Bachelor thesis
Computing Science

Radboud University

Cookie Compliance of Dutch Hospital Websites

Author:
Alessandra van Veen
4683382

First supervisor/assessor:
prof. dr. ir. A.P. de Vries
a.devries@cs.ru.nl

Second assessor:
prof. dr. mr. F.J. Zuiderveen Borgesius
frederikzb@cs.ru.nl

January 13, 2021

Abstract

We investigated to what extent Dutch hospital websites are compliant with the Algemene Verordening Gegevensbescherming (AVG), the Dutch implementation of the EU's General Data Protection Regulation (GDPR), regarding cookie consent. No previous research had been done into cookie consent in the medical sector. We limited our scope to look at cookies that were set before any consent was given. We used OpenWPM to collect our dataset and created our cookie database for cookie categorization. By combining these lists, we could then look at the statistics. We found that 42% of all Dutch hospital websites had at least one cookie that did not comply with the AVG. The result suggests that many Dutch hospital websites need to improve their websites to make them AVG compliant. It has also shown that even the medical sector still has a high rate of non-compliance with the AVG, and we hope to encourage more research into this on a larger scale.

Contents

Chapter 1

Introduction

Nowadays, whenever you visit a website for the first time, you will most likely be confronted with a cookie consent banner. A cookie consent banner is a small pop up that allows you to select which cookies you allow to be stored on your computer for that website. This is required by the General Data Protection Regulation (GDPR) for all websites that natural persons from the EU use. Your choices may range from necessary cookies to analytical or even tracking cookies. For some cookies, such as tracking cookies, consent must always be explicit. Even two and a half years after the GDPR went into effect, many companies still do not follow these rules. On the 7th of December, 2020, the CNIL, the French data privacy watchdog, gave Google a financial penalty of a total of 100 million euros [1] and Amazon a financial penalty of 35 million euros [2] for a multitude of breaches. One of the breaches was that they both deposit cookies without obtaining the prior consent of the user. The GDPR states that tracking cookies and most often analytical cookies may not be deposited before a user gives their consent.

This raises the question that if such big companies violate these laws, what about smaller ones, with often very few resources? While a lot of research has been done on cookie violations, the main focus was usually on the most popular websites. Certain sectors, such as the medical sector, have never been covered before, yet this is a sector that often concerns itself with highly sensitive personal data. Also, not a lot of research has been done on regional levels. Because of the lack of cover in the medical sector, we carried out a study to investigate cookie violations for Dutch hospital websites. Due to the lack of Dutch regional research, our solution was only semi-automated. Most of the analysis was done manually to ensure the most accurate results. We will further elaborate on these choices in Chapter 4. Due to our regional focus, we have also decided to specifically focus on the Algemene verordening gegevensbescherming (AVG) which is the Dutch implementation of the GDPR.

Our research question is as follows:

- To what extent are Dutch hospitals websites compliant with the AVG rules, specifically in regard to cookie consent?

To help answer this question, we first want to answer the following sub-questions:

- How can we measure all cookies that are present on all Dutch hospital websites?
- When is a cookie compliant with the AVG?
- Which cookies of the dataset are compliant with the AVG?

By answering these questions, we can show Dutch hospitals how to comply better with the AVG in regards to cookie consent. Also, this research can also promote more research into cookie consent research on regional levels, but also for the medical sector and other sectors that deal with sensitive data.

In the next Chapter, the preliminaries, we describe cookies and how they work. Chapter 3 discusses the related work. In Chapter 4 we describe the process to get our results, the choices we made, and show our results. In Chapter 5 we discuss the results and the implications these have, and at last Chapter 6 will form the conclusion.

Chapter 2

Preliminaries

Cookies, or more specifically HTTP cookies, have evolved over the last two decades to be essential for almost any website to work. In these two decades, their purpose has also grown. Not only do they provide basic functionality for websites, but many companies now also use cookies to track a user across the web. In this Chapter, we explain HTTP cookies in terms of what they are, their structure, and cookies as a tracking mechanism. Zombie and flash cookies (and edible ones) fall out of the scope of this thesis.

2.1 Cookies

Cookies are small pieces of data limited to 4 KB that are set by the webserver on the client [3]. This data is then saved in the browser storage. As HTTP is stateless by design, HTTP cookies were used to increase the statefulness of the web and to be able to easily identify a user. They can be used to store information such as items you put in your shopping cart or even logins. As described by Kristol [4], there are different ways to achieve a stateful web or to achieve identification, but many are far less reliable than cookies. To set a cookie, there are two methods. Either the host adds the Set-Cookie header in the HTTP reply message to user requests, or via JavaScript executing on the client computer. The browser then adds the cookie to its storage with a *same-origin* policy. This states that cookies may only be set, read or modified by domains which match with the *Domain* and *Path* attributes of the cookie [5]. For a service to read a cookie, there are again two options. All cookies are automatically added to all HTTP requests that belong to the same domain, or they can be requested via a JavaScript API.

