Privacy Considerations when Seeking Health Information Online

Linh V. Nguyen

Master of Science

School of Computer Science
McGill University
Montreal, Quebec, Canada

April 2025

A thesis submitted to McGill University in partial
fulfillment of the requirements of the degree of
Master of Science

Abstract

More people are using the Internet to seek health information due to the ease of access, the diversity of choices, and the privacy these online sources of information provide. However, surveillance on health websites supported by invasive tracking technologies has increased over time, leading to the collection and aggregation of a large amount of personal user data.

We sampled and analyzed 22 different health information websites relevant to Canadian residents to detect the levels of surveillance conducted on these sites as well as how transparent they are about their surveillance practices. Based on prior work and the results from our analysis, we developed a model to classify different levels of surveillance activities on health websites. Our results showed that most non-commercial health websites from the sample contain less invasive levels of tracking compared to commercial ones. However, only three of these 22 websites are fully transparent about their tracking activities, while the remaining 19 websites do not contain full details about who conducts surveillance on these sites and how.

We then conducted a user study with 20 Internet users to examine how they seek health information online and their awareness of current privacy issues and surveillance practices. We discovered that most participants prioritized gaining access to health content over privacy concerns, except for when it comes to information about sensitive conditions. Moreover, most participants did not plan on making major changes to how they browse online even after being provided with correct information about surveillance, citing their assumptions of the harmlessness of tracking parties, their belief in the privacy of their current seeking strategies, or their reluctant acceptance of tracking as a part of their web searching

experience. We found that these attitudes may be related either to their great demand for health information caused by healthcare inaccessibility, or to their inaccurate understanding of tracking mechanisms—a result of their limited knowledge of privacy technologies. These findings inform how surveillance is conducted on health websites relevant to Canadian residents, as well as how some Canadian health information seekers look for medical content online and their attitudes towards privacy.

Abrégé

De plus en plus de personnes utilisent l'internet pour rechercher des informations sur la santé en raison de la facilité d'accès, de la diversité des choix et du caractère privé de ces sources d'information en ligne. Toutefois, la surveillance des sites web consacrés à la santé, qui s'appuie sur des technologies de suivi invasives, s'est accrue au fil du temps, conduisant à la collecte et à l'agrégation d'une grande quantité de données personnelles sur les utilisateurs.

Nous avons échantillonné et analysé 22 sites web d'information sur la santé pertinents pour les résidents canadiens afin de détecter les niveaux de surveillance effectués sur ces sites ainsi que le degré de transparence de leurs pratiques de surveillance. Sur la base de travaux antérieurs et des résultats de notre analyse, nous avons élaboré un modèle permettant de classer les différents niveaux d'activités de surveillance sur les sites web consacrés à la santé. Nos résultats montrent que la plupart des sites web non commerciaux de l'échantillon contiennent des niveaux de suivi moins invasifs que les sites commerciaux. Cependant, seuls trois de ces 22 sites web sont totalement transparents quant à leurs activités de suivi, tandis que les 19 sites web restants ne contiennent pas de détails complets sur les personnes qui effectuent la surveillance sur ces sites et sur la manière dont elles le font.

Nous avons ensuite mené une étude auprès de 20 utilisateurs d'Internet afin d'examiner la manière dont ils recherchent des informations sur la santé en ligne et leur connaissance des problèmes actuels de protection de la vie privée et des pratiques de surveillance. Nous avons découvert que la plupart des participants donnaient la priorité à l'accès aux informations sur la santé plutôt qu'à la protection de la vie privée, sauf lorsqu'il

s'agit d'informations sur des pathologies sensibles. En outre, la plupart des participants ne prévoyaient pas de changer radicalement leur façon de naviguer en ligne, même après avoir reçu des informations correctes sur la surveillance, en invoquant leur présomption du caractère inoffensif du suivi des parties, leur conviction de la confidentialité de leurs stratégies de recherche actuelles ou leur réticence à accepter le suivi comme partie intégrante de leur expérience de recherche sur le web. Nous avons constaté que ces attitudes peuvent être liées soit à leur forte demande d'informations sur la santé causée par l'inaccessibilité des soins de santé, soit à leur compréhension imprécise des mécanismes de traçage - résultat de leur connaissance limitée des technologies de protection de la vie privée. Ces résultats éclairent la manière dont la surveillance est exercée sur les sites web consacrés à la santé des résidents canadiens, ainsi que la manière dont certains internautes canadiens recherchent des informations médicales en ligne et leur attitude à l'égard de la protection de la vie privée.

Acknowledgements

Moving to Canada to pursue a Master degree was something that 17-year-old me would never have imagined, especially as someone who had just moved to the Netherlands all the way from Vietnam on his own. Yet here I am, writing the last few words of my thesis and getting my Master degree. This was only possible with the love and support from the many people I cherish in my life, to whom I would like to express my greatest appreciation here.

First and foremost, I would like to extend my sincere gratitude to my supervisor, Professor Martin P. Robillard. I could not have finished this thesis and this degree without his constant support and guidance. His genuine dedication to his students as well as his expertise have been a tremendous help to me. It was a great honor to have worked with a supervisor as considerate and passionate as Prof. Robillard, a rare opportunity that would not have happened had he not been willing to take a chance on a student who was completely unaware of the fact that Master applicants should communicate with potential supervisors beforehand. For all of this, I am extremely and forever grateful.

I would also like to thank Professor Jin Guo, one of my two collaborators. Without her unique perspectives on numerous topics and invaluable guidance, the work that this thesis is based on would not have come to fruition. Moreover, she has been an exceptional co-leader of our research group, with whom I have also had the rare but thoroughly pleasant experience of working under as a teacher's assistant. To my other collaborator who is also a dear friend of mine, Deeksha, I am extremely grateful for her tremendous contribution to this project. Her wisdom and experience have helped me grow significantly as a researcher. Furthermore, her friendship and mentorship has not only helped me keep my

plants alive, but also supported and comforted me through different obstacles during my Master's, including the greatest of obstacles—administrative tasks from the university.

I am also thankful for my labmates at the Software Technology Lab, who have been the most supportive research group and the best working environment anyone could ask for. A special shoutout to these labmates: Mathieu, for giving me invaluable advice and teaching me Québécois culture as well as the joys of the French language; Avinash, for always helping me with an insufferable number of questions I have had over the years about literally anything, and for listening to my silly rants; Ziming, for the many post-meeting coffee sessions at Tim Hortons where we had interesting conversations that helped us procrastinate very successfully; and Sara, for being the best deskmate ever, with whom I shared many long work days at the lab. For Jaz, my labmate, friend, and roommate, I am grateful to her for giving me the best "home away from home", and for her continuous support through both the toughest and the happiest of times, including the random last-minute food excursions across the border.

I would also like to thank the friends I have made in Montréal, without whose friendship I would have never had such a memorable time living in this city for the past three years: To Quentin, the best Frenchman and the perfect lunch buddy; to Amin, the happiest and most optimistic person I got the pleasure of knowing; to Milou: my fellow bagel enthusiast who always helped me miss the Netherlands a little less; to Jamie, my Blues companion who has always been there for me; to Sahar, my first true friend in the city whom I will forever cherish; to Sophia, a friend who shared my pop culture references, big city walking tendencies, and impeccable driving skills; to Mel, my favorite "frosh kid" and the best person to have midweek pub crawls with; and to Ariana, who never declined my invitation to buffets and always tolerated my terrible attempts at a British accent.

It would not be complete without also acknowledging the close friends that I have known all my life and, despite great geographic distances, were still people I can depend on no matter what. Danke schön and merci Thea, for always supporting me from an ocean away with your humour and love. Спасибо Vic, for being my ride-or-die during the hard times of COVID and for always picking up my call and comforting me, day or night. Dankjewel Lucas, for being one of the funniest, most talented friend of mine and a pure ray of sunshine. Dankjewel Bartjan, for being one of my favorite Dutchies and a brilliant

friend/travel buddy. Mulţumesc Ana, for being one of the few people who share my music taste and for the frequent Discord calls that have been a great source of joy for me. Cảm ơn Uyên, for being my best friend of 14 years and being there for me through the many, many ups and downs, all the way from California.

Finally, I am forever indebted to my family. To bà ngoại, her words of encouragement have never failed to motivate me. To my other grandparents whom I miss dearly—ông bà nội và ông ngoại—even though they could not be here to experience this, I hope that they are proud of me, especially for keeping my promise of getting a postgraduate degree from when I was barely five. To my sister, chị Ánh, I want to thank her for being the greatest older sister: cảm ơn chị vì đã luôn quan tâm em những lúc khó khăn và là một người chị gái em luôn có thể tin tưởng, đặc biệt là những lần em tạo bất ngờ cho bố mẹ cũng như sự sẵn sàng gọi trà sữa cùng em bất kỳ lúc nào. And last but certainly not least, to my parents, I could not have accomplished anything, including this degree, had it not been for their immense sacrifices, constant love, and unwavering support throughout the years: Con cảm ơn bố mẹ vì đã luôn yêu thương, đồng hành và tin tưởng con qua mọi nẻo đường, qua 3 đất nước khác nhau, và sự hy sinh vô tận của bố mẹ để con và chị có được ngày hôm nay. Con sẽ không bao giờ quên được những gì bố mẹ đã làm cho chúng con, con yêu bố mẹ nhiều lắm!

I sincerely dedicate my thesis and my Master degree to all these wonderful people in my life—thank you.

Contents

List of Figures

List of Tables

1

Introduction

The Internet has emerged as a preferred destination for different types of health information, especially among younger adults [44]. In the European Union (E.U.), over 55% of E.U. citizens aged 16-74 stated that they used the Internet to look up information about different health topics [29]. Similarly, in Canada, it was estimated that up to 8.7 million adults also used the Internet for the same purpose [101]. Factors such as financial barriers and inadequate access to healthcare systems have increasingly driven people towards online sources for health content [16, 21]. Moreover, due to the COVID-19 pandemic in recent years, there has been a major surge in online health information seeking activities in many countries due to both the rising need for up-to-date, accurate health information and the inaccessibility of in-person medical care [63, 102, 110].

Since information seekers frequently use the Internet as a source for health content, it is understandable that they want to have a private, confidential browsing experience. That, unfortunately, is not the case for information seekers, as surveillance on websites has gotten more widespread and invasive [18, 67]. Moreover, web traffic surveillance has been shown to originate from a large number of third-party organizations: in the top one million most popular websites, 81,000 third-party organizations are present on these sites and conduct some degree of tracking [28].

This is a major cause of concern for Internet users, as such collection and sharing of user information to so many organizations can allow data collectors to gather user data from various sources to identify and target consumers for specific purposes, such as product

marketing and behavior manipulation. This problem becomes even more concerning when it involves potentially-sensitive personal health-related information. Not all Internet users are aware of this privacy invasion, however. Prior work has shown how unaware users are of tracking technologies as well as website's surveillance practices, along with the numerous misconceptions people have surrounding basic security and privacy concepts [62, 40].

As Nissenbaum pointed out in her theory of contextual integrity [72], "privacy is at best a culturally relative predilection rather than a universal human value". In other words, contextual factors such as cultural and regulatory ones influence privacy and the norms of information flow. This is also applicable to the field of online health seeking behavior, as Jia et al. have pointed out from their literature review how a person's country of residence significantly influences information seeking behavior [45]. Therefore, it is important to keep track of the status of surveillance on health websites as well as people's awareness of it, especially considering how the landscape of privacy constantly changes.

This thesis takes a detailed look at the nature of surveillance on health websites as well as privacy considerations from online information seekers, all from a Canadian perspective. More specifically, the main contributions of this work are:

1. A quantitative analysis of tracking methods employed on health websites that are relevant to Canadian residents as well as the extent of surveillance practices disclosure on these websites. Our results showed that non-commercial websites contain less user tracking than commercial ones, but that there were still major privacy implications from visiting non-commercial websites. Moreover, we also found that the majority of health websites, both commercial and non-commercial, are not fully transparent about their surveillance practices, omitting details about third-party organizations and their method of monitoring visitors from their privacy policies.
2. A model to classify the different levels of tracking observed from different websites. Each website's level of surveillance is categorized based on three main criteria: type of tracking technology, availability of defense strategies against tracking technologies, and the degree of user information dispersion to third-party organizations.
3. A qualitative analysis of an interview study of Canadian residents' considerations when seeking health information online. We found that our participants prioritized

content over privacy when seeking health information, but in some cases were aware that they were searching for sensitive information. Moreover, we observed how participants were unwilling to modify their seeking behavior even after being presented with accurate information about online user tracking, justifying it with one of three reasons: their assumption of the harmlessness of tracking organizations, their belief in the privacy of their seeking strategies, or their reluctant acceptance of tracking as a normal part of their seeking experience. We found that these reasons stemmed from either their great demand for health information due to healthcare inaccessibility, or their understanding and perception of tracking mechanisms, which was mostly inaccurate due to their limited knowledge of privacy technologies.

The rest of this thesis is organized as follows. In Chapter 2, we discuss relevant prior work in the field of online health information seeking, privacy, and web tracking. In Chapter 3, we describe our analysis of surveillance on health websites, including the methodology, results of our analysis, and the model of surveillance levels. In Chapter 4, we report on the interview study on information seekers' privacy considerations. In Chapter 5, we present our discussion of the results and the limitations of our work along with our recommendations for future work. Finally, in Chapter 6, we present a summary of the overarching results as a conclusion to this thesis.

2

Background and Related Work

Our research centers around the field of usable privacy, specifically privacy implications and considerations when seeking health information online. Therefore, the three areas of research we focus are online health information seeking behavior, end users' privacy concerns and awareness, and surveillance on websites. In this section, we explore the current state of development in these areas and how our work can potentially fill in the gaps resulting from the limitations of previous work.

2.1 Seeking Online Health Information

Previous work focused on the behavior patterns of information seekers as well as factors influencing how and why they search for health information [45]. Jia et al. reported that users rely on online health resources because they help users obtain answers to their health inquiries in a timely manner. This is especially useful when Internet users are facing many barriers trying to gain access to healthcare services, which can be caused by personal financial issues, scheduling conflicts, or difficulty in making appointments because of a shortage in personnel [16, 21].

Factors Influencing Online Health Information Behavior

In a national survey on adults living in the U.S., Jacobs et al. investigated different demographic factors and their correlation with the choice of health information source, covering

the Internet as well as more traditional sources such as healthcare professional, traditional media, and friends and family [44]. They found that a person's age, socioeconomic background and ethnicity all have a role in how they seek online health information: while younger and higher-educated adults with a higher social-economic status are more likely to choose the Internet as their source of health information, older adults who are of Hispanic descent and have low Internet skill still rely predominantly on traditional print media. Similarly in a sample of 49 Hong Kong residents, Chu et al. found that older and less educated Internet users are less likely to seek health information online [21]. LaValley also reported a discrepancy between how younger and older adults select health information sources [52]. In particular, younger health information consumers prefer content from commercial sites, while their senior counterparts focus more on academic ones.

Pian et al., with the help of an eye tracking system, investigated how 58 participants browsed a health discussion forum in three different usage contexts [80]. They found that the medical topics on which participants focused during browsing are different when they are searching for themselves than when they are searching for other people: when they do not have a particular health issue in mind, they browse more generally and visit more articles. Xiao et al., on the other hand, demonstrated the importance of trust in online health seeking behavior by finding evidence from an American national survey on cancer for the influence of a user's trust in online health information on their usage frequency. Moreover, they also found the correlations between trust in online health information and the diversity of searches as well as their choices of the channel where they search [107].

The Process of Seeking Health Information

Maon et al. discovered that 83% of their sampled population of Malaysian adults begin their search with common search engines, with Google and Yahoo being the two predominant services being used [64]. In contrast, only 15% of the respondents opted for specific websites when seeking details about health topics. In addition, some Internet users also rely on sources that are not conventionally used for health information, such as social media platforms. Augustaitis et al. conducted a focus group consisting of 26 transgender people in the U.S. and found that the majority of them go to social media platforms or online community forums for health information. Some examples for such platforms listed

by the participants included Facebook, Reddit, Instagram, and Discord [12].

When faced with multiple health information websites as a destination for health content, information seekers use a variety of selection criteria. Sillence et al. organized four 2-hour sessions where 15 British women who were experiencing menopause were asked to browse the Internet for health information related to their condition and concluded that their participants overwhelmingly based their trust and subsequently their selection of a site on characteristics of the site's content, such as the inclusion of relevant figures and the perceived objectivity of the information [91]. In contrast, their mistrust of health websites is based on the page layout. More specifically, features such as unnecessary complexity of a site design and overpopulation of pop-up ads lead to participant's rejection of a site. Similarly, Maon et al. found that the three criteria regarded as most important when choosing a health information website by half of their study's respondents are all related to a site's content, or more specifically, its professional origin, currency and ease of understanding [64]. This abundance and diversity of health information resources, however, do not come without disadvantages. Fiksdal et al. found that within a sample of 19 residents of Olmsted County, Minnesota, the main reasons for an information seeker to stop their health content search journeys are information saturation and fatigue [30]. In other words, after a long process of viewing numerous sites for health information which can get repetitive at a certain point, seekers become "tired with the screens" and "exhausted".

Sensitive Nature of Health Information Seeking

One factor that separates health information seeking from general information seeking is the inclusion of health data in the searches and the sensitive nature of such data. Rudnicka et al. defined sensitive data as data from which information about someone's behavior and routine, such as their location or their health details, can be inferred [86]. Ortega et al. extended this conceptualization by emphasizing that sensitive data retains "information on a person's behavior that is often (un)available to others" [77]. Because of how sensitive health data is, it makes sense that users are not comfortable with sharing such data in real life and therefore feel safer to search for queries that involve sensitive health data online thanks to the apparent privacy it offers [45, 64]. Some researchers attributed this uneasiness in sharing sensitive data and the subsequent preference for revealing such data

online, either on social media [10] or search engines [85], to the social stigmas linked to such data. De Choudhury et al. studied both social media platforms and search engines, specifically how users choose which type of platform for the disclosure of different kinds of sensitive data. They did this by analyzing a combination of large-scale logs from Twitter and a popular search engine, both over a period of 15 months, and a survey which was sent out to 210 Internet users [24]. In their research, their definition of sensitive health data is made up of two dimensions: the condition severity and the social stigmas linked to it. They discovered that when it comes to serious and socially stigmatized conditions as well as disabilities, users are more likely to look for more information on these topics on search engines than to disclose them on social media platforms like Twitter. Such platforms are only used to share symptoms of health conditions, as this is a way to "express the ordeals and inconveniences" that they face everyday.

2.2 End Users' Privacy Concerns and Awareness

With the rise of Internet and digital technology over the past two decades, research in the field of privacy has also grown and diversified significantly, spawning numerous subfields [93, 14].

Privacy Concerns and the Privacy Paradox

Some researchers concentrated on end user's concerns about the privacy of their online information and any potential consequences derived from the misuse of such data [43, 111]. Bergström, using data from a Swedish national survey, investigated the extent to which privacy concerns are expressed when performing four common Internet tasks [15]. They found that such concerns are varied and largely dependent on the specificity of the task in action, with the task of using debit cards for online payments being the one users are most worried about. However, the survey analysis also showed that users are not troubled by these concerns too often in their daily lives, having only thought "about them every now and then".

This aligns with the results from various studies which showed that despite expressing concerns about their online informational privacy, their real-life behavior do not neces-

sarily reflect this, as they still disclose their personal information when using the Internet [99, 95, 4]. This phenomenon is known as the "privacy paradox", first coined by Norberg et al [73]. However, some researchers believe that this paradox is not what is going on with Internet users. One such researcher is Solove, who argued that this "paradox" is merely a "myth" [94]. He believed that existing literature on the privacy paradox failed to acknowledge how much user behavior depends on context, and that the statements people made about privacy, collected through surveys and interviews, only reflect the broad concerns that they have but cannot be accurately applied to the multifaceted nature of their corresponding privacy behavior.

Nonetheless, there have been numerous attempts over the years at explaining this paradox [50, 32]. Choi et al. believed that this discrepancy in privacy concerns and behavior is due to something called *privacy fatigue* [20]. This phenomenon is described as the feeling of tiredness from users about protecting their online privacy, which is brought on due to the frequency of data breaches as well as the increasing complexity of controlling their online privacy. Using the concept of burnout and adapting their scales from prior literature, they conducted a survey on 324 Internet users and found evidence for their hypothesized impact of privacy fatigue on user's privacy protection behavior, or lack thereof.

