ. eDemocracy Open Gov.* **2021**, *13*, 133–53. DOI
37. Quintel, D.; Wilson, R. Analytics and privacy: using matomo in EBSCO's discovery service. *Information Technology and Libraries* **2020**, 39. DOI
38. Valdez, C. A glimpse at german privacy laws, from a dark past to the strictest data protection laws in Europe (but There Is Still a Long Way to Go). *Rutgers JL & Religion* **2016**, *18*, 430. Available from: https://heinonline.org/HOL/Page?handle=hein.journals/rjlr18&div=22&g_sent=1&collection=journals.