The *Path* and *Domain* attributes are far from the only attributes of a cookie. In its basic form, a cookie is defined as follows [6]:
Name=Value; Host=example.com; Path=/account; Expires=Sun, 3 Jan 2021 10:11:12 UTC; Secure;
Each attribute is defined as follows:

Name is the given name of the cookie. This helps to uniquely identify a cookie to a particular server.
Value contains the actual data. This data can appear in clear text, but may also be encrypted.
Domain is used to identify the cookie's origin server. This attribute is necessary to allow for communication between the browser and the proper server, so the cookies can be sent back. This attribute is also used to help distinguish first and third-party cookies.
Path limits the scope of a cookie to a set of paths. The browser will only include the cookie in an HTTP request if the path portion of the request-URI matches the path. If the path is not specified, it will be set to root.
Expires determines the maximum lifetime of the cookie in a datetime string format. If this attribute is missing, and the *Max-Age* attribute (maximum lifetime in seconds) is also not present, the cookie will be defined as a session cookie. This means that the cookie will be removed when the session ends.
Secure is a flag that limits the cookie to secure channels. If the secure attribute is set, the browser will only send the cookie in an HTTP request if and only if the request is transmitted over a secure channel, such as HTTP over TLS (HTTPS).
HttpOnly is a flag that limits the cookie to HTTP requests. Specifically, it means that this cookie will not be available via the JavaScript API.

2.2 Cookies as a Tracking Mechanism

The different attributes are utilized to turn cookies into a tracking mechanism, and sometimes the attributes can be used to help identify whether a cookie is a tracking cookie as well.

The *Name* attribute is commonly used as a variable name and it is often assumed to hold no personal information. This assumption was tested by Gonzalez [5] and was found to be untrue. In fact, they found that some cookies even had complicated cookie names that stored all kinds of different values. As a consequence, this also makes it harder to block certain cookies based purely on their name. It is, however, important to note that these complicated names, while often used for tracking, can also be used for operational means as well. YouTube for example uses cookies which store the video ID and a timestamp in its name.

The most common way, however, to store tracking information in a cookie is to use the *Value* attribute. Within this, it is easy to store unique identifiers and other sorts of information. In our dataset, we have found cookies that held the IP address of the user. In many cases, however, these values are encrypted. This means that most of the time, the value cannot be used to determine whether something is a tracking cookie or not.

Some papers consider cookie lifetime to help detect whether something

is a tracking cookie [7]. However, Fouad et al. [8] determined this is not a good heuristic. As they state, domains can easily keep updating the cookies, and then handle the data on the server side.

The *Domain* attribute can be used to determine if something is a tracking cookie, but this is not a foolproof method. However, it can be used to determine whether a cookie is a first or third-party cookie. This is because by definition for third-party cookies, the domain does not match the website the user actually accessed. Third-party cookies always have the potential to be tracking cookies, because whenever a third party cookie is set, updated, or read, this is always done by a different domain than the domain you are on. This means that the other domain is aware of the domain you are on. In most cases, this is used for cross-website tracking. Over the last few years browser manufacturers, especially Apple (Safari) and Mozilla (Firefox), have started to try to directly block third parties from setting or reading cookies. Though, this is something that can be bypassed using a JavaScript that redirects the user temporarily to the third-party website [7]. This would make the third-party cookie a first-party. It is more difficult to determine for first-party cookies whether they are used for tracking or not. While most first-party cookies are not used to track, they can very well be used for tracking purposes. Performance cookies are most often seen as first-party, yet some of these do have cross-web tracking capabilities as well.

Lastly, the *Path* can be used with tracking, however, it is not limited to tracking. Cahn et al. [6] found that 98.4% of their dataset had the path set to the root level. This means that a user can then be identified on every single webpage of that domain.

Chapter 3

Related Work

Advertising networks use trackers to analyse the behaviour of internet users. In 2013, Gomer et al. [9] investigated, as one of the first, the tracker networks that support Online Behavioural Advertising (OBA). First, they collected their data by running multiple Linux machines in parallel. Each machine spawned a browser instance controlled by a Python program. They used custom logging add-ons to save all cookies and the referrer. The referrer headers were used to establish the graph where each directional edge showed which node refers to which node. Their research showed that in 30 clicks on search results, there was a 99.5% chance that a user would be tracked by all top 10 most frequent trackers. This showed how extensively third-party tracking through cookies was already present in 2013 throughout many different sectors.

In 2016, Cahn et al. [6] did a large-scale study to analyse approximately 3.2 million cookies. They investigated the characteristics, placement, and information transmission. With the use of a web crawler, they went over Alexa's top 100k websites. Alexa's top website list has been commonly used by a lot of the research in this section to gather their seedlist for the crawl. Cahn et al. used Cookiepedia to categorize the cookies. They found that the majority of their cookies were performance related and that 68% of the cookies found were from a third-party. For our research, Cookiepedia has also been used as the main resource for cookie categorisation.

In the same year, Carpineto et al. [10] created a method called CooLCheck to automatically assess whether a website meets the requirements of the Cookie Law. They focus on the placement of tracking cookies in combination with user consent. The method uses a combination of cookie disclosure and classification, together with the identification of natural language consent requests. For their dataset, they used Alexa's top 500 websites in Italy. To classify the cookies, they combined two different filter lists. They discovered that of the 17073 sites with analysable content, 1930 websites placed

tracking cookies. The tool for automatic assessment was created for consent banners written in either English or Italian. It is also unknown how the implementation of the GDPR affected the usefulness of this tool.

Another important paper from 2016 was written by Englehardt and Narayanan [11]. The most important contribution of this paper was the privacy measurement tool OpenWPM, which can perform 15 different measurements. One of the measurements, stateful cookie-based measurements, has also been used in our research. The tool is still used in research to this day and at the time of writing, has been used in over 50 research projects [12]. As of April 2019, OpenWPM has moved towards Mozilla and is still being updated.