Knowles and Conchie explored the concept of trust and how it could help to explain the paradox [49]. Examining 4 older adults and 6 younger adults through interviews and a survey, their study affirmed existing theories about privacy, such as how common Internet users have difficulty understanding privacy policies and practices as well as explaining in detail their privacy concerns. However, the most important finding they had was how users have *hopeful trust* when using services with privacy risks online. In other words, they concluded that people, motivated by the utility and the social implications from using said services, developed an almost unreasonable sense of trust in these services in order for them to justify and continue their usage despite their privacy concerns and potential privacy risks. This, in their opinion, proves that "hopeful trust enables the privacy paradox".

Applying iterative thematic coding on 13 semi-structured interviews and 187 surveys on the process of selecting and using mobile applications, Shklovski et al. found that the majority of their sampled Internet user population of 50 participants thought that "tracking is disturbing and/or creepy" [89]. However, while some are greatly concerned about it,

others accepted it and felt that it was justified, citing either "nothing is free" or "user has to comply" for the service of these applications. Moreover, some participants expressed what is evidence for the relevance of *learned helplessness* in privacy. *Learned helplessness* is defined as "repeated invasions into a persons' privacy and a conviction that there is no recourse", which then caused people to "stop responding to invasions even when presented with ways to defend themselves"—making this another possible explanation for the paradoxical behavior captured in users.

Some other researchers also claimed that this paradox exists due to what is known as *privacy calculus*—first conceptualized by Dinev and Hart [25]. This theory stated that when asked to make a decision involving their informational privacy, users would weigh the risks and potential gains from disclosing the information online. Thus, when users behaved paradoxically to their stated privacy preferences, it is possible that they have compared the gains and losses from informational disclosure and came to the conclusion that the gains overpowered the losses [50].

Privacy Literacy and Informational Self-Determination

Trepte et al. suggested that the reason why people display a disparity in privacy-related attitudes and behavior is due to a "knowledge gap hypothesis" [99]. They proposed that despite wanting to take measures to protect their information online that align with their privacy preferences, users' lack of privacy literacy bars them from doing so. Therefore, the authors argued that online privacy literacy would serve as a "stopgap" for this privacy paradox.

Seeing the importance of privacy literacy and the limitations of prior instruments for measuring privacy literacy due to oversimplification and reliance on user's self-assessment, Trepte et al. developed a scale to measure such literacy, known as "Online Privacy Literacy Scale", or OPLIS [99]. They defined online privacy literacy as one that comprises of two dimensions: factual knowledge, which refers to knowledge about the technicality of security and privacy as well as related laws and common data control practices by different institutions, and procedural knowledge, which refers to knowledge about the ways in which users can apply a defense strategy to regulate and control their online data. Using this definition, Trepte et al. conducted content analysis on 2597 extracts from 395 documents

focused on privacy which resulted in the final validated scale containing five dimensions, namely knowledge (1) about the practices of organizations, institutions, and online service providers, (2) about the technical aspects of online privacy and data protection, (3) about the laws and legal aspects of online data protection in Germany (as the investigators are based in Germany), (4) about European directives on privacy and data protection, and (5) about user strategies for individual online privacy control. This scale has since then been adapted to assess user's degree of privacy literacy in other studies [92, 8, 82, 38].

One reason why online privacy literacy is one of our research focuses is its role in supporting informational self-determination. The German Federal Constitutional Court defined *informational self-determination* as "the authority of the individual to decide himself, on the basis of the idea of self-determination, when and within what limits information about his private life should be communicated to others". This concept has become an integral pillar in the field of privacy as this highlights the importance of a person's rights to regulate what information about them is shared—the essence of privacy [84]. Because of this, in recent years, the enforcement of "informational self-determination" through privacy policy requirements has become the focus of privacy lawmaking by various governments. However, privacy policy regulation has been proven to be an ineffective method to gain informed consent for personal information disclosure from users [33]. Masur argued in his work that the best way to achieve self-determination, as well as self-data protection, is through the increase of user's online privacy literacy [65].

Increased Awareness and Its Influence on Behavior

This hypothesized effect of user's unawareness of privacy risks and misunderstanding of the implications of privacy violations on their behavior has been observed by numerous researchers. Herbert et al. found from their large-scale survey study sent out to 12,351 participants from 12 countries that Internet users still have some fundamental misconceptions about privacy and security when being online, misconceptions such as the belief that HTTPS is a predictor for how trustworthy a site is, or that regular password modifications are highly needed and provide enough security [40]. Moreover, they discovered that users from the same country of residence tend to have the same misconception, making this the strongest factor to estimate a person's privacy misconception and thus demonstrating the

need for region-specific research and solutions to help people deal with privacy risks and violations.

Kang et al. conducted an interview study with 28 different Internet users that includes both lay and technical participants [47], during which they are asked to perform different drawing tasks to illustrate their knowledge about how the Internet operates in the form of mental models. They found that despite having a more accurate understanding of the workings of the Internet as well as privacy threats, participants with a technical background are not necessarily more protective of their privacy in their actions than the lay people from the same participant pool. They found that while technical knowledge does not correlate to privacy protection behavior, respondents' awareness of privacy issues are more predictive of how they act with regards to the protection of their online privacy. They also concluded that their protection behavior is more influenced by the user's complex personal contexts.

Examining 419 adult Internet users via an online survey, Park's findings also align well with that from prior research [78]. Firstly, they found that not only over 40% of their respondents have some misunderstandings of institutional data practices, but the majority of them are also unaware of technical privacy terminologies and in general do not have any regulations around their online personal information disclosure. Secondly, although there was mixed confirmation of relations between a user's knowledge about methods for privacy protection and the actual protection behavior, there was indeed significant support for the relations between their awareness of surveillance practices and their real-life behavior.

Malandrino et al. corroborated this hypothesis with their user study of 36 university students [62]. Separating this group into two groups—Information and Communication Technology (ICT) group and non-ICT group—based on their respective field, they conducted a preliminary survey to collect demographic information as well as their self-declared knowledge on the Internet and online privacy. After this step, the participants were given time to interact with NoTrace, a "privacy-enhancing tool" developed by Malandrino and Scarano [61], which provides user details about the extent of tracking as well as potential risks from such tracking thus helping increase their privacy awareness. After a 30-minute session with NoTrace, they gave the participants a summary questionnaire to examine if there were changes in awareness, attitudes and/or behavior concerning privacy. From their interviews and questionnaires, they observed an increase in participant's pri-

vacy concerns, awareness of privacy knowledge and risks, as well as their motivations in pursuing protection measures against such violations. They also found differences between the ICT-group and the non-ICT group: people without an IT background are more likely to benefit from privacy-enhancing tools and consequently experience a more significant increase in awareness and changes in behavior than those who are in the field of IT.

In a similar study, Gerber et al. developed a mobile application called "FoxIT", which was intended to educate users on different aspects of privacy and analyze data practices of applications via their device permission requests, and evaluated it against 31 users over a span of two weeks [31]. In the end, they found that by increasing their privacy literacy through the use of this application, users became more proactive in defending their privacy against potential threats.

2.3 Surveillance on Websites

Surveillance Technologies

Several researchers have examined the complex evolution of tracking technologies. Bujlow et al. surveyed the literature on the different surveillance techniques implemented on websites and developed a taxonomy of different web tracking technologies, categorized by their method of data storage [18]. The final classification comprises five types of tracking mechanisms: session-only, storage-based, cache-based, fingerprinting, and others. They also provided a summary of available defense strategies for each tracking technology as well as a list of common tracking defense tools, such as Tor [98] and AdblockPlus [5], and tracking auditing tools that scan and reveal surveillance methods found on a given website, such as OpenWPM [28]. Despite also dividing existing tracking techniques by storage, Mayer et al. took a different approach and categorized surveillance methods into two groups: stateful tracking, which refers to methods that store tracking data on the client's device, and stateless tracking, which refers to methods that conduct surveillance in real-time as the client visits a website [67]. A more detailed explanation of this categorization is also included in our Section 3.3. Mayer et al. also provided a list of known defense strategies against online surveillance, although their list only contained three options: opt-out cookies, blocking, and Do Not Track. However, as new privacy protection methods and tools are introduced,

tracking technologies continue to evolve in order to evade detection and blocking from such tools. Lin et al. examined the use of domain-changing techniques by advertisers in order to dodge ad blockers, which block ads by using filter lists to detect known domain names used by third-party organizations for advertising [59]. After crawling 50,000 sites, they reported a taxonomy of four methods that were used to create replica ad domains, also known as RAD domains, which are domains that were of same advertising purpose as the original ones but were new domains registered by companies to avoid ad blockers. Lin et al. found that 10% of these sites are impacted by RAD domains, and 24% of these RAD domains have intrusive behavior that significantly decrease ad blocker's effectiveness in defending user's privacy.

Tracking on General Websites

As tracking technologies continue to grow and evolve, many attempts have been made to investigate the amount of tracking conducted on websites and provide users with tools that enable them to obtain such insights easily. Libert, using a self-developed software called webXray, analyzed the tracking done on top one million sites determined by Alexa and found that 88% of the pages sent out connection requests to third-party organizations [57]. Among these sites, a page initiated connections to an average of 9.47 distinct domains, and 36% of the requests are Javascript requests, which can be used for fingerprinting purposes. Moreover, 63% of the analyzed sites contained third-party cookies. Libert extended his analysis by developing and adding a new module to webXray called policyXray, with the purpose of analyzing the site's privacy policy and examining to what extent the policy discloses their tracking practices to users [58]. Collecting the policies from over 200,000 websites and 25 prominent third-party data collectors, Libert reported that despite how widespread data collection and sharing with third-party organizations is, only 15% of found data transmissions to third-party organizations are disclosed in privacy policies. Similarly, Malandrino et al. utilized their tool, NoTrace, to find what types of information third-party data collectors can do, either through direct collection of such data or through the aggregation of data collected from various sources. They discovered that a plethora of private and sensitive information, such as names, location, and sexual orientation, can be found and shared among multiple third parties through various tracking mechanisms. Moreover, Google received a significant 87% of such data leaked from tracking. More importantly,

they found that health terms, which they classify in their work as highly identifiable and highly sensitive data, are surprisingly leaked to nine out of ten of the most common data collector-aggregators.

Another aspect of tracking that many researchers focus on is the owner identities of these pieces of third-party tracking technology found on websites. Libert reported that Google is the company with the highest level of surveillance presence on websites, appearing on 78% of the analyzed websites via tracking technologies installed on these sites. Google is then followed by Facebook, Akamai, Twitter, comScore, Amazon, and AppNexus. The remaining 33 companies that were found are completely overshadowed by these seven, accounting for tracking on only 1–4 % of the sampled sites. Englehardt et al. conducted a similar experiment to that of Libert on the top one million sites, albeit on a later date and with the help of a different tool called OpenWPM [28]. From their analysis, they discovered the same results: Google remains the most prominent data collector, and the top companies, which includes some similar names (Google, Facebook, Twitter, Amazon, AppNexus) and a new addition, Oracle, are present on more than 10% of websites, while the remaining third-party organizations quickly reduce to a long-tail in the distribution of tracking third-party organization presence.

Tracking on Health Information Websites

As health information websites contain highly sensitive information, health privacy on websites is an area within web surveillance that is important and needs to be examined thoroughly. In this context, Libert replicated his previous study on particularly health websites and found the same degree of privacy intrusion on these sources [57]. More specifically, 47% of sampled health websites send HTTP requests to pages that can generate HTTP requests and manipulate browser caches, while 33% of them have JavaScript files that can be used to conduct surveillance via fingerprinting. Similar to other prior work, Google, comScore, Facebook, and AppNexus remain the top companies with tracking presence on sites, with Google being found on 78% of online health information sources. Burkell and Fortier conducted a similar analysis of surveillance on health websites, with the main difference being that they classified health websites based on two dimensions [19]. The first dimension is source: a site can either be retrieved from Google's search results, from

a list of recommended health websites that was curated by health professionals, or both. The sites are also grouped by their host organization: a site can either be governmental, not-for-profit, or commercial. Focusing on the detection of web beacons and cookies, they found that sites that are recommended by health professionals, especially those from a government or a not-for-profit organization, contain less tracking than those that are returned by Google search engine. More specifically, recommended sites contain on average 6.2 problematic beacons and cookies, while those found via Google contain an average of 14.1 problematic beacons and cookies on their site. Similarly, on average, recommended sites have 2.3 advertising beacons and cookies while sites from Google have 7.5 advertising beacons/cookies. As for the number of distinct third-party organizations found with tracking mechanisms on the sites, there are on average 3.5 domains tracking on recommended sites, while their Google counterparts have 13.2 domains present on their sites.

3

Surveillance on Health Information Portals

With the wide spectrum of possible surveillance technologies available for use on health portals [18], we realized that these portals might contain varied degrees of surveillance on their pages. This makes it challenging to analyze how numerous portals track their visitors. To support a systematic analysis of tracking, we developed a model to characterize the tracking technology deployed on a health information portal, adapted from previous research [18, 67, 48].

Our model distinguishes between five levels of tracking on health portals, in increasing degree of invasive tracking: **No Tracking**, **Minimal Tracking**, **Preventable Tracking**, **Unmanageable Tracking**, and **Invasive Tracking**. We assign a level to a website based on three dimensions of tracking: the type of tracking mechanisms present, the capability of users to defend themselves from said technology, and the degree of collected data shared with third-party organizations.

In Figure 3.1, we have the model visualized in the form of a horizontal decision tree. We report the decision criteria in Section 3.3, after describing the methodology for data collection and analysis in Section 3.1 and 3.2, respectively.

We investigated 22 health information portals and the nature of surveillance on these websites, as well as the disclosure of their data practices in their privacy policies. We

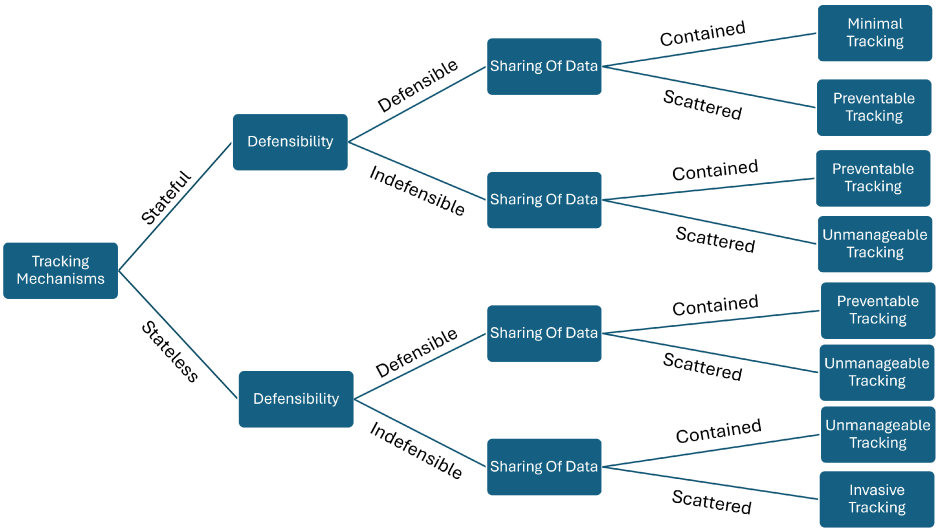

Figure 3.1: Visualization of the model of surveillance levels

covered both commercial and non-commercial sites, focusing on those that are relevant to Canadian residents. We were interested in three aspects of these websites:

1. **Organizations Observing Visitors.** Which organizations are tracking visitors of popular health information portals for Canadian residents that visitors might be unaware of?
2. **Tracking Level.** To what extent are visitors being tracked on popular health information portals for Canadian residents?
3. **Disclosure of Surveillance Practices.** To what extent are popular health information portals for Canadian residents disclosing their surveillance practices in their privacy policies to visitors?

3.1 Data Collection

For our analysis, we define a health information portal as a website that presents health information in the form of a structured collection of articles on health topics intended for the general public. Moreover, we focus specifically on portals that are relevant to Canadian residents. In this search for sites, we first sought portals that are under the authority

3.1 Data Collection

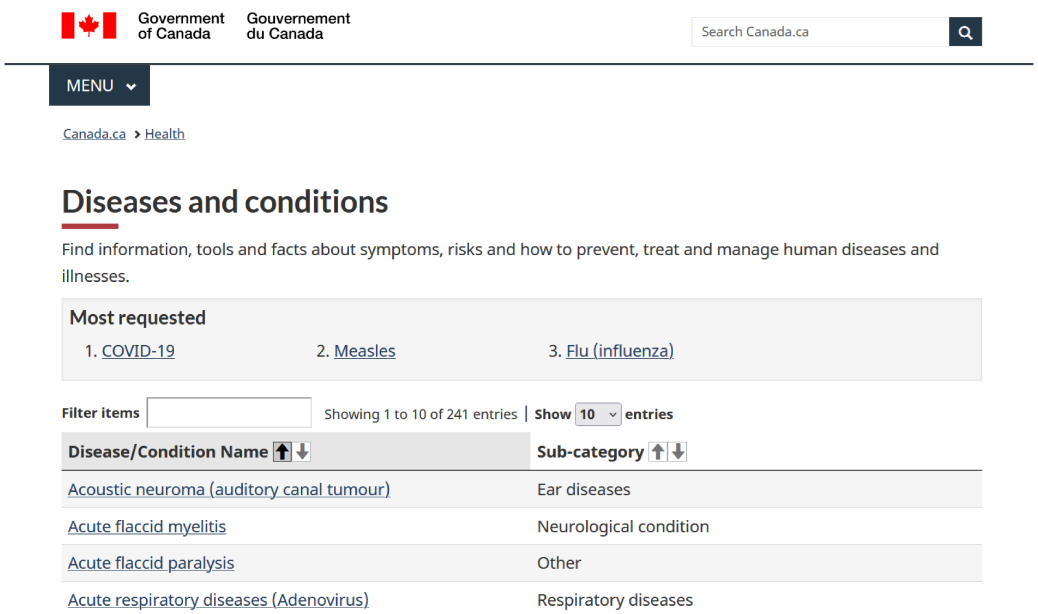

Figure 3.2: The content index page of the government of Canada's health information portal

of either the federal government or any government of the ten provinces and three territories of Canada, because in Canada, responsibilities for healthcare are shared between these two types of government [35]. Our portal search concluded that there is an official health information portal for the federal government as well as all ten provinces and one territory.[1] These portals are reported in Table 3.1, with the first twelve entries as the government portals within Canada. Figure 3.2 demonstrates the home page of the government of Canada's official health information portal. This inclusion of health information portals from provincial and federal governments illustrates the novelty of our analysis, as prior work from Burkell and Fortier on this topic focused more on health portals recommended by medical experts in both Canada and the U.S., which included portals hosted by private organizations [19]. As health information seekers might also rely on health information from sources unaffiliated with a Canadian federal or provincial/territorial government, we

[1]Nova Scotia's health information is split across two resources: one for general health articles, and for communicable diseases. For the purpose of the study we consider the aggregation of these two resources as a single portal. Although the Northwest Territories does not have a specific directory page for health articles, the government's website provides a directory page for all available articles that can be filtered by category. We consider the topics that are in the category *Diseases and Conditions* as a health information portal.

supplemented this list of portals to be analyzed with portals most likely to be visited for health information by the Canadian population. We found such portals by using the SimilarWeb service to obtain the 50 most accessed websites worldwide in the *Health* category according to their proprietary ranking algorithm (see Appendix 6.2). From this list of 50 sites, we developed a set of criteria to select the portals that align with our research goals the best as well as with the existing governmental portals we have already chosen. This set of criteria is:

- Selected health information portals must use English as a primary language.[2]
- Selected health information portal must provide information organized by health topic. For our study, we define a *health topic* as a symptom, condition, or treatment related to a person's health.
- Selected health information portal must focus on health topics rather than drugs information. An example for a portal on SimilarWeb's list that does not meet this criteria is https://drugs.com.
- Selected health information portals must be targeted to regular health information seekers rather than health professionals. An example for a portal on SimilarWeb's list that does not meet this criteria is https://medscape.com.

With these criteria, we identified ten additional portals, which includes six commercial health information portals, two US government portals (CDC and MedlinePlus), one UK government portal (NHS), and the portal of the World Health Organization (WHO), as shown in Table 3.1.

We made the distinction in this study between four different types of health information portals based on the organization operating them. For the portal under the authority of the Canadian federal government, we label it as *Federal*, while their counterparts with a Canadian provincial or territorial government is labelled as *Provincial*. As for other portals affiliated with a governmental or inter-governmental organization outside of Canada, they

[2]The portals for Canada, Manitoba, New Brunswick, Nova Scotia, Ontario, Prince Edward Island, and Québec provide information in both English and French, while the portal of the Northwest Territories provides information in English, French and different Indigenous languages. We studied the English version of the portals for consistency.

3.1 Data Collection

Portal	Type	URL
Canada	Federal	canada.ca/en/public-health/services/diseases.html
Alberta	Provincial	myhealth.alberta.ca/health/Pages/default.aspx
British Columbia	Provincial	healthlinkbc.ca/illnesses-conditions
Manitoba	Provincial	gov.mb.ca/health/publichealth/atoz_diseases.html
Newfoundland & Labrador	Provincial	centralhealth.nl.ca/health-information-a-z
New Brunswick	Provincial	gnb.ca/0051/site-e.asp
Nova Scotia	Provincial	novascotia.ca/DHW/azindex.asp novascotia.ca/dhw/cdpc/communicable-diseases.asp
Ontario	Provincial	health811.ontario.ca/static/guest/medical-library
Prince Edward Island	Provincial	www.princeedwardisland.ca/en/topic/public-health
Québec	Provincial	quebec.ca/en/health/health-issues/a-z
Saskatchewan	Provincial	saskhealthauthority.ca/your-health/conditions-diseases-services/all-z
Northwest Territories	Provincial	hss.gov.nt.ca/en/topics
CDC	Governmental	cdc.gov/health-topics.html
MedlinePlus	Governmental	medlineplus.gov/all_healthtopics.html
NHS	Governmental	nhs.uk/conditions/
WHO	Governmental	who.int/health-topics/
Cleveland Clinic	Commercial	my.clevelandclinic.org/health/diseases
Everyday Health	Commercial	everydayhealth.com/conditions/
Healthline	Commercial	healthline.com/directory/topics
Mayo Clinic	Commercial	mayoclinic.org/diseases-conditions/
Verywell Health	Commercial	verywellhealth.com/health-a-z-4014770
WebMD	Commercial	webmd.com/a-to-z-guides/health-topics

Table 3.1: Health information portals studied, sorted by type, then alphabetically.

are listed as a *Governmental* portal. As for any remaining portals, we automatically label as *Commercial*, since they involve commercial activities.[3] For consistency, we refer to the group consisting of the three types of governmental portals (both inside and outside of Canada) as *non-commercial* portals.

We saw that there is a wide variety of information each portal provides, specifically in terms of the amount offered. Some portals, such as that from the province of Prince Edward Island, only contain a few articles on public health topics relevant to their communities. On

[3]Commercial portals include portals under the authority of both companies and not-for-profit organizations, as the latter can incorporate commercial activities for fund-raising and related purposes.

the other hand, portals like Cleveland Clinic offer thousands of articles covering different health topics. Similarly, the extensiveness of the content of these articles is also quite diverse, with some portals only having articles that contain short summaries along with external links to other resources (e.g., MedlinePlus), and others having longer and more detailed articles (e.g., Mayo Clinic). Seeing how vastly different the content on each portal are from one to another, we estimated the amount of health information each portal offers.

We began our analysis by compiling an index of all articles provided on each portal. We first accessed each portal's homepage and searched for the directory page(s), as this would contain a list of all available health information articles on each portal. The identified directory pages are listed in the "URL" column of Table 3.1. The complete index of all articles can either be in the form of a single directory page, which is the case with most portals, or multi-level pages, such as the case of Canada where articles are indexed by each letter of the English alphabet. Once located, we used a Python script to scrape through the directory pages in order to extract the URLs of each article from each portal. Using a combination of heuristic-based filtering script and manual checking, we filtered through the list of extracted URLs from each portal to eliminate articles that do not fit into our definition of a *health topic*. An example is articles focusing on general lifestyle topics, such as "Active Children". After this process, we obtained a refined list of every health topic available on each portal as well as their corresponding URL. After finalizing the indices of all portals, we analyzed the portal characteristics and investigated the extent of surveillance on each of them.

3.2 Data Analysis

We employed different methods to analyze the collected data, specifically to analyze the portals' content characteristics as well as the state of surveillance on these portals.

3.2.1 Portal Characteristics

To determine how much health information each portal provides, we first randomly sampled 50 health articles from each portal's index, as a representative sample of the entire population of health topic articles from the portal. The reason for sampling the articles was

two-fold. First, some portals provide a very large number of articles, an analysis of which would require exhaustive use of web scraping technology and would therefore go against the terms and conditions of use from some of the portals. Second, a web page on which the article is published contains numerous elements, including navigation, branding, logos, tables of contents, and many more. Since not all of these elements are relevant to our analysis, combined with the difference in structures of each portal, it is difficult to automatically parse and identify parts of the articles that we are interested in. Therefore, we opted for a manual approach towards extracting and analyzing the content on these articles. We first developed a set of guidelines to define what is relevant health information on each article from the portals. More specifically, we focused on the main content from health articles including bullet points, and eliminated any navigational texts, menus, licenses, references, figure captions, table captions and tables. Then, we manually visited each article, then copied and pasted the relevant content onto markdown files. In total, we collected a sample of between 14 and 50 filtered health topic articles for each portal.[4] Using these sampled articles, we computed the average number of words and the estimated total number of words for each portal, which we calculated by multiplying the sample's average number of words with the number of articles from each portal.

3.2.2 Surveillance on Portals

For our other goal of investigating what types of tracking mechanisms and which organizations receive visitor data from each portal (i.e., answering our first and second research question), we employed the help of the Blacklight privacy inspector [66]. Blacklight[5] is a software publicly available as a web application, which automatically visits a website to monitor "which scripts on that website are potentially surveilling the user by performing seven different tests, each investigating a specific, known method of surveillance." This tool is developed and maintained by The Markup, "an American nonprofit news publication focused on the impact of technology on society" [1]. The seven surveillance methods detectable by Blacklight are: third-party cookies, ad trackers, key logging, session recording, canvas fingerprinting, Facebook pixels, and the Google analytics "remarketing audiences"

[4] Some portals provide fewer than 50 articles, which is below our sample size. In these cases, we include all articles from the portal.

[5] https://themarkup.org/blacklight

feature. Adjusting these technologies to the surveillance model in Section 3.3, we then reclassified these technologies as either stateful or stateless tracking, which can be seen in Table 3.2. In summary, with the exception of cookies, all the remaining tracking technologies are stateless tracking techniques. Keylogging and canvas fingerprinting are counted as a form of device fingerprinting and therefore stateless tracking techniques. Similarly, session recording is conducted using relay scripts [2]. These scripts are installed on the web page itself and therefore requires no data storage on the visitor's device, making it a stateless tracking technique as well. As for ad trackers, Facebook pixels, and the Google Analytics' "Remarketing Audiences" feature, they are all a subtype of web beacons and therefore are stateless tracking techniques.

Tracking Technology	**Defense Strategy**
Stateful Tracking	
Cookies	Disallow cookies [18]
Stateless Tracking	
Ad Trackers	PrivacyBadger [26]
Canvas Fingerprinting	Block JavaScript execution, use Tor browser [18]
Facebook Pixels	PrivacyBadger [26]
Google Analytics "Remarketing Audiences" Features	PrivacyBadger [26]
Keylogging	Partly by blocking JavaScript and Flash execution, Tor, VPNs, anonymous web proxies, clearing of browser web cache [18]
Session Recording	None

Table 3.2: Tracking technologies detectable by Blacklight and their available defense strategies, separated by our classification as either stateful or stateless as outlined in Section 3.3.

The Blacklight tool provides the option to inspect any website by providing the URL of the article. It can also emulate a connection from either a desktop or mobile environment, and the location can be from either the United States or Europe. For each portal, we executed the tool using as input the randomly sampled article from each portal's directory of health topic articles. We tested all portals using the desktop browser emulators, and only

the connection from the United States.

Our decision to use the United States connection instead of the European connection stemmed from our observation that Canadian privacy laws have a smaller scope of applicability than their European counterparts. The E.U.'s General Data Protection Regulation (GDPR) is a unified framework for user data regulation that is directly implemented across different countries from the E.U. However, the Canadian counterpart—the Personal Information Protection and Electronic Documents Act (PIPEDA)—is applicable to only parts of the country, as some provinces such as Alberta, British Columbia, and Québec have their own privacy laws [76]. This sectoral nature of privacy regulation from the Canadian government resembles the American government's approach to privacy, as they also leave the power to regulate the collection of personal data to state government, which resulted in the creation of different state-level privacy laws such as the California Consumer Privacy Act (CCPA) [74]. This existence of privacy regulations for different regions also meant that each framework is only applicable to a certain type of organizations. With the GDPR, any organization that either operates in the E.U. or interacts and collects data from E.U. citizens and residents, regardless of their base of operation and purpose of data collection, has to comply with their privacy regulations [105]. Privacy laws from North American governments such as PIPEDA and CCPA, however, are only relevant to private businesses that operate in specific jurisdictions and collect data commercially from residents of said jurisdictions (i.e., Canada for PIPEDA [76] and California for CCPA [74]).

Since European countries have stricter privacy regulations that apply to a broader range of data collectors, using a connection from this region on Blacklight would result in a GDPR-compliant version of the website that potentially has more limited surveillance. Therefore, we believe that despite Blacklight not offering an option for emulating a connection to these portals from a Canadian location, the American connection from Blacklight and the surveillance on health information portals resulted from such a connection would be a closer approximation to the surveillance taking place on these portals when accessed from Canada.

We ran the tests in May 2024 and collected the detailed reports produced by Blacklight. After each run of the tool, aside from the summary of the type(s) and the quantity of tracking mechanism out of the mentioned seven present on the input webpage, Blacklight

also provides raw results of the surveillance analysis on the input web page in the form of a downloadable archive.

With our script, we extracted from the raw analysis result files the domain name of each tracker and each cookie, the quantity for each type, whether the domain name is a third-party domain or not, and the owner organization of the domain. In order to obtain the owner organization information, we used the DuckDuckGo's tracker-radar database of tracker information.[6] This open-source database was established in 2020 and regularly updated by DuckDuckGo, a software company known for its privacy-focused search engine, to be a "data set of the most common third party domains on the web with information about their behavior, classification and ownership". The available details about these domains include the owner, domain's prevalence among top sites, source of information, and many other fields. As the entries on this data set are separated based on domain names without any subdomain, we removed the subdomains before using the filtered domain name to look up for the owner information on the tracker-radar data set. For example, if the extracted domain name in its full format is "prefix.example.com", the filtered version of this domain name would be "example.com". On the tracker-radar dataset, domain data is organized by region subdirectories, which includes the subdirectory US for the United States and subdirectory CA for Canada. As we adjusted the settings on Blacklight to emulate a connection from the U.S. yet our study wants to focus on Canadian residents, we pulled owner information from both subdirectories and compared them. The majority of the owner organization information are identical, with only seven domains found in the US subdirectory but not in the CA subdirectory. In these cases, we decided to adopt the owner information from the US subdirectory.

Given the evolving nature of the business landscape, we investigated the parent organizations of these owners as well, in order to have the full picture of the network of third-party domains on health information portals. To achieve this, we looked for the ownership information of each domain's owner organization using the Crunchbase database,[7] a company specialized in providing information on businesses, acquisitions, mergers, etc. With each organization, we verified the parent organization information, if there is one, obtained from

[6] https://github.com/duckduckgo/tracker-radar
[7] https://crunchbase.com

3.2 Data Analysis

Crunchbase by either examining the organization's official websites, or checking on search engines and news articles.[8]

This process of automatic parsing and manual investigation is repeated for both commercial and non-commercial portals, resulting in a compiled data set of all cookies and trackers found on these 22 health information portals. Once done, we aggregated information to compile a list of all organizations present with tracking technologies on each portal - by owner organizations and by their parent organizations.

We analyzed the portals' privacy policies by first obtaining the policies via the URLs listed in Table 3.1 and reading through the policies to look for any mention of two specific topics, all of which derived from our analysis of surveillance on these portals:

- Whether the policy contains detailed information on the third-party organizations present on their site with surveillance technologies: We read through each policy, then highlighted and extracted any explicit mention of which third-party organizations can collect users' information. We recorded all organizations mentioned and compared them to the list of third-party organizations actually present on the site from our Blacklight analysis and from that determine whether this portal contains partial or full information on data-collecting third-party organizations.
- Whether the policy reveals that data of the visit session is collected implicitly: Similarly, we read through the policy manually and we highlighted and extracted any explicit mention of data collection method in the policy, specifically for visitor data that is not directly collected via user input, such as trackers, cookies, key-logging, etc. We compared these mentions to our results from the Blacklight analysis to see if the policy is consistent with the actual tracking mechanisms being employed on the site.

From the results of these two aspects, we assess if the portal is consistent with results from our Blacklight analysis of the portals and their surveillance practices.

[8]For our study, we view parent organization as one that has fully acquired the child organization, and not just one who has obtained a majority stake in it. In one specific case (Mayo Clinic and Ziff Davis LLC), despite not being actually acquired, Mayo Clinic is officially declared to be a part of Ziff Davis LLC's health properties. This partnership allowed them to have direct access and control over collected visitor information, which is why Ziff Davis is categorized as the "Parent Organization" of Mayo Clinic.

3.3 Modeling Levels of Surveillance on Health Information Portals

We analyzed the tracking installed on each portal based on three dimensions: type of tracking technology, availability of defense strategies against tracking technologies, and the degree of user information dispersion to third-party organizations.

Types of Tracking Technology. We based our classification of the type of tracking mechanisms on the work of Bujlow et al. [18] and Mayer and Mitchell [67]. We distinguish between two categories of tracking technologies, depending on whether the underlying technology is *stateful* or *stateless* [67]:

- Stateful tracking: "Stateful tracking techniques store information on the user's computer and later retrieve it to recognize the user" [17]. Therefore, we consider stateful tracking the type of tracking where it is necessary to store files or data on user's device storage. As tracked data is locally stored and therefore can be located by users through the usage of publicly available tools, stateful tracking in our model is considered to be less harmful. Technologies in this category include:
  - window.name DOM property [18] - this tracking mechanism "uses a special Document Object Model property to store up to 2 MB of data" [18], meaning that there is storage of tracking data on the user's device, making them a type of stateful tracking
  - Storage-based tracking technologies from the work of Bujlow et al. [18], which includes HTTP cookies, Flash cookies and Java JNLP PersistenceService, etc.
  - Cache-based tracking technologies from the work of Bujlow et al. [18], which includes web cache, DNS cache, and operational cache.
  - Evercookies (supercookies) - A combination of various storage-based tracking mechanisms and therefore is "stateful".
- Stateless tracking: "Stateless technologies allow trackers to recognize users without storing any information on the user's machine" [17]. Using these technologies, the tracker can collect data from the users purely from their visit of the site and does

not need to locate previously placed files to perform their tracking operations. Since this type of tracking technology is not easily detectable from their storage system but rather through intricate forms of website analysis, it poses a greater threat to user's privacy. Technologies in this category include:

- Session-only tracking technologies [18], with the exception of window.name DOM property tracking, which includes session identifiers stored in hidden fields and explicit web-form authentication, both of which are performed directly on the website and therefore is stateless.
- Fingerprinting technologies [18], which includes network and location fingerprinting, device fingerprinting, etc.
- Web beacons/pixels, which are regarded as "ad trackers" by Blacklight [66]. They are "small 1px by 1px images that are placed on a website for tracking purposes by third parties" [66], which is loaded upon each visit by the user. Despite this technology not being explicitly mentioned as "web pixels" or "web beacons" by Bujlow et al., it is a very common tracking mechanism which, as seen by definition mentioned earlier, operates purely from the loading of visited pages and no storage of data is involved.
- Other tracking mechanisms (Category VI) from the work of Bujlow et al. [18], with the exception of evercookies, which includes, but is not limited to, timing attacks, headers attached to outgoing HTTP requests, and clickjacking.

Availability of Defense Strategies against Tracking Technologies. Tracking technology can be defeated by users with different degrees of ease, from trivial (e.g., cookies) to very challenging (e.g., browser fingerprinting). We keep in consideration whether there is a way for users to defend themselves from the employed technologies on the sites when needed. To assess a site's defensibility, we check the availability of a (proposed) defense strategy by Bujlow et al. [18]:

- If **all** tracking mechanisms on a site have a defense strategy based on Bujlow et al., the site is **defensible** for users.
- If **one** of the tracking mechanisms on a site does not have any defense strategy based on Bujlow et al., the site is **indefensible** for users.

- In the case of web beacons, which is not mentioned in the article by Bujlow et al., there is an available defense strategy, PrivacyBadger [26], which was also referenced by Bujlow et al. as a tool for protecting online privacy [18].
- In the case of session identifiers stored in hidden fields, although no defense strategy is known, Bujlow et al. explained that a strategy for this technology is not needed due to its limited usefulness in tracking visitors. Therefore, this mechanism is classified as **defensible**.

Degree of User Information Dispersion to Third Parties. The final aspect we consider is the information flow to third parties. More specifically, we check the spread of data sharing, or where the collected data is directed to. If the site shares data with organizations who are well-known data collectors with significant presence in the current website population, this site is high risk, because these organizations potentially already have detailed profiles from the various sources where they have their tracking mechanisms placed. Therefore, data collected from this site can be aggregated with that from other sites to create a more detailed profile of the user.

- Organizations that are regarded as major data collectors with significant tracking presence are the ten companies from the work of Karaj et al. [48] - **Google, Facebook/Meta,**[9] **Amazon, ComScore, Twitter/X,**[10] **Criteo, Microsoft, Adobe, Oracle** and **AppNexus**. These companies are the ones that dominate the distribution of tracking technology ownership detected on Karaj et al.'s sample of 1.5 billion web pages. After these ten companies, the distribution descended into a long tail, showing that each of the remaining companies with tracking technologies found on web pages only account for an insignificant share of the tracking distribution.
- If a site contains tracking originating from **more than one of these organizations**, we label the degree of information dispersion as **scattered**.
- If this is not the case, we label it as **contained**.

Using these three characteristics, we categorize a site's level of surveillance, as depicted in Figure 3.1. In total, we defined five levels of surveillance:

[9]Meta Inc. is the parent organization of Facebook
[10]X Corp. is the parent organization of Twitter

1. **No Tracking.** To be classified as "No Tracking", the website should **not have any tracking mechanism present on their pages**.
2. **Minimal Tracking.** Such websites can be viewed as only having necessary tracking mechanisms present for functioning, and this level of tracking is manageable, i.e., users can manage the tracking technologies and remove them if needed. To be classified as "Minimal Tracking", the website should only have **stateful tracking** mechanisms present, all of which must be **defensible** for the users, and user information collected when using the portal must be **contained**.
3. **Preventable Tracking.** To be classified as "Preventable Tracking", the website should:

    - Either have **stateless tracking** mechanisms present on the site, all of which must be **defensible** for the users, and user information collected when using the portal must be **contained**.
    - Or have only **stateful tracking** mechanisms present on the site. Should this be the case, then either all mechanisms present should be **defensible** for the users yet user information, collected through their visit of the portal, is **scattered**, or all mechanisms present are **indefensible** for the users but the user information collected when using the portal is **contained**.

4. **Unmanageable Tracking.** To be classified as "Unmanageable Tracking", the website should:

    - Either have **stateful tracking** mechanisms present on the site, all of which must be **indefensible** for the users, and the user information collected when using the portal is **scattered**
    - Or have **stateless tracking** mechanisms present on the site. Should this be the case, then either all mechanisms present should be **defensible** for the users yet user information, collected through their visit of the portal, is **scattered**, or all mechanisms present are **indefensible** for the users but the user information collected when using the portal is **contained**.

5. **Invasive Tracking.** The surveillance on these websites are alarmingly invasive, due to the advanced technologies being used as well as how widespread the collected

traffic data is and the unavailability of defense strategy against these technologies. To be classified as "Invasive", the website should have **stateless tracking** mechanisms present, all of which must be **indefensible** for the users, and user information collected when using the portal is **scattered**.