In 2017, a few more large-scale studies have investigated how cookies were being used. A notable example is from Gonzalez et al. [5] They performed a study on 5.6 billion HTTP requests to gather HTTP cookies. On these cookies, they analysed the cookie structure and found this structure is more complex than thought previously. They also mention how this undermines some assumptions about cookies from previous research projects, such as that the cookie name does not contain relevant information about the user. They show that cookie names can indeed have identifiers within them.

During this year, two other papers got published that analysed cookies on a national scale [13][14]. The first paper by Aladeokin, Zavarsky, and Memon [13] looked at 4 different countries and 5 different categories. From each category, it selected the top 20 websites per country. They then compared the Cookie laws of each website and manually investigated whether cookies were compliant. They found that in developed countries, a user has a higher probability to be tracked (around 70%) than in developing countries (around 55%). Ruohonen and Leppänen's research [14] investigated the amount of persistent third-party cookies present on the Finnish web and compared this with the global web. For this, they used around 206 popular Finnish websites. They did not use Alexa's popular website list, but they used a Finnish equivalent instead. Similar to our paper, the sample is small compared to a global worldwide scale. To generate their results, they retrieved their cookies from Firefox's cookies.sqlite file and cleaned up all Firefox-specific data from the local hard disk after every visit. An interesting result is that 91.4% of their cookies were third-party.

Shuford et al. [15] performed a small scale study in 2018 to investigate the effect of browser settings and browser extensions on the surveillance of users. With a browser extension to detect third-party trackers, they manually investigated 10 different websites. With this, they also discovered that 'vanilla tracking' was the most common form of tracking, namely the use of cookies of a third-party to track a user across websites.

As the GDPR went into effect, researchers also began to study the effects of the new law. In 2019, Sakamoto and Matsunaga [16] investigated opt-out states in regards to OBA post-GDPR. Their method consisted of the OpenWPM framework, which they used to perform multiple crawls. Like many other papers, they used Alexa's top 100 news websites list as their base. The first crawl phase involved crawling the websites with third-party cookies enabled. In the second phase, they used a browser with all cookie data from phase one. They used Digital Advertising Alliance (DAA) and Network Advertising Initiative (NAI) opt-out websites to automatically opt-out of all third-party cookies. Both the DAA and the NAI are groups with self-regulatory programs focused on internet-based advertising. In phase three, they used the same browser again and re-crawled all the websites. These phases were performed in both July 2017 and June 2018. Using McNemar's test, they found that the GDPR had no effect at the time. They determined that only about half of the ad agencies stop web tracking when a user enables OBA opt-out.

In the same year, Bailey, Laakso, and Nyman [17] conducted a study on the 500 most frequently visited websites by Finnish users. This makes it one of the few studies that focus on trackers with a regional focus. Their goal was to investigate the difference in tracking between Finnish and non-Finnish sites. In contrast to the other Finnish paper [14], which had a similar goal, Bailey et al. do use Alexa's top websites list. To obtain their cookie data, they used the tool Tracker Tracker that is based on the Ghostery browser extension. Lastly, to identify tracking ownership, they used Disconnect. The results were that there was a similar amount of tracking on Finnish and non-Finnish webpages, but there was a difference in the frequency of tracker owners. For example, Google was much more present on Finnish ones. An important note to make is that even though the study was published after the GDPR went into effect, the data was collected before that time period.

Research into tracking is still active to this day. In February 2020, two studies on this topic were published that also use data of 2018 or later. Matte, Bielova, and Santos' [18] study investigated how to detect suspected violations of the GDPR and the ePrivacy Directive in cookie banners. As part of their methodology, they created a crawler called Cookieinspect to perform large-scale automatic and semi-automatic crawls. The crawler also has the ability to detect the presence of a TCF (Transparency & Consent Framework) banner. To choose their websites, they used Tranco. To find any suspected violations, they first performed an automatic crawl to gather all websites with a TCF banner. This is then followed by a semi-automatic crawl which requires interaction with the cookie banner. The last phase a verification procedure to cross-check the results. This resulted in 54%

websites of the checked websites contains at least one suspected violation.

Fouad et al. [8] also published a study in 2020 in which they propose an alternative method to detect trackers. Using Alexa's top 10,000 domains, they performed two stateful crawls with OpenWPM. Each crawl was performed on two different machines with two different IP addresses. They then developed a method to detect identifier cookies, which are cookies that store identifiers. They use different qualities to determine whether a cookie is an identifier cookie or not. According to their study, 91.92% of the websites incorporate at least one type of cookie-based tracking, even after the GDPR came in force.

Maris, Libert, and Henrichsen [19] published a paper in October 2020 that investigated the tracking and privacy risks on pornography websites. One of the things they studied was to what extent do these websites reveal user data and allow for third-party identification and tracking. To investigate this, they used the webXray software platform, which allowed them to capture network traffic and record cookie data. Part of the identification involved cookies and they found that 79% of the pages had at least one third-party cookie present. This is lower than our previously discussed paper by Fouad et al. [8]

All of these papers are about or heavily feature (tracking) cookies, similar to our paper. While many of the papers investigate the presence of tracking cookies, our focus is unique. Not many papers focus on regional tracking [10][13][14][17] nor do they focus on a specific sector or category [9][13]. Our paper is the first paper to have a regional focus on the medical sector when it comes to the presence of tracking cookies.