3.4 Results

In this section, we describe the results we have obtained from our analysis of surveillance on health portals. More specifically, we report the characteristics of the portals' content, the presence of third-party organizations on these health portals, the tracking level of each portal based on our model in Section 3.3, and the extent of surveillance practices disclosure in portals' privacy policies.

3.4.1 Portal Characteristics

In Table 3.3, we report our computations of the portal characteristics in terms of the amount of relevant content provided. The table is sorted by the size of each portal, i.e., the overall number of relevant words in all health topic articles on each portal.

During our process of analyzing the content from the portals, we observed an interesting similarity between the articles from three different portals, particularly those from Alberta, British Columbia, and Saskatchewan. A deeper investigation of these articles revealed that these health articles were all content licensed from the same source, namely "Ignite Healthwise, LLC". Upon further investigation, we learned that Healthwise is a non-profit company specialized in providing customized consumer health content for different healthcare organizations. Healthwise was acquired in 2024 by Ignite, a subdivision of Internet Brands—the owner of WebMD.[11]

3.4.2 Organizations Observing Visitors

In Figure 3.3, we show the number of organizations with tracking technologies present on the health portals. We observed that on the 22 chosen health information portals, there

[11] https://webmdignite.com/news/webmd-health-corp-acquires-healthwise-incorporateds-operating-assets-building-leadership

		Article Size		
Portal	**Articles**	**Mean**	**SD**	**Total Size**
Cleveland Clinic	3653	1681±187	673	≈6.14M
Alberta	2031	1034±216	779	≈2.10M
WebMD	1609	1147±199	717	≈1.85M
CDC	1580	758±388	1399	≈1.20M
Everyday Health	618	1853±353	1273	≈1.15M
Mayo Clinic	1152	843±92	332	≈970k
NHS	1202	643±88	316	≈770k
British Columbia	1140	563±156	563	≈640k
Verywell Health	326	1322±199	720	≈430k
Ontario	1381	262±39	141	≈360k
Healthline	150	1963±330	1190	≈290k
Medline Plus	1023	275±68	246	≈280k
Saskatchewan	293	747±213	770	≈220k
Canada	242	664±152	550	≈160k
WHO	192	507±30	108	≈100k
Québec	90	756±141	510	≈70k
Manitoba	78	540±103	371	≈40k
Newfoundland & Labrador*	26	509	296	≈10k
Nova Scotia*	31	414	253	≈10k
New Brunswick*	25	368	282	9189
Northwest Territories*	22	338	207	7427
Prince Edward Island*	14	475	314	6654

Table 3.3: Content metrics for selected health information portals, in order of decreasing total size. The article size metrics represent the number of words. The statistics for the portals indicated with an asterisk were computed using the entirety of the document population. For the other portals, we indicate the sample mean with the bounds of the 95% confidence interval computed using the sample standard deviation as an estimate of the population's, while calculating the total size of each portal by multiplying the number of articles and the mean article size.

are a total of 159 distinct organizations with tracking technologies present on the portals. On average, a non-commercial portal has 4.4 organizations tracking their visitors, while this number for a commercial portal is 70. However, these figures are lower when taking into consideration the parent organizations. In that case, there are only 137 organizations

3.4 Results

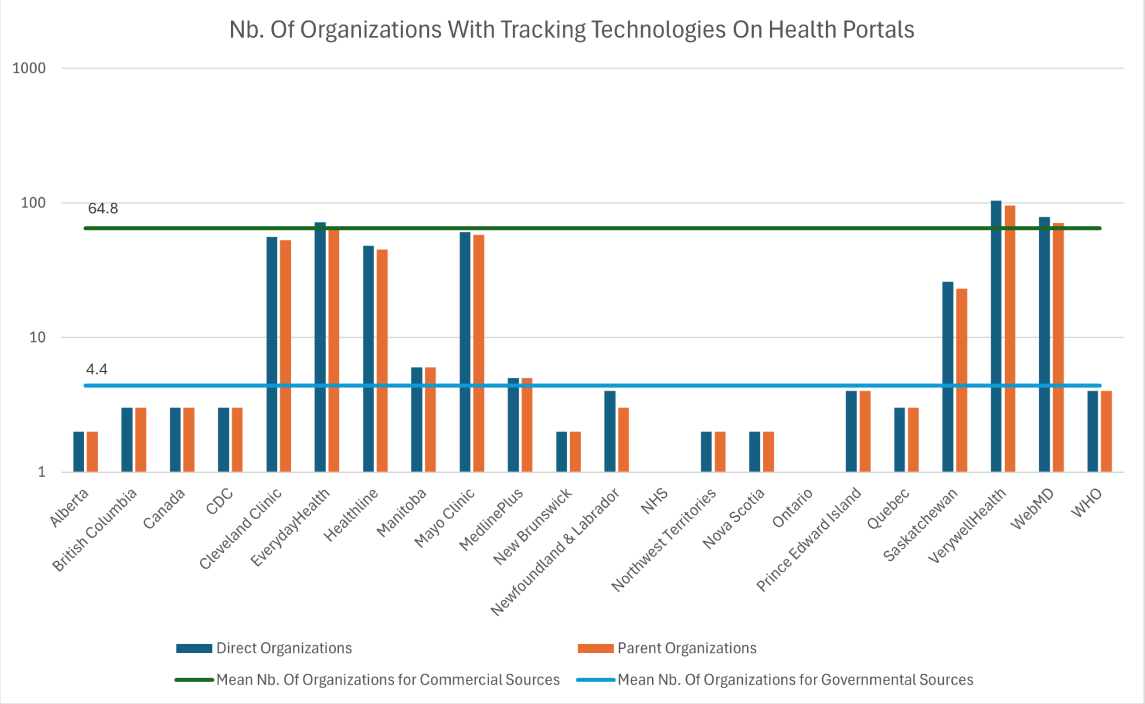

Figure 3.3: Bar chart of the number of organizations that has tracking technologies installed on health portals, in log scale.

with access to visitor surveillance on these portals. The average number of organizations present on the portals thus drops down to 4.1 for non-commercial portals, and 64.8 for commercial ones, as can be seen in Figure 3.3. The portal for Ontario has the lowest number of tracking organizations overall, with none detected. As for the portal with the highest number, Saskatchewan's portal has the highest number of organizations present on their site among the non-commercial portals (26), which is significantly higher than the remaining non-commercial portals. Overall, VerywellHealth has the highest number of organizations found on their site with tracking technologies, specifically 104 organizations.

Among the organizations detected, Google is the one with the most dominating presence on all portals, having had some types of tracking mechanisms installed on 18 portals out of 22. Following Google are Microsoft, with their tracking available on 10 portals, and Adobe, on 8 portals. Table 3.4 have details on the organizations with the highest num-

ber of appearances on health information portals on both commercial and non-commercial portals.

Organization	**Nb. of non-commercial portals ($N = 16$) they appeared on**	**Nb. of commercial portals ($N = 6$) they appeared on**	**Total**
Google LLC	12	6	18
Microsoft Corporation	4	6	10
Adobe Inc.	3	5	8
Magnite, Inc.	1	6	7
PubMatic, Inc.	1	6	7
TripleLift	1	6	7
OpenX Technologies Inc	1	6	7
LiveRamp Holdings, Inc.	1	6	7
Sovrn Holdings	1	6	7
Tapad, Inc.	1	6	7
Verizon Media	1	6	7

Table 3.4: Organizations with the highest number of portals on which they have tracking technologies installed

Taking a look at the 511 distinct domains found on these portals which are tracking portal visitors through cookies and trackers, we see some consistency with the results presented above. Google, Microsoft, and Adobe remain within the top organizations with most domain names registered to them, each with 34, 14, and 12 domain names, respectively. An interesting pattern we see, however, is that domain names are spread out when studying under the lens of child organizations - some of these child organizations only have a couple of domain names registered directly through them. However, when we combine the registrations from all child organizations into the parent organization's count, this number increased significantly. This is shown best in Figure 3.4: for example, Criteo SA only has seven domain names registered directly to them, but when adding the number of registrations from its child organizations through mergers and acquisitions, we see that they have a total of 13 registered domain names. Similarly, Ziff Davis LLC only has five directly registered domain names, but changes into nine domain names when including all subsidiaries.

We also investigated the degree of user information dispersion to third-party organiza-

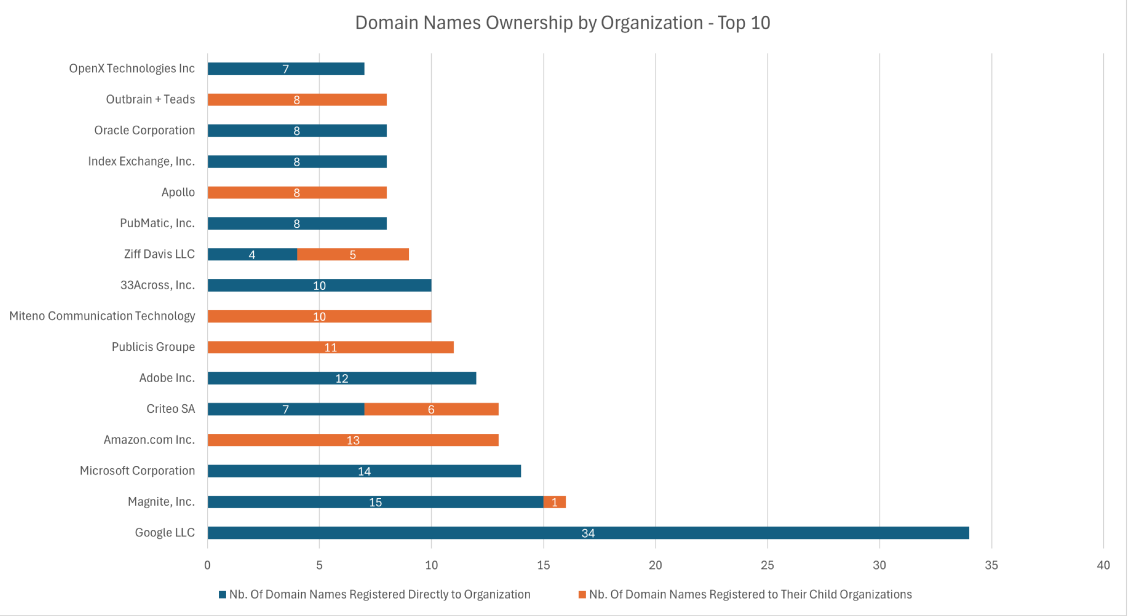

Figure 3.4: Number of domain names by organizations. The chart shows the organizations that have the top 10 highest number of domain names registered either directly to them or to their parent organizations. Blue parts in the chart indicate the number of domain names directly registered to them, while orange parts symbolize the number of domain names registered to other organizations that are a part of the parent organization

tions based on our method outlined in Section 3.3. We examined the list of organizations found with tracking technology on each portal and cross-referenced it with the aforementioned list of ten major data collectors - Google, Facebook/Meta, Amazon, ComScore, Twitter/X, Criteo, Microsoft, Adobe, Oracle, and AppNexus. Thence, any portal with two or more major data collectors found on their pages would lead to their classification as a site with **scattered** user information dispersion. The remaining would then be classified as **contained**. We show the results from this analysis in Table 3.5.

3.4.3 Tracking Level

To study how portals and third-party organizations conduct visitor tracking, we look into the results from our Blacklight analysis. The tool indicated that among the 22 chosen health information portals, surveillance from third-party organizations is conducted on nearly every health information portal we analyzed, with the exception of that from the UK's NHS

Health Portal	Major Data Collector Found on Portal
Contained	
Alberta	Google
British Columbia	Adobe
Canada	Adobe
MedlinePlus	Google
New Brunswick	Google
Newfoundland & Labrador	Google
NHS	(-)
Northwest Territories	Google
Nova Scotia	Google
Ontario	(-)
Prince Edward Island	Google
Scattered	
CDC	Adobe and Google
Cleveland Clinic	Adobe, Amazon, comScore, Criteo, Google, Microsoft, and Oracle
EverydayHealth	Adobe, Amazon, comScore, Criteo, Google, Microsoft, and Oracle
Healthline	Amazon, Criteo, Google, and Microsoft
Manitoba	Facebook/Meta, Google, Microsoft, Oracle, and Twitter/X
Mayo Clinic	Adobe, Amazon, comScore, Criteo, Google, and Microsoft
Québec	Google and Microsoft
Saskatchewan	Facebook/Meta, Google, Microsoft, and Oracle
VerywellHealth	Adobe, Amazon, comScore, Criteo, Google, Microsoft, and Oracle
WebMD	Adobe, Amazon, comScore, Criteo, Facebook/Meta, Google, Microsoft, and Twitter/X
WHO	Google and Microsoft

Table 3.5: Major data collectors with tracking technologies found on each portal. Portals are separated by their degree of user information dispersion determined by our model described in Section 3.3. (-) denotes that there were no major data collector found on that portal.

and the province of Ontario. Excluding tracking technologies from the portal's host organization, we saw that seven portals had ad trackers from third-party organizations installed on their sites. As for the remaining 13 portals, they contain not only ad trackers but also

cookies from third-party organizations, with seven of them being governmental sources and the other six being commercial ones. Moreover, the portal of Mayo Clinic was discovered by Blacklight to be the only one among the 22 portals to contain indications of session recording. This technology is also known as session replay scripts, and it has been shown to be highly invasive for end users, as it can easily lead to leakage of sensitive data from visitors to unauthorized third-party organizations [27].

From these findings of surveillance technologies present on health portals along with the calculated degree of user information dispersion shown in Table 3.5, we applied our model from Section 3.3 to these portals to categorize each portal's level of surveilling visitors. The resulting classification is presented in Table 3.6. As we can see, most portals are classified as either "Preventable Tracking" (9 portals) or "Unmanageable Tracking" (10 portals). Only Ontario is categorized as "No Tracking", NHS as "Minimal Tracking", and Mayo Clinic has the highest level of surveillance - "Invasive Tracking".

3.4.4 Disclosure of Surveillance Practices

We also inspected the privacy policies from these 22 portals to determine whether the details of such tracking have been fully disclosed to their visitors. A portal is only classified as "consistent" (marked as ✓ in Table 3.7) with our analysis if both the list of third-party organizations with surveillance presence on the portal and the list of tracking technologies employed on the portal obtained from the privacy policy match exactly to the list of organizations and tracking technologies we obtained from our analysis. If either of these condition is not met, the portal is then classified as "inconsistent" with our analysis. As Table 3.7 demonstrates, we found that out of the 22 portals we analyzed, only three sources - Québec, Ontario and Nova Scotia - are consistent with our analysis. In other words, only three health information portals disclosed their data practices fully and openly with their visitors.

We see that most portals did not disclose the full extent of customer tracking on their sites. However, a clear distinction between the commercial and non-commercial sources we observed is the difference in extensiveness of the policies. More specifically, non-commercial (governmental, federal, and provincial) portals overall contain more concise

3.4 Results

Health Portal	Type of Tracking	Defense Strategy Availability	Degree of User Information Dispersion
No Tracking			
Ontario	(-)	(-)	(-)
Minimal Tracking			
NHS	Stateful (Cookies)	Yes	Contained
Preventable Tracking			
Alberta	Stateless (Ad Trackers)	Yes	Contained
British Columbia	Stateless (Ad Trackers)	Yes	Contained
Canada	Stateless (Ad Trackers & Cookies)	Yes	Contained
MedlinePlus	Stateless (Ad Trackers)	Yes	Contained
New Brunswick	Stateless (Ad Trackers)	Yes	Contained
Newfoundland & Labrador	Stateless (Ad Trackers & Cookies)	Yes	Contained
Northwest Territories	Stateless (Ad Trackers)	Yes	Contained
Nova Scotia	Stateless (Ad Trackers)	Yes	Contained
Prince Edward Island	Stateless (Ad Trackers)	Yes	Contained
Unmanageable Tracking			
CDC	Stateless (Ad Trackers & Cookies)	Yes	Scattered
Cleveland Clinic	Stateless (Ad Trackers & Cookies)	Yes	Scattered
EverydayHealth	Stateless (Ad Trackers & Cookies)	Yes	Scattered
Healthline	Stateless (Ad Trackers & Cookies)	Yes	Scattered
Manitoba	Stateless (Ad Trackers & Cookies)	Yes	Scattered
Québec	Stateless (Ad Trackers)	Yes	Scattered
Saskatchewan	Stateless (Ad Trackers & Cookies)	Yes	Scattered
VerywellHealth	Stateless (Ad Trackers & Cookies)	Yes	Scattered
WebMD	Stateless (Ad Trackers & Cookies)	Yes	Scattered
WHO	Stateless (Ad Trackers & Cookies)	Yes	Scattered
Invasive Tracking			
Mayo Clinic	Stateless (Ad Trackers, Cookies & Session Recording)	No	Scattered

Table 3.6: Health portals and their level of surveillance classified based on our model described in Section 3.3, sorted by increasing level of tracking. Italicized portals are commercial, while non-italicized ones are non-commercial portals.

3.4 Results

Portal	Third-party Organizations Mentioned in Policy	Tracking Mechanisms Mentioned in Policy	Privacy Policy Consistent With Our Analysis
Nova Scotia	✓	✓	✓
Ontario	✓	✓	✓
Québec	✓	✓	✓
Alberta	✗	✗	✗
British Columbia	✗	✓	✗
Canada	!	✓	✗
CDC	!	✓	✗
Manitoba	✗	✗	✗
MedlinePlus	!	✓	✗
New Brunswick	✗	✗	✗
Newfoundland	✗	✗	✗
NHS	!	✗	✗
Northwest Territories	✗	✓	✗
Prince Edward Island	!	✓	✗
Saskatchewan	!	✓	✗
WHO	✗	✓	✗
Cleveland Clinic	✗	✓	✗
EverydayHealth	✗	✓	✗
Healthline	!	✓	✗
Mayo Clinic	!	!	✗
VerywellHealth	!	!	✗
WebMD	!	!	✗

Table 3.7: Results of our privacy policy analysis compared to our results from analyzing tracking mechanisms used on health portals. When a portal is marked "**!**", it signifies that there are some differences between the list of third-party organizations (or the list of implemented tracking technologies) obtained via the privacy policy and via our analysis in Section 3.4. Italicized portals are commercial, while non-italicized ones are non-commercial portals.

policies than their commercial counterparts. We can see this result in Table 3.8, as policies from commercial portals contain on average 7044 words. This number is almost five times higher than the average length of 1474 words from non-commercial portals' policies. Despite being beneficial from a readability viewpoint, such conciseness is only useful when they still contain clear and full details on the tracking practices on these sites. In contrast, non-commercial sites tend to not include some, or in some cases, any information on the

third-party organizations that are tracking visitor's data. For example, despite having the same trackers and cookies from LiveRamp—an ad technology company—Saskatchewan's health information portal did not include any details on this company in their policy, yet WebMD, VerywellHealth, and Healthline all mentioned it directly.

Portal	Nb. of Words in Privacy Policy	Average Nb. Of Words Among Portals of the Same Type
Non-Commercial Portals		
Alberta	332	
British Columbia	990	
Canada	1614	
CDC	3228	
Manitoba	320	
MedlinePlus	2140	
New Brunswick	578	
Newfoundland & Labrador	1329	1474
NHS	3769	
Northwest Territories	1699	
Nova Scotia	1487	
Ontario	1954	
Prince Edward Island	745	
Quebec	948	
Saskatchewan	1549	
WHO	902	
Commercial Portals		
Cleveland Clinic	2194	
EverydayHealth	12487	
Healthline	6265	7044
Mayo Clinic	4286	
VerywellHealth	8505	
WebMD	8529	

Table 3.8: Number of words in privacy policy of each portal, separated by the type of portal (Commercial vs. Non-Commercial portals), and average number of words in commercial and non-commercial privacy policies.

Another difference we noticed in the policies from non-commercial sources and commercial sources is that policies from the commercial sources contain various versions on

the same policy page to cater to different privacy regulations from different authorities, such as the E.U.'s GDPR and California's CCPA.

4

Privacy Considerations when Seeking Health Information

We conducted an interview study to investigate how people react to information about surveillance on health portals. In particular, we were interested in four research questions:

RQ1: What factors do users consider when selecting a site to visit for health information?

RQ2: How does a user perceive and understand the surveillance practices of the sites they visit for health information?

RQ3: What are the different strategies that users employ when seeking information they perceive as sensitive?

RQ4: How do users react to detailed information about privacy violations on health portals?