Chapter 4

Research

In this Chapter, we describe the methods we used and present our results. The Chapter has been divided into three sections, each corresponding to one of the sub-questions mentioned in the introduction. In Section 4.1, we describe how we collected our list of websites and our cookies. In Section 4.2, we first discuss when a cookie is compliant with the AVG. After, we follow with how we created our filter list, a list of all unique cookies categorized. In Section 4.3, we have applied our filter list to our cookie list, and describe the results.

4.1 Data Collection

4.1.1 Retrieving Websites

Before we can retrieve our cookies, we must first establish a list of all Dutch hospital websites. We decided to base our result list on a website called Zorgkaart Nederland [20]. This is a well-known Dutch Web directory of healthcare that has an overview of all healthcare providers in the Netherlands and allows people to rate different providers. It also provides the ability to search based on specific organisations.

We wrote a script using Python 3.8.3 and a Python library called BeautifulSoup4 4.9.3, which allowed us to parse HTML and XML web pages. The script first extracted all links to the Zorgkaart Nederland hospital pages that are divided over 17 pages. Then, we made a request to each page and used BeautifulSoup to extract the link to the hospital website and gather the name of the hospital. This gave us a list of the websites of 301 hospitals. We found that this automated approach missed the links for two hospitals. These were added manually. Due to how Zorgkaart Nederland works, listing each hospital location seperately, a singular hospital could appear multiple times in the list. To solve this issue, we decided to manually filter the list.

During filtering, we also came across other issues that made manually filtering the better choice. A hospital with multiple locations often appeared

with different links and names. Sometimes the link led to the same web page, and there are cases it did not. Two hospitals with non-working links and two hospitals that had closed down were excluded from this analysis. We were left with 81 hospitals from the 301 results.

4.1.2 Retrieving Cookies

With our list of websites, we could now automatically gather all cookies found on these websites before user interaction. We decided to perform our measurements with the web privacy measurement framework OpenWPM [11] version 0.12.0.[1]

OpenWPM is a tool designed to perform automated web scraping and is able to perform large-scale measurements. The browser automation and data collection infrastructure of the tool are divided into three main modules: browser managers, a task manager, and a data aggregator. A browser manager ensures stability. OpenWPM is using the Selenium WebDriver as its browser driver. Selenium is often known to hang indefinitely. [11] A browser manager provides an abstraction layer surrounding the driver and allows for recovery when a browser fails or hangs. The task manager is the controlling instance. It is responsible for all browser instances and it can spawn and control multiple browser instances simultaneously. Lastly, the data aggregator logs all data in a standardized format.

OpenWPM is a tool commonly used in research [16][8][12]. The tool is suitable for research into cookies, as unlike many other similar tools, OpenWPM allows us to gather not just cookies gained through HTTP requests, but also those through JavaScript. As stated by Englehardt and Narayanan [11], executing JavaScript and other dynamically loaded content is important for a privacy study, as that often includes advertising resources otherwise missed. OpenWPM is, however, not without its limitations. It is unable to perform actions such as selecting and accepting options on cookie banners. While there has been research into tools that can analyse, and possibly use, cookie banners [10][18], it is often limited or unable to work with the Dutch language. All cookies we will collect with OpenWPM are cookies that are set before any consent has been given.

For our research, we have created a setup to gather the required cookies. The source code of OpenWPM includes a python demo. We modified this code so it referred to our list of websites. We also set most of the *browser_params* to false. Specifically, we disabled all measurements except the *cookie_instrument*. This browser parameter allows us to record cookie changes. With these settings, we could perform a simple stateless crawl. The crawl must be stateless to ensure each new page visit uses a fresh browser

[1] We retrieved our cookie list on November 19 2020, the day version 0.13.0 was released. As the changes would not affect our results, we decided to stick with 0.12.0.

profile. The process of recording all cookies of all 81 websites took less than an hour.

The result was a total of 709 cookies from 78 different websites [21]. Only three websites did not give any cookies.

4.2 Cookie Categorization

4.2.1 Cookie Compliance

In the Netherlands, two laws determine the validity of the use of cookies. The first being the Telecommunicatiewet and the second is the Algemene verordening gegevensbescherming (AVG). The Telecommunicatiewet is the Dutch implementation of the EU ePrivacy Regulation. The rule about cookies defined by Article 11.7a [22]. This states cookies may only be placed if and only if the user has given their informed consent. There are two exceptions to this rule. A website does not have to receive permission from the user if the cookie's only goal is communication. Also, a website does not need to gain permission first if the cookie is necessary to provide the service with the user requests. With the second exception, there must be no or very little consequences for the privacy of the user.

Since the implementation of the GDPR, the Telecommunicatie wet was adjusted to follow the AVG's definition of consent [23]. This is a very strict definition. First, consent must be given through a clear active action, for example through a written declaration. It must also be clear that the user gives their consent out of free will, the consent is specific and informed, and it must be clear to what the user consents to. In terms of cookies, this would be, for example, clicking on a box. With this, it is important that, where possible, the box is not filled out yet. Inactivity nor silence do not count as consent either. If the user uses the website without checking any boxes, the website is not allowed to set any cookies. If the processing of personal data has multiple purposes, this must be clear to the user and the user must consent for each individual purpose. The AVG has more rules on consent, however, for this paper, only the previously mentioned will be relevant.

To easily identify whether a cookie needs consent or not, the Autoriteit Consument & Markt (ACM), a Dutch business regulation agency, divides cookies into three categories: functional, performance and other [24]. The first category, functional, are cookies that are necessary for a website to provide the services the user requested. These are cookies that are, for example, necessary to log in, or to save language choices. The second category, performance, is a special case. In 2015, they added a short sentence to the Telecommunicatiewet that allowed performance cookies, but this exception does come with a limitation. Performance cookies do not need prior consent if and only if the cookies are privacy-friendly. The third and last category, other, can also be called tracking cookies. These cookies are all cookies that

require consent before setting and that are not privacy-friendly for the user.

In the case of our paper, we are interested in tracking cookies and to some degree, performance cookies. As the cookies we have collected were all obtained without providing consent, all tracking cookies from our dataset do not comply with the AVG nor the Telecommunicatiewet. Performance cookies are not clearly in breach. In most cases it is nearly impossible to determine whether or not a performance cookie can be considered privacy-friendly. Google Analytics, for example, has the ability to be privacy-friendly, but it also has the ability not to. The name may sometimes give away if it is not privacy-friendly. Sometimes, there are unique identifiers in the name. Unless otherwise known, these often have tracking capabilities. In the end, without knowing how any performance cookie was set up, we cannot determine whether they are the privacy friendly variant or not. Due to these limitations, we go by the assumption that performance cookies are privacy-friendly. The exception of this rule is if the cookie was set by a third party. If it is set by a third party, we consider the cookie to not be privacy-friendly. First-party performance cookies allow only for minor tracking, namely that within the domain itself [8]. So we can assume these are privacy-friendly. A third-party performance cookie is not restricted to the domain itself, and thus there is a possibility of cross-website tracking that also involves usage data. Because of this, we say these are not privacy-friendly by default.

Lastly, if a cookie's consent category cannot be determined, it will be marked unknown. These cookies will be deemed safe as we cannot be sure what its purpose is.

4.2.2 Cookie Database Creation

To categorise our cookie dataset, we decided to create a manual cookie database [21]. The database contains the name of the cookie, category, host, service, retention period, and description. The host is equivalent to the domain attribute. This is the domain that set the cookie. The service determines who provides the service for these cookies. For example, *ARRAffinity* is a cookie originally provided by Microsoft Azure. is the original service provider for google ana. The retention period describes the period of time for which the cookie would be stored in the browser. If it is a session cookie, it will say session instead of the time. While some of the columns are irrelevant for the current research, they did help to determine the cookie category. The inclusion of this information can be useful for future research. We chose a manual approach because it does give the most accurate results. It allowed us to cross-reference multiple databases to determine the most accurate category. This was also a viable approach as we only processed 119 unique cookies.

Most of the information could be extracted from our dataset. To determine the category, the service, and the description, we used Cookiepedia [25]

as our main source. We found that this was an extensive database of both common and uncommon cookies. This resource was also used by Cahn et al. [6]. In case a cookie was not present on Cookiepedia, or information was missing, we used two additional sources. The first is cookiedatabase.org [26] which was founded by a group of volunteers, and supported by Complianz, a Privacy Suite for WordPress and SIDN Fund. Our second source was an open-source Github database [27], which we partially cross-referenced with the earlier two sources to confirm its validity.

Both Cookiepedia and Cookiedatabase.org define their consent categories differently from the ACM. Cookiepedia uses the classification system developed by The UK International Chamber of Commerce (ICC). It consists of four categories, strictly necessary, functionality, performance, and targeting/advertising. Strictly necessary cookies are those necessary for the website to work. Functionality cookies are those which remember the choices you made and provide an enhanced user experience. Both of these cases are privacy-friendly cookies and thus would be equivalent to the Dutch category of functional cookies. Performance cookies follow the same classification as with the ACM classification. Any cookie with a performance-related task but which is not privacy-friendly would be classified as a targeting/advertising cookie. Targeting/advertising cookies are equivalent to the tracking classification. Cookiedatabase.org follows a similar classification system. They categorise cookies as functional, preferences, statistics, statistics-anonymous, and marketing/tracking. Functional and preferences are once again equivalent to ACM's classification of functional cookies. Both statistics cookies are equivalent to performance cookies, however, Cookiedatabase makes a clear distinction whether it is anonymous or not. Marketing/tracking is equivalent to tracking cookies. As Cookiepedia is the main source for our database, we decided to use their categories.