4.1 Data Collection

We recruited participants from departmental mailing lists at McGill University, as well as through personal contacts. This study was open to all Canadian residents 18 years or older, and we requested participants to complete an online consent form and questionnaire to be invited for the interview (see Appendix 6.1). In this questionnaire, we asked if the participant has anxiety browsing health information, then proceeded with questions pertaining

to their basic personal information. As we also wanted to investigate participant's privacy knowledge, attitudes and online interactions with websites, we complemented the questionnaire with sixteen questions on privacy attitudes and behavior from the 2022–23 Survey of Canadians on Privacy-Related Issues, conducted by the Office of the Privacy Commissioner of Canada (OPC) [75]. We selected this specific set of questionnaire items as this is the most recent official survey on the subject of privacy for the Canadian population—the same population which our study is also investigating.

In our first round of recruitment, we received 107 completed questionnaires, with fourteen additional respondents who did not finish the questionnaire since they answered "Yes" to having anxiety when reading health information. As we are interested in how participants react when presented with information on different aspects of health portal tracking, we wanted to make sure that there is a diversity in not only demographic factors such as age but also the levels of digital and privacy literacy among the participants. Therefore, we used a combination of convenience and theoretical sampling and designed a sampling framework to help us select participants in a way that allowed us to achieve balanced groups for the factors of age as well as self-reported level of digital and privacy literacy.

Since most of the 107 respondents from the first round were between 18 and 34 years old, we sent a second recruitment message to departmental mailing lists and personal contacts, changing the eligible age group for the study to people aged 35 and above to recruit older respondents. We received eight more questionnaire responses, making our final number of received responses for the questionnaire 115, excluding the unfinished questionnaires from respondents who stated that they are anxious looking at health information. From this pool of 115 respondents, we purposively sampled 20 participants based on their age as well as their digital and privacy literacy to ensure that we obtained the privacy perspectives of a diverse set of participants. All 20 participants agreed to take part in our online video-interview, and no one withdrew from the study. We began with 20 interviews, as an initial target number to balance a variety of perspectives with the data collection effort. However, after interview 15, we observed that additional interviews did not generate any new codes, and five further interviews confirmed the data saturation. Therefore, we concluded that 20 was a sufficient number of participants for our study.

Table 4.1 provides the characteristics of our participants as collected through our pre-

4.1 Data Collection

ID	Gender	Age	Education	Digital Literacy	Privacy Literacy
1	♂	65 or above	PhD	●	◕
2	♀	18–24	Bachelor's degree	●	◓
3	♂	25–34	Bachelor's degree	◔	○
4	♂	25–34	Master's degree	●	◕
5	nb	18–24	College	●	◔
6	♂	25–34	College	◕	◓
7	nb	18–24	College	●	○
8	♀	18–24	High school	◔	◓
9	♂	25–34	PhD	◔	◔
10	♀	18–24	Master's degree	○	○
11	♀	18–24	College	◕	◕
12	♀	35–44	PhD	●	◕
13	♂	25–34	College	●	◕
14	♀	55–64	Master's degree	◔	◔
15	♀	35–44	Bachelor's degree	●	◕
16	♀	55–64	Bachelor's degree	◔	◔
17	♀	35–44	Master's degree	◔	◔
18	♀	55–64	Master's degree	◔	●
19	♀	35–44	Master's degree	◔	◔
20	♂	65 or above	Master's degree	◕	◕

Table 4.1: Characteristics of the study participants.

interview questionnaire (see Appendix 6.1). For *digital literacy*, we asked participants to report their level of literacy using a 4-level ordinal scale (see question 9 in Appendix 6.1). As for their self-reported *privacy literacy*, we used a five-point scale, adapted from the OPC's privacy survey on Canadian residents (Question 12). The columns *Digital Literacy* and *Privacy Literacy* report on their answers, respectively, with proportionally darker icons representing higher levels of literacy.

The author conducted all 20 individual interviews through the Microsoft Teams video-conferencing software, with audio- and video-recording turned on for each of the interviews. Additionally, we also set up Teams to allow the participants to browse the web on the interviewer's computer. All the interviews were conducted in English. The inter-

views lasted on average 42 minutes ($SD = 8.2$ minutes). This protocol was reviewed and approved by the Research Ethics Board of McGill University.

We followed a semi-structured interview format [53] to encourage participants to speak freely about their experience seeking health information online. The interview was organized in four parts:

1. The first part was an administrative overview of the interview. We asked the participants to use "asthma" and "hypertension" as a placeholder for an anonymous condition, in cases where they need to discuss a specific condition.
2. In the second part, we asked the participant to assume that they suspect they have shortness of breath and asked them to walk us through a health information experience that mimics their last time searching for health information, by performing the search on the interviewer's shared screen from their device. As the participant sought health information, they were asked to verbalize their thoughts (RQ1). Afterwards, the interviewer asked the participants about their privacy awareness (RQ2) and their health seeking behavior (RQ3). This part of the interview was designed to learn about the participant's approach to health information seeking as well as their knowledge on privacy and visitor surveillance on health websites.
3. In the third part, we gave the participant a short presentation on surveillance technology present on most websites. This presentation included four major points:
    - Loading a web page connects visitors to third parties through various components on the page, such as social media buttons and ad placeholders.
    - Third parties collect data, such as URL requested, IP address, system configuration information, and interaction data, implicitly as soon as a page is loaded, despite visitors not actively providing them.
    - Web browsing data can identify visitors personally by linking collected data from various sources
    - Surveillance has real impacts on users. These include the risk of identity theft and scams [3], differential pricing [37], behavior manipulation [100, 108], and algorithmic discrimination [22].

This presentation was developed as a form of an intervention, where we wanted to provide users with information on contemporary surveillance technologies and the impacts on their privacy. With this intervention, we hoped to increase the awareness of privacy issues for participants who did not know such information before the interview. For participants who did, we wanted to remind them of these four points. In either case, we wanted to use this intervention as a precursor for the last part of our interview, where we further investigated reactions to this presented information.

4. For the fourth part of the interview, the interviewer asked the participant what information from the presentation was new to them, and how they felt about the provided information (RQ4). We included this part to explore how increasing such awareness might impact their future attitudes and actions when seeking health information.

4.2 Data Analysis

The author of this thesis went through the first five transcripts and extracted all quotes made by the participants that relate to one of the four research questions. Based on the extracted quotes, the author built a code book using a *descriptive coding* approach [70] where a short label summarizing the main point of the quote in relation to the specific research question is created. The author sent this initial code book to their supervisor and another collaborator for feedback, based on which the author updated the coding scheme. The author then continued with the quote extraction of the next five transcripts, revised the code book to also include the newly extracted quotes, and sent this version of the code book to the collaborators for another round of feedback. This process was repeated two more times, as the total number of transcripts was twenty. Overall, the code book was constructed in four iterations. We adopted this method of code book construction to ensure that the code book is not overgeneralized and could fully capture the ideas from all quotes. Moreover, by doing it in iterations, we also allowed the collaborators to be involved in the construction of the code book, instead of just including them at the reviewing process of the final version of the code book.

Once we extracted the relevant quotes from the twenty transcripts (297 quotes in total), the author and the two other collaborators validated the code book. We went through

this process by using the code book to assign codes to quotes from five transcripts randomly sampled from the original set of 20. We then compared our codes, discussed any disagreements we found and either resolved them or recorded quotes where a unanimous code could not be agreed on. Overall, out of the sampled 67 quotes from five transcripts, there were 22 quotes where there were initial discrepancies between the coders, and after further discussion, this number was reduced to only eight instances where we could not agree on a single code. For these eight quotes, we recorded and considered the codes given by all three authors during our analysis. This decision of allowing multiple codes for certain quotes resulted in certain implications for the final results, which is discussed further in Section 5.2. Once we finalized the code book, the author used the code book to assign code to the quotes from the remaining 15 transcripts.

4.3 Results

In this section, we describe the observations we made from our participant quotes in relation to the four research questions. Each subsection, from 4.3.1 to 4.3.4, corresponds to each research question. These results are a synthesis of the codes we have assigned to the quotes using the code book, as described in Section 4.2, but not the codes themselves. The main observations that are the answers to each research question are set in **bold**.

4.3.1 Website Selection (RQ1)

When faced with multiple health portal options provided by search engines after searching for a health topic, we observed that our participants evaluated health information portals by using eight different criteria: content quality, level of detail, recommendations from known sources, familiarity, ranking from search engine, name branding, and privacy and surveillance practices. We recognized that these factors can be further grouped into one of two categories: content-related factors, or source-related factors, which is similar to the findings from previous work on website selection criteria [97].

We found that for many of our participants, their decision to visit a health information portal depends on content-related factors. For instance, most of our participants focused on **the quality of the content**, specifically on the information's perceived accuracy and up-to-

dateness. For example, P_{16} explained why they preferred to visit information portals from health clinics: "*Yes, I prefer the clinical ones because I have more confidence in the [fact] that the information they're presenting is relatively unbiased and factual and quantified*". Similarly, P_8 said that they would select a portal, if "*it's a specific article with a published date, anything that's more recent is better*", citing that "*science is such a field where it changes all the time*". The **level of detail** of the information a portal provides is also of importance to some participants: "*So [specific health portal], I assume it's just gonna be bullet points, quick access information. So I would choose based on expecting [that assumption to be true]*" (P_3); "*[This portal,] typically I go there also because it's very quick and succinct information*" (P_{10}). These factors are identical to those mentioned from the prior work of Sillence et al. as indicators of a trustworthy health information portal that they would visit for health information, such as "informative content" and "unbiased information" [90].

On the other hand, we observed that our participants also considered a wide range of source-related factors when it came to the decision of choosing which portal to visit. For many, their trust in a health information portal, which eventually would lead to their visit of the site, depends on how they became aware of such portals. Some participants preferred portals that they knew from the **recommendations of a source they already know**, such as health professionals, friends and family members, as well as a previously visited article: "*So for example, I have friends that are doctors. If they one day told me, you know, [specific health portal] is not to be trusted for such and such, I would reconsider [visiting it] because I'm not a doctor*" (P_4); "*Yeah, [a site that I visit] would have to be referred from, say, a research article that I'm looking at, that's in a trusted source already*" (P_{20}). Others prioritized portals with which they had **familiarity**, due to previous experiences of using them: "*I've looked up symptoms on these [specific portals] before. [..]. And usually they're pretty helpful. So those I kind of trust, but I would have no proof or reasoning or anything to tell you why they're trustworthy*" (P_5), "*So I feel any websites that I've seen before, I tend to go back to [them], if they gave me the correct answer at the time*" (P_8). Participants also relied on **search engines and their page-ranking algorithm** for the choice of health information portals—"*I would say it's more of the top ones [that I visit] because I won't feel I'm expert enough to judge which one is more authoritative*" (P_{17}). This influence of search engines on website selection is consistent with the findings of Kammerer and Gertjets, who found that participants were less likely to

choose trustworthy health portals when they were presented toward the bottom of a search engine result page than when they were at the top of the page [46]. However, there are also cases where participants simply chose a site because of the **name branding of a site and its authors**, without any recommendations or previous familiarity: "*But honestly, if I were to click on American Lung Association or say, Diabetes Canada or whatever, I would probably not [check] the legitimacy of it further than just the name*" (P_{10}).

Some participants stated that they would deliberately avoid sites with a **commercial mission**. For instance, P_1 said: "*I tend to avoid things that say sponsored. They were often pushing some specific drug or cure*". Similarly, P_{13} explicitly mentioned that they had higher trust in any site that is not a commercial website, because "*if they're trying to get money [from you], they are not your friend. Simple as that*". This avoidance of commercially-motivated portals extended to sites that are marked as "sponsored" by search engines and therefore appear at the top or bottom of the result page from these engines: "*Because of my experience with just any other normal searches, that when I do search for something, Google tends to present me with paid [promoted websites]. So yeah, it says sponsored here. So those at the bottom there, I would completely avoid*" (P_9). This observation mirrors the findings of Lewandowski and Kammerer, who concluded, from their literature survey of research using eye-tracking to study viewing behavior on general search engine results page, that participants tended not to click on these "sponsored" sites despite paying some visual attention to them [55].

Two participants also used artificial intelligence (A.I.)-powered chatbots for seeking health information. P_{17} explained why they opted for these tools, saying that: "*It's easier to navigate, it's more user friendly. I can just keep asking questions. I don't need to go through that process of screening the information by myself because I feel it sort of can prioritize the important things for me and give me the integrated version of what those individual answers from Google or the search engine could generate*". P_6 demonstrated similar sentiments about A.I. chatbots, specifically ChatGPT: "*it summarizes all of this research, all of these results [from] the Google results and a lot of other sources in way faster [approaches] and more compact form than any human can ever [do]*". These testimonies echo the results of the work of Al Shboul et al. [7] as well as Shahsavar et al. [88], who both discovered that participants are increasingly employing chatbots like ChatGPT for seeking health content and using that to aid their self-diagnosis, due to their accessibility and efficiency.

Few participants considered a portal's state of security and privacy when deciding which portal to visit. In fact, only three participants exhibited any signs of acknowledgement of a portal's surveillance practices when evaluating health information portals. These participants include P_{20}, who said that they would only go to portals where "*they're not gonna be tracking people to follow up with you*". Similarly, P_4 stated: "*If I was aware of a website that scraped information or was doing something a little bit less trustworthy, I might stay away from it. The other thing I might avoid is if I was aware of a recent data breach or hack, I might stay away from it, because even though the web page is loading, who knows [...] whatever [security flaws] are still lingering there, right?*". These statements beg an interesting question—how would these participants know about the tracking practices conducted by their visited portals? What are the steps, if any, that they take to investigate and analyze the surveillance that takes place on these portals? Moreover, was the lack of consideration for the privacy safety when choosing a health information portal observed in the remaining participants influenced by their understanding of visitor tracking on such portals, or lack thereof? We investigated further the comprehension of our participants of health information portals' surveillance practices, the answers to which are detailed in the following Section 4.3.2.

4.3.2 Surveillance Awareness (RQ2)

When examining the extent of the participants' knowledge on online tracking during their health information seeking process, we discovered that despite their differences in backgrounds, digital literacy as well as privacy literacy, the majority of our participants actually had a similar understanding of how their web browsing activities are observed and collected. More specifically, we found that these perceptions of how surveillance works were superficial and, in many cases, inaccurate.

When explicitly asked if they knew the kinds of user information tracked by websites during their visits, many participants gave concrete examples of the types of collected data. However, we noticed that **participants tended to only focus on one single category of information that is collected by health portals**. For instance, some participants only brought up information specific to each browsing session such as cursor movements or link navigation: "*I assume that what's being observed is what I click or how long I stay on a particular page. I would assume that once I'm on a page, it can kind of track in a holistic sense*

what pages I'm going in between, mostly associated with the time spent on a page or how I navigate up and down the page and what I click on" (P_9). Others would only mention the collection of device information: "*I knew that, on average, advertisers and analytics would collect data and I knew they would know my system type, the version of the browser, and my configuration or whatever*" (P_7). This focus on specific categories of information helped explain why many of our participants were shocked to find out from the intervention how extensive the collection of user information by health portals is, with some even surprised to learn that these different data fields can be combined and aggregated from different sources to build consumer profiles for targeted marketing: "*I think I wasn't aware of the way that they have and can match your data from visiting other places too. I would have expected that they do [the matching], but maybe not to the point of being able to get your name and stuff like that*" (P_3).

Similarly, **although our participants did know about the existence of third-party organizations with surveillance presence on health portals, they only named Google when asked if they knew where the collected data went to specifically**: "*But I do believe that the companies, like Google, do have access simply because that's the exchange that we have for having [the service]. I feel like it's an implied deal that we have with Google or Firefox or any other company*" (P_{19}). This level of awareness, specifically the sole focus on Google as the major data collector, is similar to that from the majority of Kang et al.'s participants [47]—most of their participants were also only aware of Google as a destination for collected data. This, once again, is not fully accurate: our results from Chapter 3 have shown that while Google is indeed the dominating organization when it comes to user tracking across multiple health portals, there are hundreds of other organizations with varied degrees of surveillance presence on these portals as well. From this observation, we can see that our participants were not fully aware of the diversity of data collectors that are responsible for user information collection on health portals.

This superficial knowledge of tracking organizations became more apparent later in the interviews, as many showed signs of surprise and shock having learned that there are more companies than just Google that conduct invasive tracking on sites: "*OK, one of the ones that I'm a little surprised about is probably slide number 1 [which talked about the presence of third parties on health portals]*" (P_{14}). P_9 also talked about how incomplete their understanding of these organizations might be, saying that "*I'm aware of [these tracking third parties], but*

I don't know that much about them, I kind of just think of them in a very general sense as these companies who sort of collect as many data points as possible".

When it comes to the mechanisms behind how tracking is conducted on health portals, we found that **many participants were in the dark about how such online tracking takes place**. For example, when asked if they knew how information is tracked and collected from them, P_{11} said outright: "*I don't know what kind of devices the sites themselves have in place to track that kind of thing. I'm not super tech savvy*". The only participants who were aware of the technologies used to conduct surveillance on health portals are ones with the highest degree of digital literacy (i.e., a background in Information Technology (I.T.) or Computer Science). However, even in these cases, we saw that this knowledge of tracking mechanisms can be limited both in depth and diversity. More specifically, most of these digitally literate participants neither knew about the mechanisms behind tracking techniques—"*Cookies, I actually don't really know what it does. I just see it all the time on websites, so I always accept it*" (P_2)—nor were aware of the wide variety of tracking techniques available on sites, as they only mentioned cookies as the technology being used by health portals to collect user information. However, our analysis of tracking technologies used on health portals in Chapter 3 showed that portals employed a wide range of techniques to monitor visitors and collect data, including canvas fingerprinting and ad trackers, which were only mentioned by three of these digitally literate participants. This lacking comprehension of how web tracking works was even more apparent when we look at how they misused specific privacy-enhancing tools to seek sensitive information (detailed in Section 4.3.3), as well as their conviction that there is nothing further than can be done to prevent tracking (detailed in Section 4.3.4).

In contrast to their knowledge about how tracking technologies work, we observed that **most of our participants have a higher, yet still limited, level of awareness of available privacy protection strategies**. We also observed a connection between this knowledge on protection strategies and age groups: while almost all of the younger participants (18–34) were able to name a tool that they knew and/or used to defend their online privacy, only one older participant (35 and above) mentioned something about protection strategies and their experience with using such strategies. This result is consistent with the findings of previous work that older adults are less likely to be aware and adopt strategies to help them

protect their online privacy [96, 112].

Among those who did talk about protection strategies, we saw that they only mentioned one of three strategies: virtual private networks (VPN), private browsing mode, or privacy-enhancing browser extensions, such as ad blockers. For example, P_{13} talked about how they used VPNs to protect their identities: "*No, they can't track me by my IP. They think I'm in 13 different countries at once because I've got a two-step VPN always enabled*". Other participants, like P_9, stated that they blocked ads with browser extensions: "*So on my web browser for example, I have multiple extensions that I use. One of them is a sort of ad block just to sort of navigate without having to interact with as many ads*". This preference for such tools could stem from the popularity of these three protection methods, as prior work from Story et al. have shown that the majority of their survey respondents have heard of and used VPNs, ad blockers, and private browsing modes [96].

However, as Story et al. have also pointed out, a person's awareness and experience with a protection strategy does not necessarily mean that they correctly understand how it works. We observed this discrepancy in our participants, with some participants having clear misconceptions about the mechanisms behind their adopted protection tools, which led to them believing that such tools had provided them with their desired level of protection, when in reality they do not. One such case is P_{12}, who, after being presented with factual information about what type of information is collected on portals and how, were still convinced that "*this data collection doesn't happen when you are using private mode*". These misconceptions are quite dangerous, as participants could become inappropriately confident in the safety of their browsing behavior, even though the mentioned strategies have been shown to only be partially effective (or not at all in the case of private browsing) in protecting and providing anonymity for users' online activities [69, 96]. This inappropriate confidence is more evident later in the interview session (see Section 4.3.4), as some participants expressed that their existing precautions for privacy protection were sufficient enough for them not to improve their browsing behavior, even when presented with information about how invasive tracking can be on health portals.

Overall, we can see that the majority of our participants demonstrated low levels of privacy literacy, expressing only surface-level awareness of the technicality of surveillance on web portals. More specifically, we observed that most of them were unable to elicit

concrete information about different aspects of surveillance, such as tracking organizations or types of collected data. In line with the pre-interview questionnaire, we did observe a significant correlation between a participant's digital literacy and privacy literacy: people with a higher level of digital literacy, specifically those with a background in I.T., are more knowledgeable about the mechanisms behind online tracking ($\phi = 0.58$, $p = .007$). As privacy literacy has been shown to influence how people browse and interact online [13, 92], we were interested in learning how such low levels of awareness of surveillance practices manifest in their information seeking behavior, especially when it concerns information they perceived as "sensitive" and thus have a higher need for privacy. We examined this in the following Section 4.3.3.

4.3.3 Strategies for Seeking Sensitive Information (RQ3)

In this section, we look into the different strategies, if any, our participants employed when seek sensitive information. We employ the same definition of a "sensitive health topic" as the one used by De Choudhury [24]—a health topic's sensitivity is made up of two dimensions: the severity of the condition involved and how socially stigmatized that condition is. A health topic is considered "sensitive" when they are highly stigmatized and/or they concern a serious condition. With this definition, we found that some participants employ specific protection strategies, such as using private browsing mode and refrain from disclosing sensitive data online when seeking highly sensitive health information.

Some participants also proactively protected their sensitive information seeking session by employing a specific privacy protection strategy. One strategy is **using private browsing mode**—P_3, who would go into private browsing mode on their browser when seeking sensitive health information, said: "*I have the intuition, [...], I'll open Incognito if I expect it to be a thing that they may track more*". P_{13} stated that they would act similarly, that "*If it's very private, I usually switch to a burner browser, not on a burner phone or anything. Just no cookies, no history*". However, this belief that private browsing provides them protection against tracking is incorrect and a common misconception, as private browsing mode is not able to defend users against tracking technologies such as fingerprinting and ad trackers that can ultimately still collect these sensitive fields of information [106]. The fact that these participants misunderstood how protection strategies work not only align well with the

limited awareness of surveillance previously observed in Section 4.3.2, but it also demonstrated how surveillance misconceptions like these could lead our participants to engage in interactions that ultimately make them vulnerable to privacy threats.

Meanwhile, other participants employed the strategy of simply **not searching for sensitive health topics online altogether**. For example, P_{17} would skip online searches and directly make appointments with health professionals when they are concerned about a serious condition: "*If it's very everyday illness like some common illness, I'll use what I have [to] describe to Google, but if things get very serious or I know that it requires very professional help, I won't go there in the first place*". P_4 shared a similar mindset, focusing more on their desire to hide such sensitive information from other parties on the Internet: "*If it was something more sensitive, I may not look it up at all. I may go directly to trusted sources, like my doctor or trusted friends who are medical professionals and reach out directly to them*" (P_4). This belief is another myth, however: although active withholding of data can indeed reduce the amount of information online, especially sensitive ones, research have shown that by aggregating data from various visited sites, data collectors can easily infer different information about user, even those that were not directly disclosed by them [103], showing a weakness of this strategy for privacy protection.

For some participants, sensitive health topics did not lead to the implementation of specific privacy protection strategies. However, it did influence them to alter their website selection process and prioritize specific types of information portals, based on the quality of the portal's content. Although these strategies are more motivated by the participant's information needs than their sensitivity to surveillance, we found that there were some major privacy implications to the reported website selection strategies reported.

For instance, with serious conditions and problem such as COVID-19 (public health crisis) or cancer, several participants said that they would slightly tweak their website selection criteria as discussed in Section 4.3.1 to focus more on authoritative sites, such as governmental or medical institutional sources, because they wanted to obtain the most accurate, up-to-date information that was developed and verified by health professionals aimed at the public: "*But the more serious and important information I want, the more I'll be looking for, [in] public journals or a health authority. Like during COVID, I would only ever look [in health authority websites], I was in Vancouver and I would only be looking at the British*

Columbia website. I don't remember ever looking anywhere else" (P_3). A problem with this preference for authoritative portals, however, is that authoritative health portals are not free of surveillance, as our analysis from Chapter 3 has demonstrated (see Table 3.6). In fact, many of these governmental and institutional portals contain tracking technologies from major data collectors that can also be found in commercial ones, showing that it is not necessarily safer privacy-wise to look up sensitive health information in authoritative (also known as "non-commercial" in Chapter 3) portals compared to their commercial counterparts.

Similarly, some participants would opt for social media platforms when seeking sensitive topics such as mental health issues, which has been shown to be publicly stigmatized [11, 87, 109]. They explained that this decision of moving to social media is due to their preference for personal testimonies: "*[Searching on forum]'s something I would do for mental health issues [...] Because I think it's really important to get the perspective of people with the disorder, to understand the inner experience*" (P_5). This reasoning resembles that given by transgender participants from the work of Augustaitis when they were asked why they would prefer social media platforms for health information and advice [12]. However, visiting social media platforms specifically for sensitive health information could cause major problems for the users, as social media platforms have previously been shown to consist of numerous privacy and security issues [9].

As we have seen, despite taking specific steps to separate between seeking sensitive and regular health information, these strategies are insufficient in protecting users against the surveillance conducted on health portals. We observed that participants employed these limited protection strategies largely due to their incorrect and incomplete understanding of web tracking, as observed in Section 4.3.2. Prior work has shown how people with lower levels of online surveillance awareness became more concerned about their privacy and expressed intentions of changing their online behavior after learning more about tracking practices [31, 62]. Motivated by these findings, we wanted to see if our participants would have similar reactions when we intervened and provided them with accurate, up-to-date information about surveillance on health portals. Thus, we investigated the participant's reactions in RQ4, the results of which are described in the following Section 4.3.4.

4.3.4 Reactions to Surveillance on Health Portals (RQ4)

We found that in spite of their inadequate knowledge on health portal surveillance, all of our participants expressed that they did not anticipate any significant changes in their information seeking patterns even after knowing more about web tracking from the intervention. Taking a closer look at the participants' elaborations, we noticed that they rationalized their unwillingness to modify their information seeking behavior with one of three reasons: their trust in the tracking parties, their belief in the safety of their browsing behavior, and their reluctant acceptance of tracking.

Several participants said that they did not expect a change in how they seek information after the interview because of their **assumptions about the harmlessness of tracking**. Some participants did not think that online surveillance is concerning because the parties in charge of such surveillance are not malicious organizations. P_9 communicated this sentiment, believing that surveillance on portals are conducted by governmental agencies, which to them are not worrisome: "*I know that there's a lot of creepy data that can get really precise about me and my movement in the world and the things that I do. And that's just generally weird to me and uncomfortable, but for some reason I don't really know what would the government do with it that would negatively affect me*". We observed this trust in tracking parties, specifically governmental ones, in the pre-interview questionnaire as well, where the majority of our participants responded that they believed the amount of online activity tracking conducted by governmental parties are either equal or, in most cases, less to that done by commercial companies and organizations. For other participants, user tracking on websites was innocuous because they believed that the motives behind such surveillance of these organizations are inherently innocuous. For instance, P_{13} said: "*Ultimately, I think it depends largely on where it goes and the larger analytics websites that are making a profile on you. They're not making a profile that's going in a manila folder in some sketchy organization somewhere that's [going to] send a guy in a trench coat to follow you. They're doing it to get ad revenue ultimately, and I don't necessarily feel that that is a nefarious goal*". These findings resemble those of Melicher et al., who concluded that users are comfortable with tracking as long as they originate from sources they were personally familiar with and the intentions behind such collection are reasonable [68].

The question here, however, is how do they know about the tracking parties that are present on health portals and their surveillance practices in order to evaluate it as trustworthy? We observed that these judgements do not come from concrete evidence of tracking, but rather from misunderstandings that participants had about tracking parties. In particular, we noticed that participants thought that it was the hosts of the portals themselves that conducted harmful visitor tracking. We saw this way of thinking with the previous example of P_9, who assumed that "*government website would have less ad tracking [or] cookies*", and even then, such levels of tracking were not worrisome because the government was in charge of the user tracking and there was nothing that the government could do with the collected data to "*negatively affect*" them. This perception of tracking is inaccurate, however, as our analysis on health portal surveillance in Chapter 3 have shown that it was not the portal hosts like governmental agencies that were in charge of data collection and user surveillance, but third-party organizations, such as Google, who provided websites with analytics services or advertisements. In fact, portal hosts only claim to have access to very general statistics about their visitors from these third-party organizations. These organizations, on the other hand, were the reasons why visitors should be wary of online tracking, as they had access to a significant amount of fine-grained user data from the various sources on which they installed numerous pieces of tracking technologies that can identify each visitor personally. Moreover, the fact that multiple third parties had access to user data meant that privacy breaches to such parties would then become much more disruptive to users, demonstrating that our participant's other argument of "it is just about ad revenue and thus not worrisome" is problematic. These findings that their attitudes were influenced by their own misconceptions about tracking are in line with the observations seen in Section 4.3.2 about how limited our participant's understanding of third-party tracking organizations and their methods of surveillance could be.

Other participants were not worried about surveillance and subsequently did not expect major changes to their online information seeking behavior after the intervention, not because they trusted the tracking parties, but because they have **belief in the effectiveness of their privacy protection strategies**. We found that the participants who had this perspective, most of whom reported in the questionnaire to have high levels of digital literacy, employed various types of strategies. Some reported that they used specific tools to aid

their private searching session—"*Oh, I mean, I'm already using a meta search engine, which is a search engine proxy of Google. So I think it already cuts off a lot of trackers and my information. Because they don't directly communicate this, there's a layer [preventing this]*" (P_7). Others expressed that they were not too worried about tracking since they would not provide sensitive information online at all and other fields of information cannot be misused in a way that is as harmful as sensitive ones—"*At this point, if it's just browsing, I'm fine with it, unless it asks me to show my face, show some very concrete image of myself or get some very concrete information about myself*" (P_{17}). By implementing these strategies and seeing no negative consequences in their lives, participants were thus convinced that their current methods for seeking health information is effective enough in protecting their privacy and therefore is not in need of modification: "*So far I don't see it [changing]. It has not impacted me or come back to me or I did not have to address anything*" (P_{15}). These sentiments are understandable, as Internet users have been shown to have different privacy expectations for different types of information—users expect higher levels of privacy and safeguard for sensitive user data [47], and they would only be concerned and uncomfortable if such user data is tracked by websites [68].

However, we noticed that the strategies mentioned here by these participants are the same strategies mentioned in Section 4.3.3 that were used to seek specifically sensitive health information, i.e., the use of private browsing mode and active non-disclosure of sensitive data online by the users. In fact, the same P_4, whom we quoted earlier about their decision to not search sensitive information online at all, cited the same strategy as the reason why they did not feel the need for taking on more protection strategies: "*But I don't do any form of ultimately very sensitive browsing, I don't do anything really besides going onto various news sites, social media and work-related things, and entertainment. So none of those have really merited me paying for a VPN*". However, tools such as private search engines mentioned and used by P_7 could only protect them during the initial search but not during the browsing session on the portals themselves, where numerous pieces of tracking technologies are installed as pointed out in Chapter 3 and prior work [28, 56]. Similarly, the strategy of actively not disclosing sensitive information online through queries and inputs only has a minimal effect when it comes to anonymizing user identity, as portals can still infer such information from the huge amount of data across different sources [103], the same types of

data that participants like P_{17} viewed as: "*in general, this kind [of data] is harmless*".

This observation that some of our participants are falsely convinced that their protection strategies are strong enough for them not to worry about tracking is not new, however. As Section 4.3.2 showed, most of our participants are under-informed when it comes to the surveillance practices on health portals. More specifically, we saw that they were not aware of what tracking technologies are being used on portals and how. They also seemingly did not know about the existence of protection strategies other than the common methods of VPNs, ad blocker and private browsing modes. Therefore, it makes sense that they potentially overestimated the effectiveness of their current privacy protection strategies. What is concerning, however, is how this erroneous confidence in their protection strategies remained intact even after being shown the correct facts about surveillance and specifically how tracking is conducted on portals. Story et al. also observed the same kind of overconfidence in their participant's inaccurate knowledge of protection tools and their effectiveness [96], which showed how such attitude could expose them to major privacy threats.

Not all participants were this indifferent about surveillance, as some participants did express a certain degree of concerns over how extensive tracking was on health portals. However, these participants then stated that they **accepted tracking as a normal part of their browsing experience, albeit reluctantly**. For some participants, they attributed this viewpoint to their belief that there is nothing they can do to stop the extensive online surveillance that is happening on health portals. We saw this mindset in P_{10}, who said: "*There's very little that I can actually control. Yes, I can deny cookies, but there's so much that's being tracked that it's kind of almost pointless. Because they're tracking, they're either tracking it from a different source or they're tracking everything else*". Similarly, P_3 explained: "*I think that a good amount of people, I put myself kind of in that camp, are having a cynicism around it, where it's almost a fight not worth fighting at an individual level, in terms of not wanting to bother to do all the steps to hide my own usage because everyone's doing it anyway, and they're gonna still get all that info*". This cynical attitude towards privacy can lead to a refusal to take on stronger tracking protection strategies in some cases, as shown by Hoffmann et al. [42]. We saw this kind of behavior with our participants as well, like P_{10}: "*I don't use [VPNs], mostly because they cost something, and my ratio of caring to wanting to pay is [not enough]*". These observations

are consistent with the findings of Shklovski et al., who found that their participants also felt uncomfortable about tracking yet did not expect to make modifications to their phone usage habits [89]. In his participant's words, "*It's unsettling and not ok, but I feel very powerless against it*".

One possible explanation for why these participants had such a pessimistic view of the state of privacy is due to their limited information about how online tracking works, as outlined in Section 4.3.2. It is their lack of awareness, particularly about the technologies used for tracking by portals currently and its mechanisms, that further solidified this mindset that privacy can no longer be maintained due to how ubiquitous tracking is. For instance, P_6, who early on in the interview could not give a correct and specific answer when asked if they knew how surveillance is conducted online, later commented: "*I've heard [that] they're even able to surpass that [level of protection], even if you try to hide your location, they can still triangulate your location back if they wanted to. So no, [...] even if I try, they're always one step ahead*". Such limited understanding of tracking technology, combined with the basic, surface-level familiarity with protection strategies, could explain why these participants were so convinced that there is no meaningful way to protect their privacy in this day and age anymore— a common myth that has been debunked by researchers [41]. This finding thus supported the argument from Trepte et al. that people refuse to take measures to protect their online privacy due to their lack of privacy literacy [99].

We also observed that with certain participants, their acceptance and normalization of surveillance came from their need for the health information that these portals provided. Because of this, they were willing to make the trade off in order to get access to such content: "*My thoughts are that I don't like [tracking] but I'm not sure how I could stop it. I mean my desire to get this medical information overrides those concerns*" (P_1), "*I look at [these health] websites in the same way as [I do with] contracts: I'm agreeing to something without knowing it, in exchange for a service that I really want. If I really want to get services, in exchange, they're following me and collecting data*" (P_{19}). This preference for health content is consistent with what we have previously observed in Section 4.3.1 about how many participants, including P_1 and P_{19}, prioritized content-related factors, such as its quality or the level of detail, over a portal's privacy and surveillance practices to help them choose from a wide selection of health information portals from search engines. We also saw a similar pattern of prioritizing

content over privacy in the questionnaire responses from the participants who had this mindset: despite stating that they were quite concerned about the protection of their privacy, the majority of them reported that they had never stopped doing business with a company that experienced privacy breaches. When we asked our participants to further elaborate on this significant demand for online health information of theirs, we discovered that this major need for health information stems from the barriers they face when trying to access healthcare services in real life, similar to findings from previous work [16, 21]. As P_{12} explained, "*At the end of the day, you are tired and you really want to look for information that is really critical. Like, you want to know what kind of medication you need to take or to relieve the pain, or if you know any family member that needs help, you just look for information, right? Especially now that it's really hard in [their province], unfortunately, to reach out to a doctor or to a nurse*".

In closing, we note a potential relation with age groups. More specifically, we observed that participants who expressed that they had come to sympathize with tracking because they were in great need of health information tended to be older adults, while those more cynical about tracking and believed that they were powerless against surveillance were mostly younger participants under the age of 35. This difference is understandable and in fact aligns well with prior research: Pourrazavi et al. have shown that older adults relied significantly more on online health information due to the prevalence of medical problems [81], while younger adults were found to be more apathetic and skeptical about privacy due to their early exposure to digital technologies as well as privacy issues [39, 51].

5 Discussion

In this chapter, we discuss the results we have obtained in previous chapters. More specifically, in Section 5.1, we talk about the results from our analysis of surveillance on health portals in Chapter 3, along with the limitations of our analysis and future directions for this line of research. In Section 5.2, we discuss the observations from our interview study and their implications in Chapter 5.2, the limitations of our research and recommendations for future work on this topic.

5.1 Surveillance on Health Information Portals

As discussed in Section 3.4, we observed that all commercial health portals contained an extensive amount of tracking, especially Mayo Clinic which contained the invasive tracking technology of session recording [27]. This mirrors the findings of Burkell and Fortier, who similarly found that commercial health websites have significantly more ad trackers than governmental ones [19].

When we examined the amount of health content available on these portals, we found that these commercial portals provided significantly more content than their non-commercial counterparts. Using the data from Table 3.3, we discovered that a commercial portal has on average 1251 articles on health topics, twice as many as the average number of articles found on a non-commercial size, which is 586 articles. The amount of content in these articles is also larger on commercial portals than non-commercial ones, with commercial

portals having on average 1468 words and non-commercial portals only having 553 words. This is understandable, as commercial portals would focus more on supplying consumers with as much health information as possible to maintain a high level of customer retention and traffic monetization. Governmental portals, on the other hand, would be expected to concentrate on providing their citizens up-to-date health information to assist with public health. This means that if information seekers want to access more detailed information on health topics, they have to subject themselves to a higher amount of tracking, as the portals with a higher amount of health information are the ones with a higher level of visitor surveillance.

Commercial portals are not the only ones with alarming privacy issues, however. Although most non-commercial portals were found to contain a minimal or preventable degree of tracking, which is consistent with findings from previous work [19], we noticed that five specific governmental portals contained a level of tracking that is comparable to that from their commercial counterparts, namely the portals from CDC, Manitoba, Québec, Saskatchewan, and WHO. Furthermore, we also observed that despite non-commercial portals having lower levels of tracking, all but two of them (NHS and Ontario) contained either cookies or ad trackers from major data collectors—companies that have tracking technologies installed across a significant number of websites. Such presence of major data collectors is highly concerning, as prior work has shown how collectors like Google can aggregate the data that they have collected from the numerous sources they appear on to create more accurate and complete visitor profiles for marketing purposes [34, 60]. These observations thus suggested how visiting non-commercial portals, such as those from governmental agencies, instead of their commercial equivalents does not necessarily provide users with higher levels of privacy, in contrast to what some, including our participants (Section 4.3.4), might believe.

With the increase in the number of privacy protection laws passed by various governments around the world over the past few decades [36], many regulations now require websites to fully disclose their surveillance practices to the users by including details about tracking operations and purposes in privacy policies. Thus, we were interested in seeing if the chosen health portals were transparent about the significant amount of visitor monitoring we have observed in our analysis. After collecting privacy policies from these

portals and thoroughly examining them, we found that the majority of health portals did not fully disclose the extent of their surveillance practices—out of 22 portals, only three were transparent about who conducts tracking on their pages and how. This resembles the result of Libert's work, where they also discovered from auditing 200,000 privacy policies that websites failed to disclose information about 85% of the third-party organizations that were found to be actively collecting user data on these portals [58].

In addition, several policies contained potentially misleading information about their tracking practices. For example, the privacy policy of Québec's health portal claims that the collected user information on their site by Google "cannot be linked to an individual", and that "Google will never link the information collected with any other data or information that it preserves". However, in the same policy, they also list the wide range of user information collected on their site, such as IP address, system information, and previously visited websites. Although these fields seem to be vague, untraceable at first sight, research has shown how third-party analytics services, specifically Google, can aggregate these fields of data and infer information about various aspects of users, such as their interests and general online activities, which can effectively be used to identify visitors [103]. These findings thus showed that privacy policies from health portals are lacking when it comes to providing elaborate information about their surveillance practices, essentially preventing information seekers from having an accurate and complete understanding of the user tracking and its consequences on their privacy that happens during their portal visits.