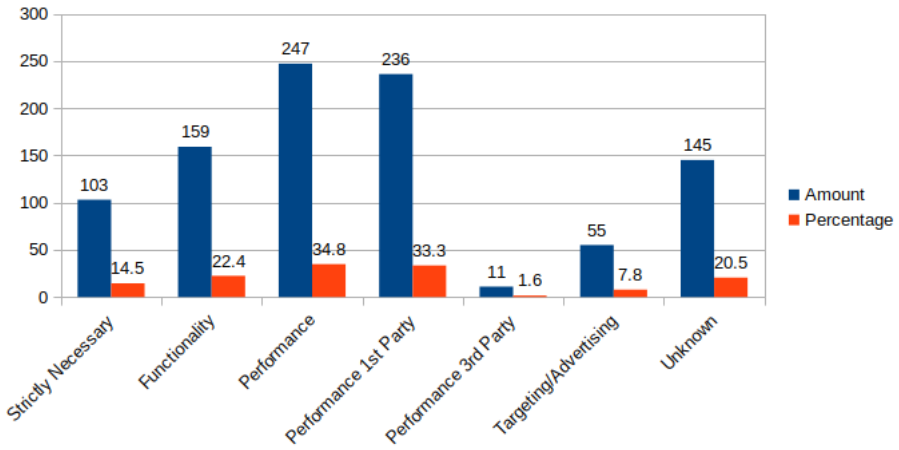
Some cookies are not recorded or had missing information on all three sources. In these cases, we would first see if we can find the service of the cookie. If we could, we would look at the website of that service to see if they could offer clarification. In the cases we could not, we would see if the service could be that of a generic cookie. A generic cookie is a cookie that is not from a particular service provider, but rather of the website itself. They are also unique to a domain. Strictly necessary cookies and functional cookies are generic cookies if the website did not use a particular service. If we determined a cookie was a generic cookie, we would look at the name, the retention period, the value and the host, to determine the consent category. We also looked at the privacy policy of the website for more information. Determining the category was not possible in all cases, but in some cases, we could determine with certainty what the purpose, and thus the consent category, of the cookie was. A common example was cookies that saved the language preference of the user. The name and the value of these cookies often indicate this.

4.3 Results

To combine the cookie database and the dataset, we wrote a Python script that joins the two based on the name. We limited the columns to the website, name, category, service, and host [21] as that is all we needed to gather our results. With the join, we did have to consider two special cases. The first case are cookies with a pattern in their name. Sometimes cookies have a identifier in their name which always follows a specific pattern. In these cases, the pattern is replaced by a *. To illustrate, a cookie from our dataset called *_gat_UA-8527914-1* would be joined with *_gat_UA-** from our cookie database. The second case is about a very specific cookie, which also contains a pattern, called *NSC**. Cookies with this name can exist as a session cookie, or as one that expires in a year.

We first looked at how many first-party cookies were present versus how many third-party cookies were present. This was 547 first-party against 162 third-party. This means that around 77.2% of our cookies were first-party. It is, however, important to note that a first-party cookie called *bovenij_guid* was present 103 times. This cookie was also limited to a singular website and was most likely from the website itself. While there were more cookies that had duplicates on a singular website, those only had at most five duplicates. If we would only have the textitbovenij_guid cookie a singular time, as is more reasonable, we would have 445 first-party cookies against 162 third-party. This means that around 62.8% are first-party. We also looked at how many webpages had at least a first-party cookie present and how many webpages had at least one third-party cookie present. 78 webpages, so all pages that had at least one cookie, had a first-party cookie present. 21 pages had at least one third-party cookie present. That's equivalent to 96.3% and 25.9% of the websites respectively.

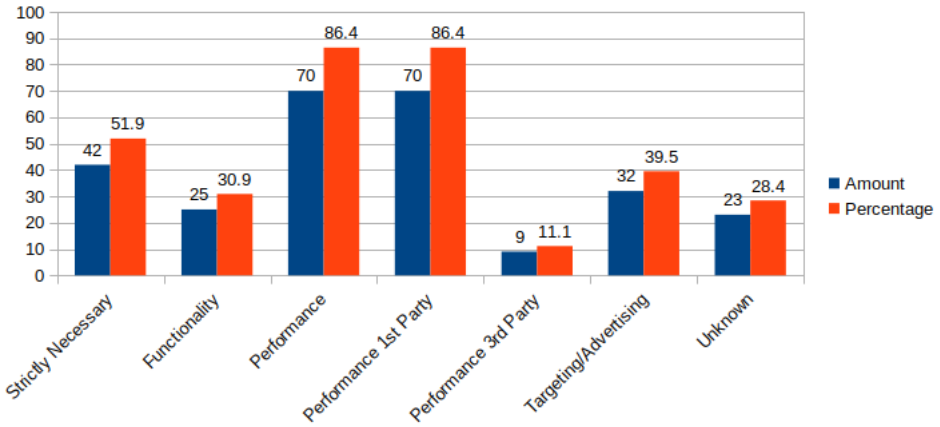
Figure 4.1: Total amount of cookies per consent category



Subsequently, we looked at how many cookies from each category were present. This can be seen in Figure 4.1. As we consider first-party performance cookies to be safe, and third-party to be a violation of the Telecommunicatiewet and the AVG, we have also shown how many were present for each.

In total, 90.7% of the cookies were safe, and 9.4% of the cookies were not privacy-friendly, meaning they have the ability to track the user across the web. These numbers sum up to 100.1% due to rounding to the first decimal place. With the first-party and third-party performance cookies, both rounded up, causing them to not be equivalent to the percentage of all performance cookies.

Figure 4.2: Amount of websites with cookies by cookie consent category



When we compare these results to how many websites had at least one cookie present per consent category, we get quite different percentages. This can be seen in Figure 4.2. The most common cookie type, both in amount and how many websites have at least one, are performance cookies. 86.4% of the websites have a performance cookie present. Each website of our dataset with a performance cookie is also guaranteed to set a first-party performance cookie. Not shown in Figure 4.2 is how many websites have at least one privacy-friendly cookie. All Dutch hospital websites that set a cookie, namely 78, have at least one such cookie.

While there are only 55 targeting/advertising cookies, which made up for 7.8% of all the cookies, they are present on 39.5% of the websites. Third-party performance cookies make up 1.6% of the database, yet these are present on 11.1% of the websites. If we take the union of the set of websites with third-party performance cookies and the set of websites with targeting/advertising cookies, we would have 34 websites. This means that 42.0% of Dutch hospital websites place one or more cookies that are not compliant with the AVG consent regulations, because these cookies are placed before

these users could give their consent.

Chapter 5

Discussion

Our results show that 42.0% of the websites have at least one cookie that is not compliant with the AVG regarding cookie consent. This number is high, considering we do not even consider the full scope of what is consent and there is a reasonable expectation that hospital websites have a higher consideration for privacy than most corporate websites. This expectation comes forth from the fact that hospitals as an entity handle highly sensitive data and those who visit hospital websites for information tend to be in a more vulnerable position. It raises the question of why any form of targeting/advertising cookie is necessary on a hospital website. Even the use of performance cookies can be taken into question. While performance cookies can certainly be useful to help improve and maintain the website and the traffic to it, this should always be done in a privacy-friendly manner. This is even more important considering that these are hospitals. As we stated before, even with first-party cookies we cannot be sure if they are set up in a privacy-friendly way. Our results imply that even some of the institutes we trust with some of our most sensitive data are not always up to date when it comes to privacy laws on their websites.

In our data, a few cookies stood out and were quite unexpected. Two different hospital websites had cookies from Doubleclick, an advertising company by Google. While this is lower than what Bailey, Laakso, and Nyman [17] recorded, namely 63% of the websites had Doubleclick presence, it is still concerning. There is a reasonable expectation that ads or ad-related cookies are not present on hospital websites. Doubleclick functions by allowing third parties to execute code on the website. This can not only lead to potential data leaks, but it has also in the past led to the infection with malware [28]. On the same note of advertising cookies, there was one cookie called *NID*. This cookie is from Google Ads Optimization and is used to enable ad delivery or retargeting. It can also be used to store user preferences. There were also some cookies present from Youtube and Vimeo that can be classified as either third party performance cookies or targeting/advertising

cookies, such as Youtube's *GPS* cookie, which shares your physical location with Youtube. Another website had the presence of some Facebook cookies such as *fr* which contains an encrypted Facebook ID and Browser ID and it is used to track logged out Facebook users. The last of our tracking cookies were from Google Analytics, however, these contained unique identifiers within the cookie name and were suspected by Cookiepedia to be targeting/advertising cookies.

While there is no research related to other hospitals we can compare our results with, we can look to compare our results with others on a larger scale. In a study by Fouad et al. [8] they found that 91.92% of the websites incorporated at least one type of cookie-based tracking. Another study [19] found that 79% had at least one-third party cookie present. They counted third-party cookies as tracking cookies, or at least as cookies with a high potential to track. Both of these studies had different rules for what counted as tracking, though both did include performance cookies as tracking. These results are much higher than our result of 42.0%, however, it is important to take into account the different laws. If we counted first-party performance cookies as tracking, our result would be much higher as 86.4% of our websites already include first-party performance cookies. Interestingly, if we compare our results with those of Aladeokin, et al. [13], who performed a study on a regional scale, the numbers become a bit closer. In their study, 70% of their UK websites were compliant with the UK privacy protection law. This is higher than our 58% compliance. While they also investigated other countries, we felt the UK was closest to the Dutch laws in terms of privacy protection. A good note is that these UK websites reached a relatively high compliance rate despite having stricter rules as to what is compliant compared to our study.

As has already been mentioned, our study had its limitations. We limited our scope of consent to only the cookies that were placed before any further user interaction due to the scope of this paper. This means that our amount of consent violations is a minimum. We were unable to measure whether all consent was properly informed and if choices were made individually using a cookie consent banner. We were also unable to measure what would happen if we were able to first consent to cookies and to then take that consent away. If we were able to measure this, we would have been able to provide a more complete picture of the state of cookie consent for Dutch hospitals. While the percentage is already high, this percentage may be even higher upon further investigation.

Chapter 6

Conclusions

We investigated to what extent Dutch hospital websites are compliant with the AVG rules concerning cookie consent. We found that 42% of all Dutch hospital websites place third-party performance cookies or tracking cookies before any form of consent is given. So while the majority of the hospitals has already taken a step in the right direction, there is still a lot of work to be done. However, consent law is more extensive than we have investigated. We did not account for the existence of cookie banners and we did not consider the ability to first opt-in and then opt-out. With the opt-in, we could have compared the results with both the opt-out, but also with the cookies before any consent has been asked. Nonetheless, we still recommend that these websites investigate their policies and make sure that their website is compliant with both the AVG and the Telecommunicatiewet. As sensitive data can be involved on these websites, it is even more so important that these websites are compliant. This study implies that research into cookie consent of the medical sector is important, because the compliance rate is lower than what we can reasonably expect on a regional scale.

Future research could investigate the compliance rate when the full scope of consent is taken. Another path future research could take is repeating the study in other countries, or even on a more global scale. The study can also be repeated for other medical institutes such as dentists and therapists.

Bibliography

[1] CNIL, "Cookies: financial penalties of 60 million euros against the company google llc and of 40 million euros against the company google ireland limited." `https://www.cnil.fr/en/cookies-financial-penalties-60-million-euros-against-company-google-llc-and-40-million-euros-google-ireland`, accessed 2020-12-22.