We also noticed, however, that there is a major difference in the length of the policy between commercial and non-commercial health portals: on average, policies from non-commercial portals are significantly shorter and more concise than those from their commercial counterparts. Upon further investigation, we discovered that the privacy policies from all six commercial portals were detailed and extensive when talking about the tracking mechanisms as well as third-party organizations present on the portal for analytics purposes, even going as far as citing the privacy policies from the analytics service providers themselves. Their non-commercial counterparts were notably briefer, with some policies containing virtually no information on visitor monitoring technologies or third-party tracking organizations present on their sites, such as the policies from Alberta and Manitoba.

A possible explanation for this is the business nature of commercial portal's organizations. While non-commercial portals' main aim is to provide residents with health information, commercial portals are hosted by private organizations who have a financial stake in the sites. Therefore, visitors to these health portals are essentially consumers of these portals, meaning that there are more legal implications for these organization and their relationships with the consumers, such as consumer protection laws. Moreover, in many cases like Canada's PIPEDA [76], privacy laws have a larger focus on regulating data collection and usage from commercial entities rather than governmental ones, with more restrictions and requirements in these cases in order to protect consumers. Therefore, it is understandable that the policies from commercial portals tended to be longer and more detailed, as these private organizations needed to be more exact and precise for legal purposes.

The results from our analysis of surveillance on health portals thus revealed the extensive and invasive nature of surveillance on both commercial and non-commercial health portals, as well as the portals' failure to fully disclose their surveillance practices to visitors via privacy policies. These findings are greatly concerning, especially considering the increasing dependency of Internet users on online resources for health information [44, 81]. Such demand for health content could potentially lead information seekers to prioritize the obtaining of health information over their privacy concerns. We observed this preference for content in real life with our user study, where many participants similarly mentioned that they would disregard their worries about tracking when they visit health portals due to their pressing need for health information (see Section 4.3.4). When doing so, however, information seekers are unknowingly opening themselves up to significant privacy intrusions that would have major consequences on their lives.

Limitations and Future Work

We conducted our study of surveillance on health portals by analyzing data on health portals that was collected from April to August 2024. This is to a certain extent inevitable, as our research involves manual analysis. However, this lag introduced the potential for discrepancies between the data we report and the current state of surveillance on the portals we studied. Therefore, a possible future direction for this work is to replicate our analysis at a later point in time and use these results as a baseline for comparison, providing a

longitudinal overview of the evolution of user tracking on health portals.

Another limitation is that we did not consider privacy policies from web analytics service providers like Google and Microsoft themselves, especially when they were directly mentioned by some health portals in their respective policies. These policies might have provided some more insights into the level of transparency tracking organizations have about the disclosure of their tracking operations. However, this is less important for our analysis, as we were more interested in learning if the health portals included complete details about the tracking that takes place on their pages, as it is the policy directly from the health portal where typical information seekers would go to when they want to learn more about this topic and not the third-party organizations' policy. This limitation does show the potential for the inclusion of these third-party organization's policies in future research on this topic. This process of analyzing service provider's policies can be automated further to obtain faster results, similar to the approach of Libert [58].

As for our model of the different levels of surveillance on health information portals (Section 3.3), we based our model on the categorization of tracking technologies and their corresponding list of available defense strategy provided by Bujlow et al. [18], which does not cover any new tracking technology or defense strategy that might have been developed in the eight years since the publication of the paper. Our decision to use this taxonomy from Bujlow et al. is because to the best of our knowledge, this work is the most recent comprehensive taxonomy of web tracking technologies, making it the best option for a detailed classification of web tracking technologies as well as defense strategies. Moreover, as we classified tracking technologies based on their data storage method—stateless (no need for storage on device) or stateful (storage on device needed)—which is consistent with the classification of Mayer et al. [67], we made it possible for the inclusion of any new tracking technology, since any technology must operate in one of these two modes. Similarly, if there is a new defense strategy for a tracking technology previously classified as "indefensible", we can alter their categorization to reflect the user's newly established defensibility against said technology.

Since we relied on the Blacklight tool for our analysis of surveillance on portals, our results are also impacted by the limitations of the tool itself. The developers of Blacklight have acknowledged several limitations, such as the heuristic nature of their tracking detec-

tion methods as well as the potential differences between surveillance responses by portals to simulated visitor behavior and actual visitor behavior [66]. As they have stated, "For this reason, Blacklight results should not be taken as the final word on potential privacy violations by a given website. Rather, they should be treated as an initial automated inspection that requires further investigation before a definitive claim can be made."

Since our analysis and modelling is tailored specifically to the 22 chosen health portals, we recognize that there is a threat to our external validity, meaning that our aggregated observations, such as averages, along with the model of surveillance levels are not generalizable to other health portals. However, this lack of generalizability does not significantly affect our findings, as our goal was never to analyze the levels of surveillance on all health portals, but to investigate surveillance specifically on portals that are relevant to Canadian residents. Thus, the implications of the results of our analysis are only relevant to Canadian residents, and the model of surveillance levels are therefore also only applicable to this sample of 22 health portals applicable to information seekers who live in Canada. Moreover, our methodology was applicable to any website and not restricted to only these portals: aside from Blacklight, which is also widely available for any site, we manually investigated tracker ownership as well as collected and analyzed privacy policies. Therefore, any future project can easily extend this work by applying our research methods to other health portals, such as portals from specific countries or international organizations.

5.2 Privacy Considerations when Seeking Health Information

Overall, our results showed that factors relating to a portal's content and source informed our participant's decision making process much more than privacy concerns when it comes to seeking and browsing health information online. However, when it came to seeking sensitive health information about stigmatized and serious conditions, some employed specific strategies to protect the privacy of such searching sessions. We also found that our participants had a limited understanding of how surveillance works with many misconceptions about tracking mechanisms and practices as well as protection strategies, which was apparent in the way the employed strategies for seeking sensitive information failed to

provide them with the desired level of privacy. We observed that although surprised, our participants were not overly worried about such tracking on portals and did not express any intentions of modifying their information seeking behavior. When examining the explanations participants gave for such perceptions, we observed that there are two main factors that have been a common theme throughout the interviews and have guided participant's attitude and behavior when seeking health information significantly: their need for health content, and their lower levels of privacy literacy.

Firstly, we discovered that the health content on these portals plays quite a role in how people interact with said portals. We saw this impact in the way participants prioritized characteristics of content on a site, such as its accuracy and level of details, over its privacy safety as a selection criteria to help them determine which information portal to visit. We also saw this impact in how the need for health content have led to participants forfeiting their privacy and unwillingly accepting online surveillance as part of their browsing experience in exchange for access to medical information. The question here, however, is why do our participants prioritize health information so much to the point of willingly enabling privacy violations? We hypothesize that this is because not only is online health information very useful in enhancing their real-life doctor visits [30], such information also helps mitigate an increasing number of barriers to healthcare services [16, 21], especially when healthcare inaccessibility has significantly worsened over the years in Canada [6, 104] as well as other countries [54, 79]. However, it should not be the case that users have to exchange their privacy for vital health information—information seekers should be entitled to both. Therefore, it is of utmost importance that involved parties, such as governments, focus on both improving the state of their healthcare systems as well as providing secure platforms that offer health information, so that information seekers can receive essential medical information without knowingly endangering their own privacy.

Secondly, we surmise that knowledge of surveillance practices plays an important role in how users think about privacy. More specifically for our case, we saw how participants had many misconceptions about surveillance on health portals and how to effectively protect their privacy. However, without any help to rectify these misunderstandings overtime, our participants appear to have developed either a false sense of trust in the privacy of their information seeking practices, or the incorrect belief that there is nothing that can be done

to stop tracking. In either cases, these attitudes may be related to users' nonchalance about tracking, which explained why participants expressed no intentions of changing how they seek health information, even after we showed them how extensive surveillance really is on health portals and the consequences of it. With this finding, we further support the idea that privacy literacy can influence how people behave and interact online, an idea that prior work has established in several different research projects [31, 62]. These results reveal how dangerous lower degrees of privacy literacy can be, as they can make users unknowingly perform actions that would expose them to substantial privacy threats while still believing that they are fully protected from such intrusions. Therefore, it is essential that there should be ways for Internet users to access factual, up-to-date information about online privacy in order for them to increase their privacy literacy and thus engage in safe information seeking activities, especially when it comes to such sensitive and vital information like health.

Limitations and Future Work

Our observations might be affected by some limitations related to the study design. One limitation is our assessment of participant's digital literacy via the questionnaire. As this served as a basis of their knowledge on privacy and surveillance pre-interview, it would be beneficial to use an established, multidimensional scale for a concept as multifaceted as privacy literacy, such as OPLIS [99], rather than relying on a single question based on a person's self-reporting of their awareness of privacy matters. By employing scales like OPLIS, we could then have a verified and objective assessment of their privacy literacy than we could then use as a basis to compare with the statements made by the participants when asking about their awareness of surveillance practices in Section 4.3.2. However, scales such as OPLIS are long and contain a large number of questions in order to capture the multidimensional nature of privacy literacy. Therefore, we intentionally made this trade-off by not applying such a long instrument in our recruitment process, as lengthy questionnaires would increase the cost of participation and can be a deterrent to joining the interview for potential participants.

A threat to our results' internal validity is that we did not include information about privacy protection strategies in our intervention, or more specifically, information on which

strategies are available and effective at defending against the presented tracking technologies. Our participants were unaware of helpful protection strategies outside of the mentioned strategies of VPNs, ad blockers, private browsing mode, and active withholding of data. Consequently, some participants may not have expressed any interest in improving the privacy of their online interactions by taking on better protection strategies, as they were not aware of how to change their behavior to incorporate privacy protection. Another intervention-related issue is also the generic nature of the presented information. Despite talking in detail about the mechanisms and consequences of surveillance, we acknowledge that the provided information was general and did not tailor to each participant's personal circumstances. This might have impacted how they perceived our presentation and might be associated with their underestimation of how personally disruptive such privacy intrusions can be. Both of these shortcomings were addressed by previous work like that from Malandrino et al. as their intervention was in the format of an interactive tool that provide details about tracking customized to each participant and their browsing session [62], giving us a possible explanation as to why their participants exhibited a higher degree of willingness to adopt stronger privacy protection strategies. However, we intentionally made the decision not to include such information because this is a trade-off to keep the interview focused and tractable.

As for the analysis methodology, we employed the manual annotation method of descriptive coding, performed by the main author. This can lead to biases and create subjective interpretations of participant's expressions. To minimize this bias, we constructed a code book with well-defined guidelines on how and when to assign specific codes, which was validated by a team of three collaborators, as discussed in Section 4.2. For each of these disagreements, we accepted all codes given by the three annotators and essentially viewed it as the exceptional cases where the quote was assigned more than one code.

Since we only interviewed a small population of 20 Canadian residents, we recognized that our findings are not statistically generalizable to the broader population of Canadian health seekers. However, this threat to external validity is more applicable to quantitative studies than to qualitative studies like ours, as the goal of our research is to synthesize the different perspectives and attitudes of our diverse set of participants into a series of rich descriptions and conclusions about how Canadian health information seekers consider

privacy, rather than making generalizations about the interaction patterns of a sample to a wider population.

6

Conclusion

With this thesis, we set out to learn more about surveillance on health portals in Canada as well as how website visitors consider privacy during their health information seeking process. Based on an analysis of the tracking technology deployed on 22 health portals and conducting interviews with 20 Internet users, we make observations about the state of privacy intrusion on health portals and users' perceptions of it. First, we saw that all commercial portals that Canadian residents could be likely to rely on for health information deploy a significant amount of tracking. Similarly, many non-commercial portals, such as those from provincial governments, also contain an extensive amount of tracking on their sites, albeit to a lesser extent compared to their commercial counterparts. This tracking is done through the use of invasive technologies owned by a large number of organizations, mainly advertisement providers and behavior analytics services. Despite this extensive degree of surveillance, the majority of these portals were not transparent about their visitor monitoring practices and included few details about third-party tracking in their privacy policies.

We learned from the interviews that participants used different criteria to select a health information portal from the diverse resource of health websites, focusing more on factors about the content as well as the source of the portals, rather than their privacy. This disregard for privacy is potentially related to their limited knowledge on surveillance practices on health portals, only understanding tracking organizations and technologies on a surface level. These lower levels of privacy literacy could be associated with misconceptions about

privacy and effective protection against it, which help explain why participants employ strategies that are not fully effective at protecting their privacy. However, when we corrected their perception of surveillance by providing accurate information about tracking mechanisms and its implications, we observed that participants were still not overly concerned about tracking and expressed that they did not expect to make any changes to how they currently seek health information. We hypothesize that this attitude could be related to their faith in the tracking organizations and their lack of malicious intentions, their confidence in how effective and secure their current information seeking strategies are, or their reluctant acceptance of tracking as a regular part of seeking health information.

As implications of these findings, we first recommend that there should be a way for Internet users to have access to the latest, most accurate privacy information. Such information includes details on the privacy implications of visiting each site—what tracking technologies are used on the site, what types of data are being collected, what could be the consequences of access to such data—as well as information about the possible privacy protection strategies against each of these tracking methods. One way of supplying privacy information could be through an interactive tool that can be installed on each user's device. Users can use such a tool to easily access privacy information, helping them make more informed decisions on how to interact privately with health information portals while still allowing them to get the needed medical content. Moreover, as suggested by the protection motivation theory for privacy [71, 83], actively supplying users with details about the risks of their online activities through these interactive tools can help users have a better understanding of threats and become more willing to adopt stronger privacy protection strategies.

Secondly, we also need to make sure that these interactive tools for privacy information are widely available and accessible, as the wide tool availability could lead to the improvement of knowledge about privacy protection strategies for a larger population, who in turn could potentially inform others in their social circle and thus spread awareness to even more people. Social triggers, such as recommendations from friends and family as well as discussion about privacy topics with acquaintances, have been shown to significantly influence people's decisions to change their online behavior regarding privacy and security [23].

It is also worth mentioning that efforts to improve the public's awareness of privacy issues and encourage people to enhance their online behavior should not only rely on the users themselves. Governments and other regulating bodies can be involved in the process of mitigating privacy intrusions. These parties should continually invest in improving existing regulations in order to make sure they are in the interests of the users so that they do not have to make trade-offs in order to access products and services, while still ensuring the up-to-dateness of these laws, considering the current growth rate of invasive tracking technologies.

Finally, this thesis illustrates a need for governments to address the problem of extensive user tracking on health portals, either by providing its residents with a source of health information that is free of surveillance, or investing in the improvement of the safety and privacy on existing portals. After all, it is high time that information seekers be able to access something as vital and sensitive as health information without having to sacrifice their own privacy.

Contributions

The author is responsible for the entirety of this thesis, working under the supervision and guidance of Professor Martin Robillard. For the user study in Chapter 4, this project was a collaboration between the author, their supervisor—Professor Robillard, Professor Jin Guo, and a PhD student, Deeksha Arya. In particular, Professor Robillard and Professor Guo collaborated with the author in the validation of the code book, while D. Arya worked together with the author for the design of the pre-interview questionnaire as well as the interview and analysis protocol. All three collaborators also participated in the construction of the code book by giving feedback for each iteration of the book, and later provided significant feedback for the reporting of the interview analysis results.

Bibliography

[1] The Markup. Wikipedia, Verified 2023-10-26. https://en.wikipedia.org/wiki/The_Markup.

[2] ACAR, G., ENGLEHARDT, S., AND NARAYANAN, A. No boundaries: data exfiltration by third parties embedded on web pages. *Proc. Priv. Enhancing Technol. 2020*, 4 (2020), 220–238.

[3] ACQUISTI, A., FRIEDMAN, A., AND TELANG, R. Is there a cost to privacy breaches? an event study. *ICIS 2006 proceedings* (2006), 94.

[4] ACQUISTI, A., AND GROSSKLAGS, J. Privacy and rationality in individual decision making. *IEEE Secur. Priv. 3*, 1 (2005), 26–33.

[5] ADBLOCK PLUS. Adblock plus—surf the web without annoying ads! https://adblockplus.org, 2014.

[6] AHMED, S., SHOMMU, N. S., RUMANA, N., BARRON, G. R., WICKLUM, S., AND TURIN, T. C. Barriers to access of primary healthcare by immigrant populations in canada: a literature review. *Journal of immigrant and minority health 18* (2016), 1522–1540.

[7] AL SHBOUL, M. K. I., ALWREIKAT, A., AND ALOTAIBI, F. A. Investigating the use of chatgpt as a novel method for seeking health information: a qualitative approach. *Science & technology libraries 43*, 3 (2024), 225–234.

[8] ALFNES, F., AND WASENDEN, O. C. Your privacy for a discount? exploring the willingness to share personal data for personalized offers. *Telecommunications Policy 46*, 7 (2022), 102308.

[9] ALI, S., ISLAM, N., RAUF, A., DIN, I. U., GUIZANI, M., AND RODRIGUES, J. J. P. C. Privacy and security issues in online social networks. *Future Internet 10*, 12 (2018), 114.

[10] ANDALIBI, N., ÖZTÜRK, P., AND FORTE, A. Sensitive self-disclosures, responses, and social support on instagram: The case of #depression. In *CSCW* (2017), ACM, pp. 1485–1500.

[11] ARBOLEDA-FLÓREZ, J., AND STUART, H. From sin to science: fighting the stigmatization of mental illnesses. *The Canadian Journal of Psychiatry 57*, 8 (2012), 457–463.

[12] AUGUSTAITIS, L., MERRILL, L. A., GAMAREL, K. E., AND HAIMSON, O. L. Online transgender health information seeking: Facilitators, barriers, and future directions. In *CHI* (2021), ACM, pp. 205:1–205:14.

[13] BARTSCH, M., AND DIENLIN, T. Control your facebook: An analysis of online privacy literacy. *Comput. Hum. Behav. 56* (2016), 147–154.

[14] BÉLANGER, F., AND CROSSLER, R. E. Privacy in the digital age: a review of information privacy research in information systems. *MIS quarterly* (2011), 1017–1041.

[15] BERGSTRÖM, A. Online privacy concerns: A broad approach to understanding the concerns of different groups for different uses. *Comput. Hum. Behav. 53* (2015), 419–426.

[16] BHANDARI, N., SHI, Y., AND JUNG, K. Seeking health information online: does limited healthcare access matter? *J. Am. Medical Informatics Assoc. 21*, 6 (2014), 1113–1117.

[17] BIELOVA, N. Web tracking technologies and protection mechanisms. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (2017), pp. 2607–2609.

[18] BUJLOW, T., CARELA-ESPAÑOL, V., SOLE-PARETA, J., AND BARLET-ROS, P. A survey on web tracking: Mechanisms, implications, and defenses. *Proceedings of the IEEE 105*, 8 (2017), 1476–1510.

[19] BURKELL, J., AND FORTIER, A. Could we do better? behavioural tracking on recommended consumer health websites. *Health Information & Libraries Journal 32*, 3 (2015), 182–194.

[20] CHOI, H., PARK, J., AND JUNG, Y. The role of privacy fatigue in online privacy behavior. *Comput. Hum. Behav. 81* (2018), 42–51.

[21] CHU, J. T., WANG, M. P., SHEN, C., VISWANATH, K., LAM, T. H., AND CHAN, S. S. C. How, when and why people seek health information online: qualitative study in hong kong. *Interactive journal of medical research 6*, 2 (2017), e7000.

[22] CUSTERS, B. Data dilemmas in the information society: Introduction and overview. In *Discrimination and privacy in the information society: Data mining and profiling in large databases*. Springer, 2013, pp. 3–26.

[23] DAS, S., KIM, T. H., DABBISH, L. A., AND HONG, J. I. The effect of social influence on security sensitivity. In *SOUPS* (2014), USENIX Association, pp. 143–157.

[24] DE CHOUDHURY, M., MORRIS, M. R., AND WHITE, R. W. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the Conference on Human Factors in Computing Systems* (2014), p. 1365–1376.

[25] DINEV, T., AND HART, P. J. An extended privacy calculus model for e-commerce transactions. *Inf. Syst. Res. 17*, 1 (2006), 61–80.

[26] ELECTRONIC FRONTIER FOUNDATION. Privacy badger. https://privacybadger.org/##How-does-Privacy-Badger-work, Accessed 2024.

[27] ENGLEHARDT, S., ACAR, G., AND NARAYANAN, A. No boundaries: Exfiltration of personal data by session-replay scripts. *Freedom to Tinker 15* (2017).