[2] CNIL, "Cookies: financial penalty of 35 million euros imposed on the company amazon europe core." `https://www.cnil.fr/en/cookies-financial-penalty-35-million-euros-imposed-company-amazon-europe-core`, accessed 2020-12-22.

[3] Internet Engineering Task Force, "Http state management mechanism." `https://tools.ietf.org/html/rfc6265`, accessed 2020-12-31.

[4] D. M. Kristol, "Http cookies: Standards, privacy, and politics," vol. 1, p. 151–198, Nov. 2001.

[5] R. Gonzalez, L. Jiang, M. Ahmed, M. Marciel, R. Cuevas, H. Metwalley, and S. Niccolini, "The cookie recipe: Untangling the use of cookies in the wild," in *2017 Network Traffic Measurement and Analysis Conference (TMA)*, pp. 1–9, 2017.

[6] A. Cahn, S. Alfeld, P. Barford, and S. Muthukrishnan, "An empirical study of web cookies," WWW '16, (Republic and Canton of Geneva, CHE), p. 891–901, International World Wide Web Conferences Steering Committee, 2016.

[7] T. Bujlow, V. Carela-Espanol, J. Solé-Pareta, and P. Barlet-Ros, "A survey on web tracking: Mechanisms, implications, and defenses," *Proceedings of the IEEE*, vol. 105, pp. 1–35, 03 2017.

[8] I. fouad, N. Bielova, A. Legout, and N. Sarafijanovic-Djukic, "Missed by filter lists: Detecting unknown third-party trackers with invisible pixels," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 2, pp. 499 – 518, 01 Apr. 2020.

[9] R. Gomer, E. M. Rodrigues, N. Milic-Frayling, and M. C. Schraefel, "Network analysis of third party tracking: User exposure to tracking cookies through search," in *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 01*, WI-IAT '13, (USA), p. 549–556, IEEE Computer Society, 2013.

[10] C. Carpineto, D. Lo Re, and G. Romano, "Automatic assessment of website compliance to the european cookie law with coolcheck," WPES '16, (New York, NY, USA), p. 135–138, Association for Computing Machinery, 2016.

[11] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, (New York, NY, USA), p. 1388–1401, Association for Computing Machinery, 2016.

[12] Princeton University, "Software: Openwpm." `https://webtap.princeton.edu/software/`, accessed 2020-12-14.

[13] A. Aladeokin, P. Zavarsky, and N. Memon, "Analysis and compliance evaluation of cookies-setting websites with privacy protection laws," pp. 121–126, 09 2017.

[14] J. Ruohonen and V. Leppanen, "Whose hands are in the finnish cookie jar?," pp. 127–130, 09 2017.

[15] E. Shuford, T. Kavanaugh, B. Ralph, E. Ceesay, and P. Watters, "Measuring personal privacy breaches using third-party trackers," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, (Los Alamitos, CA, USA), pp. 1615–1618, IEEE Computer Society, aug 2018.

[16] T. Sakamoto and M. Matsunaga, "After gdpr, still tracking or not? understanding opt-out states for online behavioral advertising," in *2019 IEEE Security and Privacy Workshops (SPW)*, pp. 92–99, 2019.

[17] J. Bailey, M. Laakso, and L. Nyman, "Look who's tracking – an analysis of the 500 websites most-visited by finnish web users," vol. 38, 12 2019.

[18] C. Matte, N. Bielova, and C. Santos, "Do cookie banners respect my choice? measuring legal compliance of banners from iab europe's transparency and consent framework," 2020.

[19] E. Maris, T. Libert, and J. R. Henrichsen, "Tracking sex: The implications of widespread sexual data leakage and tracking on porn websites," *New Media & Society*, vol. 22, no. 11, pp. 2018–2038, 2020.

[20] Zorgkaart Nederland, "Ziekenhuizen in nederland." https://www.zorgkaartnederland.nl/ziekenhuis, accessed 2020-11-16.

[21] A. van Veen, "Cookie database." https://gitlab.science.ru.nl/avveen/cookie-database, accessed 2021-1-10.

[22] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, "Telecommunicatiewet artikel 11.7a." https://wetten.overheid.nl/jci1.3:c:BWBR0009950&hoofdstuk=11&paragraaf=11.1&artikel=11.7a&z=2020-12-21&g=2020-12-21, accessed 2021-1-3.

[23] EUR-Lex, "VERORDENING (EU) 2016/679 VAN HET EUROPEES PARLEMENT EN DE RAAD." https://eur-lex.europa.eu/legal-content/NL/TXT/?uri=CELEX\%3A32016R0679, accessed 2021-1-3.

[24] Autoriteit Consument & Markt, "Cookies - ACM.nl." https://www.acm.nl/nl/onderwerpen/telecommunicatie/internet/cookies, accessed 2021-1-3.

[25] OneTrust, "Cookiepedia." https://cookiepedia.co.uk/, accessed 2020-12-17.

[26] Cookiedatabase, "Cookiedatabase.org." https://cookiedatabase.org/, accessed 2020-12-17.

[27] jkwakman, "Open cookie database." ttps://github.com/jkwakman/Open-Cookie-Database/blob/master/open-cookie-database.csv, accessed 2020-12-17.

[28] Jérôme Segura, "Malvertising campaign leads to doubleclick ad fraud." https://blog.malwarebytes.com/cybercrime/2016/06/malvertising-campaign-leads-to-doubleclick-ad-fraud/, accessed 2021-1-10.