[28] ENGLEHARDT, S., AND NARAYANAN, A. Online tracking: A 1-million-site measurement and analysis. In *CCS* (2016), ACM, pp. 1388–1401.

[29] EUROSTAT. One in two eu citizens look for health information online. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20210406-1. Accessed 2025.

[30] FIKSDAL, A. S., KUMBAMU, A., JADHAV, A. S., COCOS, C., NELSEN, L. A., PATHAK, J., AND MCCORMICK, J. B. Evaluating the process of online health information searching: a qualitative approach to exploring consumer perspectives. *Journal of medical Internet research 16*, 10 (2014), e224.

[31] GERBER, N., GERBER, P., DREWS, H., KIRCHNER, E., SCHLEGEL, N., SCHMIDT, T., AND SCHOLZ, L. Foxit: enhancing mobile users' privacy behavior by increasing knowledge and awareness. In *STAST* (2017), ACM, pp. 53–63.

[32] GERBER, N., GERBER, P., AND VOLKAMER, M. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Comput. Secur. 77* (2018), 226–261.

[33] GHARIB, M. Privacy and informational self-determination through informed consent: The way forward. In *CyberICPS/SECPRE/ADIoT/SPOSE/CPS4CIP/CDT&SECOMANE@ESORICS* (2021), vol. 13106 of *Lecture Notes in Computer Science*, Springer, pp. 171–184.

[34] GILL, P., ERRAMILLI, V., CHAINTREAU, A., KRISHNAMURTHY, B., PAPAGIANNAKI, K., AND RODRIGUEZ, P. Follow the money: understanding economics of online aggregation and advertising. In *Internet Measurement Conference* (2013), ACM, pp. 141–148.

[35] GOVERNMENT OF CANADA. Canada's health care system, Apr. Verified 2023-08-11. https://www.canada.ca/en/health-canada/services/canada-health-care-system.html.

[36] GREENLEAF, G. Global data privacy laws 2017: 120 national data privacy laws, including indonesia and turkey. *Including Indonesia and Turkey (January 30, 2017) 145* (2017), 10–13.

[37] HANNAK, A., SOELLER, G., LAZER, D., MISLOVE, A., AND WILSON, C. Measuring price discrimination and steering on e-commerce web sites. In *Internet Measurement Conference* (2014), ACM, pp. 305–318.

[38] HARBORTH, D., AND PAPE, S. How privacy concerns, trust and risk beliefs, and privacy literacy influence users' intentions to use privacy-enhancing technologies: The case of tor. *Data Base 51*, 1 (2020), 51–69.

[39] HARGITTAI, E., AND MARWICK, A. "what can i really do?" explaining the privacy paradox with online apathy. *International Journal of Communication 10* (2016), 21.

[40] HERBERT, F., BECKER, S., SCHAEWITZ, L., HIELSCHER, J., KOWALEWSKI, M., SASSE, M. A., ACAR, Y., AND DÜRMUTH, M. A world full of privacy and security (mis)conceptions? findings of a representative survey in 12 countries. In *CHI* (2023), ACM, pp. 582:1–582:23.

[41] HOEPMAN, J. H. Privacy is hard and seven other myths. *European Data Protection Law Review 9*, 2 (2023), 104–111.

[42] HOFFMANN, C. P., LUTZ, C., AND RANZINI, G. Privacy cynicism: A new approach to the privacy paradox.

[43] HONG, W., AND THONG, J. Y. L. Internet privacy concerns: An integrated conceptualization and four empirical studies. *MIS Q. 37*, 1 (2013), 275–298.

[44] JACOBS, W., AMUTA, A. O., AND JEON, K. C. Health information seeking in the digital age: An analysis of health information seeking behavior among us adults. *Cogent social sciences 3*, 1 (2017), 1302785.

[45] JIA, X., PANG, Y., AND LIU, L. S. Online health information seeking behavior: A systematic review. *Healthcare (Basel, Switzerland) 9*, 12 (2021).

[46] KAMMERER, Y., AND GERJETS, P. The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *International Journal of Human-Computer Interaction 30*, 3 (2014), 177–191.

[47] KANG, R., DABBISH, L., FRUCHTER, N., AND KIESLER, S. "my data just goes everywhere": user mental models of the internet and implications for privacy and security. In *Proceedings of the Eleventh USENIX Conference on Usable Privacy and Security* (July 2015), pp. 39–52.

[48] KARAJ, A., MACBETH, S., BERSON, R., AND PUJOL, J. M. Whotracks. me: Shedding light on the opaque world of online tracking. *arXiv preprint arXiv:1804.08959* (2018).

[49] KNOWLES, B., AND CONCHIE, S. Un-paradoxing privacy: Considering hopeful trust. *ACM Trans. Comput. Hum. Interact. 30*, 6 (2023), 87:1–87:24.

[50] KOKOLAKIS, S. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Comput. Secur. 64* (2017), 122–134.

[51] KURKOVSKY, S., AND SYTA, E. Digital natives and mobile phones: A survey of practices and attitudes about privacy and security. In *ISTAS* (2010), IEEE, pp. 441–449.

[52] LAVALLEY, S. A., KIVINIEMI, M. T., AND GAGE-BOUCHARD, E. A. Where people look for online health information. *Health Information & Libraries Journal 34*, 2 (2017), 146–155.

[53] LAZAR, J., FENG, J. H., AND HOCHHEISER, H. Chapter 8 - interviews and focus groups. In *Research Methods in Human Computer Interaction (Second Edition)*. 2017.

[54] LAZAR, M., AND DAVENPORT, L. Barriers to health care access for low income families: a review of literature. *Journal of community health nursing 35*, 1 (2018), 28–37.

[55] LEWANDOWSKI, D., AND KAMMERER, Y. Factors influencing viewing behaviour on search engine results pages: a review of eye-tracking research. *Behaviour & Information Technology 40*, 14 (2021), 1485–1515.

[56] LIBERT, T. Exposing the invisible web: An analysis of third-party http requests on 1 million websites. *International Journal of Communication 9* (2015), 18.

[57] LIBERT, T. Privacy implications of health information seeking on the web. *Commun. ACM 58*, 3 (2015), 68–77.

[58] LIBERT, T. An automated approach to auditing disclosure of third-party data collection in website privacy policies. In *WWW* (2018), ACM, pp. 207–216.

[59] LIN, S., CHOU, K., CHEN, Y., HSIAO, H., CASSEL, D., BAUER, L., AND JIA, L. Investigating advertisers' domain-changing behaviors and their impacts on adblocker filter lists. In *WWW* (2022), ACM, pp. 576–587.

[60] MALANDRINO, D., PETTA, A., SCARANO, V., SERRA, L., SPINELLI, R., AND KRISHNAMURTHY, B. Privacy awareness about information leakage: who knows what about me? In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society* (Nov. 2013), pp. 279–284.

[61] MALANDRINO, D., AND SCARANO, V. Supportive, comprehensive and improved privacy protection for web browsing. In *SocialCom/PASSAT* (2011), IEEE Computer Society, pp. 1173–1176.

[62] MALANDRINO, D., SCARANO, V., AND SPINELLI, R. How increased awareness can impact attitudes and behaviors toward online privacy protection. In *Proceedings of the International Conference on Social Computing* (2013), p. 57–62.

[63] MANGONO, T., SMITTENAAR, P., CAPLAN, Y., HUANG, V. S., SUTERMASTER, S., KEMP, H., AND SGAIER, S. K. Information-seeking patterns during the covid-19 pandemic across the united states: Longitudinal analysis of google trends data. *Journal of Medical Internet Research 23*, 5 (2021), e22933.

[64] MAON, S. N., HASSAN, N. M., AND SEMAN, S. A. A. Online health information seeking behavior pattern. *Advanced Science Letters 23*, 11 (2017), 10582–10585.

[65] MASUR, P. K. How online privacy literacy supports self-data protection and self-determination in the age of information. *Media and Communication 8*, 2 (2020), 258–269.

[66] MATTU, S., AND SANKIN, A. How we built a real-time privacy inspector. https://themarkup.org/blacklight/2020/09/22/how-we-built-a-real-time-privacy-inspector, 2020.

[67] MAYER, J. R., AND MITCHELL, J. C. Third-party web tracking: Policy and technology. In *2012 IEEE symposium on security and privacy* (2012), IEEE, pp. 413–427.

[68] MELICHER, W., SHARIF, M., TAN, J., BAUER, L., CHRISTODORESCU, M., AND LEON, P. G. (do not) track me sometimes: Users' contextual preferences for web tracking. *Proc. Priv. Enhancing Technol. 2016*, 2 (2016), 135–154.

[69] MERZDOVNIK, G., HUBER, M., BUHOV, D., NIKIFORAKIS, N., NEUNER, S., SCHMIEDECKER, M., AND WEIPPL, E. R. Block me if you can: A large-scale study of tracker-blocking tools. In *EuroS&P* (2017), IEEE, pp. 319–333.

[70] MILES, M. B., HUBERMAN, A. M., AND SALDAÑA, J. *Qualitative Data Analysis: A Methods Sourcebook*, 3 ed. Sage, 2014.

[71] MOUSAVI, R., CHEN, R., KIM, D. J., AND CHEN, K. Effectiveness of privacy assurance mechanisms in users' privacy protection on social networking sites from the perspective of protection motivation theory. *Decis. Support Syst. 135* (2020), 113323.

[72] NISSENBAUM, H. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[73] NORBERG, P. A., HORNE, D. R., AND HORNE, D. A. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs 41*, 1 (2007), 100–126.

[74] OFFICE OF THE ATTORNEY GENERAL OF THE STATE OF CALIFORNIA. California Consumer Privacy Act (CCPA). https://oag.ca.gov/privacy/ccpa, March 13, 2024.

[75] OFFICE OF THE PRIVACY COMMISSIONER OF CANADA. 2022-23 Survey of Canadians on Privacy-Related Issues. https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2023/por_ca_2022-23/, 2023.

[76] OFFICE OF THE PRIVACY COMMISSIONER OF CANADA. PIPEDA requirements in brief. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/, Accessed 2025.

[77] ORTEGA, A. G., BOURGEOIS, J., AND KORTUEM, G. What is sensitive about (sensitive) data? characterizing sensitivity and intimacy with google assistant users. In *CHI* (2023), ACM, pp. 586:1–586:16.

[78] PARK, Y. J. Digital literacy and privacy behavior online. *Commun. Res. 40*, 2 (2013), 215–236.

[79] PEPRAH, P., BUDU, H. I., AGYEMANG-DUAH, W., ABALO, E. M., AND GYIMAH, A. A. Why does inaccessibility widely exist in healthcare in ghana? understanding the reasons from past to present. *Journal of Public Health 28* (2020), 1–10.

[80] PIAN, W., KHOO, C. S., AND CHANG, Y.-K. The criteria people use in relevance decisions on health information: An analysis of user eye movements when browsing a health discussion forum. *Journal of medical Internet research 18*, 6 (2016), e136.

[81] POURRAZAVI, S., HASHEMIPARAST, M., BAZARGAN-HEJAZI, S., ULLAH, S., AND ALLAHVERDIPOUR, H. Why older people seek health information online: A qualitative study. *Advances in gerontology 11* (2021), 290–297.

[82] PRINCE, C., OMRANI, N., MAALAOUI, A., DABIC, M., AND KRAUS, S. Are we living in surveillance societies and is privacy an illusion? an empirical study on privacy literacy and privacy concerns. *IEEE Trans. Engineering Management 70*, 10 (2023), 3553–3570.

[83] ROGERS, R. W. A protection motivation theory of fear appeals and attitude change1. *The journal of psychology 91*, 1 (1975), 93–114.

[84] ROUVROY, A., AND POULLET, Y. The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy. In *Reinventing data protection?* Springer, 2009, pp. 45–76.

[85] ROWLANDS, I. J., LOXTON, D., DOBSON, A., AND MISHRA, G. D. Seeking health information online: association with young australian women's physical, mental, and reproductive health. *Journal of medical Internet research 17*, 5 (2015), e4048.

[86] RUDNICKA, A., COX, A. L., AND GOULD, S. J. J. Why do you need this?: Selective disclosure of data among citizen scientists. In *CHI* (2019), ACM, p. 392.

[87] SCHOMERUS, G., SCHINDLER, S., SANDER, C., BAUMANN, E., AND ANGERMEYER, M. C. Changes in mental illness stigma over 30 years–improvement, persistence, or deterioration? *European Psychiatry 65*, 1 (2022), e78.

[88] SHAHSAVAR, Y., CHOUDHURY, A., ET AL. User intentions to use chatgpt for self-diagnosis and health-related purposes: cross-sectional survey study. *JMIR Human Factors 10*, 1 (2023), e47564.

[89] SHKLOVSKI, I., MAINWARING, S. D., SKÚLADÓTTIR, H. H., AND BORGTHORSSON, H. Leakiness and creepiness in app space: perceptions of privacy and mobile app use. In *CHI* (2014), ACM, pp. 2347–2356.

[90] SILLENCE, E., BRIGGS, P., FISHWICK, L., AND HARRIS, P. Trust and mistrust of online health sites. In *Proceedings of the Conference on Human Factors in Computing Systems* (2004), p. 663–670.

[91] SILLENCE, E., BRIGGS, P., FISHWICK, L., AND HARRIS, P. R. Trust and mistrust of online health sites. In *CHI* (2004), ACM, pp. 663–670.

[92] SINDERMANN, C., SCHMITT, H. S., KARGL, F., HERBERT, C., AND MONTAG, C. Online privacy literacy and online privacy behavior - the role of crystallized intelligence and personality. *Int. J. Hum. Comput. Interact. 37*, 15 (2021), 1455–1466.

[93] SMITH, H. J., DINEV, T., AND XU, H. Information privacy research: an interdisciplinary review. *MIS quarterly* (2011), 989–1015.

[94] SOLOVE, D. J. The myth of the privacy paradox. *Geo. Wash. L. Rev. 89* (2021), 1.

[95] SPIEKERMANN, S., GROSSKLAGS, J., AND BERENDT, B. E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior. In *EC* (2001), ACM, pp. 38–47.

[96] STORY, P., SMULLEN, D., YAO, Y., ACQUISTI, A., CRANOR, L. F., SADEH, N. M., AND SCHAUB, F. Awareness, adoption, and misconceptions of web privacy tools. *Proc. Priv. Enhancing Technol. 2021*, 3 (2021), 308–333.

[97] SUN, Y., ZHANG, Y., GWIZDKA, J., AND TRACE, C. B. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. *Journal of medical Internet research 21*, 5 (2019), e12522.

[98] TOR PROJECT. Tor project: Anonymity online. https://www.torproject.org, 2014.

[99] TREPTE, S., TEUTSCH, D., MASUR, P. K., EICHER, C., FISCHER, M., HENNHÖFER, A., AND LIND, F. Do people know about privacy and data protection strategies? towards the "online privacy literacy scale"(oplis). *Reforming European data protection law* (2015), 333–365.

[100] TUFEKCI, Z. Engineering the public: Big data, surveillance and computational politics. *First Monday* (2014).

[101] UNDERHILL, C., AND MCKEOWN, L. Getting a second opinion: Health information and the internet. *Health reports 19*, 1 (2008), 65.

[102] VISMARA, M., VITELLA, D., BIOLCATI, R., AMBROSINI, F., PIROLA, V., DELL'OSSO, B., AND TRUZOLI, R. The impact of covid-19 pandemic on searching for health-related information and cyberchondria on the general population in italy. *Frontiers in Psychiatry 12* (2021).

[103] WEINSHEL, B., WEI, M., MONDAL, M., CHOI, E., SHAN, S., DOLIN, C., MAZUREK, M. L., AND UR, B. Oh, the places you've been! user reactions to

longitudinal transparency about third-party web tracking and inferencing. In *CCS* (2019), ACM, pp. 149–166.

[104] WILSON, K., AND ROSENBERG, M. W. The geographies of crisis: exploring accessibility to health care in canada. *Canadian Geographer/Le Géographe canadien 46*, 3 (2002), 223–234.

[105] WOLFORD, B. Does the gdpr apply to companies outside of the eu? https://gdpr.eu/companies-outside-of-europe/, Accessed 2025.

[106] WU, Y., GUPTA, P., WEI, M., ACAR, Y., FAHL, S., AND UR, B. Your secrets are safe: How browsers' explanations impact misconceptions about private browsing mode. In *WWW* (2018), ACM, pp. 217–226.

[107] XIAO, N., SHARMAN, R., RAO, H. R., AND UPADHYAYA, S. J. Factors influencing online health information search: An empirical analysis of a national cancer-related survey. *Decis. Support Syst. 57* (2014), 417–427.

[108] YEUNG, K. 'hypernudge': Big data as a mode of regulation by design. In *The social power of algorithms*. Routledge, 2019, pp. 118–136.

[109] YIN, H., WARDENAAR, K. J., XU, G., TIAN, H., AND SCHOEVERS, R. A. Mental health stigma and mental health knowledge in chinese population: a cross-sectional study. *BMC psychiatry 20* (2020), 1–10.

[110] ZHAO, X., FAN, J., BASNYAT, I., AND HU, B. Online health information seeking using "# covid-19 patient seeking help" on weibo in wuhan, china: descriptive study. *Journal of Medical Internet Research 22*, 10 (2020), e22910.

[111] ZHOU, T., AND LI, H. Understanding mobile SNS continuance usage in china from the perspectives of social influence and privacy concern. *Comput. Hum. Behav. 37* (2014), 283–289.

[112] ZOU, Y., ROUNDY, K. A., TAMERSOY, A., SHINTRE, S., ROTURIER, J., AND SCHAUB, F. Examining the adoption and abandonment of security, privacy, and identity theft protection practices. In *CHI* (2020), ACM, pp. 1–15.

Appendix

6.1 Pre-Interview Questionnaire

1. Do you get anxious when browsing medical information? *Participation is not recommended if you are anxious about your health.*
    - Yes
    - No
2. Full name *last, first*
3. E-mail address *The address that was used to contact us*
4. Verification code *The code you received by email after contacting the investigators*
5. What is your age group? *You must be 18 or over to participate.*
    - 18–24
    - 25–34
    - 35–44
    - 45–54
    - 55–64
    - 65 or above
6. What is your gender identity? *If you prefer to self-describe please use the last option to enter your preferred term*
    - Woman
    - Man
    - Non-binary
    - Prefer not to say

- [Other]

7. In which province or territory do you live?
    - Alberta
    - British Columbia
    - Manitoba
    - New Brunswick
    - Newfoundland and Labrador
    - Northwest Territories
    - Nova Scotia
    - Nunavut
    - Ontario
    - Prince Edward Island
    - Quebec
    - Saskatchewan
    - Yukon
8. What is your highest level of educational attainment?
    - High school
    - Post-secondary college or equivalent
    - Bachelor's degree or equivalent
    - Master's degree or equivalent
    - PhD or equivalent
9. What is your highest level of digital literacy?
    - I can use simple software tools: a web browser, an email program, etc.
    - I can use various software tools and internet services to do a range of tasks, but I am not an advanced user of any of them.
    - I would consider myself an advanced user of one or more software technologies, but this is not my primary field of activity.

- I am an IT professional or student in a field related to Computer Science.

10. How closely, if at all, do you follow news about privacy issues?
    - Very closely
    - Somewhat closely
    - Not too closely
    - Not at all
11. In general, are you concerned about the protection of your privacy?
    - Extremely concerned
    - Concerned
    - Somewhat concerned
    - Not concerned
12. How much do you agree with the following statement - "I am confident that I have enough information to know how new technologies might affect my personal privacy"
    - Strongly agree
    - Agree
    - Neutral
    - Disagree
    - Strongly disagree
13. As far as you know, how much of what you do online or on your smartphone is being tracked by companies or organizations?
    - All or almost all of it
    - Most of it
    - Some of it
    - Very little of it
    - None of it

14. As far as you know, how much of what you do online or on your smartphone is being tracked by the government?

    - All or almost all of it
    - Most of it
    - Some of it
    - Very little of it
    - None of it

15. Thinking about the information available about you online, please tell me whether you're concerned about social media companies gathering your personal information from their platform to create a profile of your interests and personal traits for marketing purposes

    - Extremely concerned
    - Concerned
    - Somewhat concerned
    - Not concerned

16. Thinking about the information available about you online, please tell me whether you're concerned about companies or organizations using information available about you online to make decisions about you, such as for a job, an insurance claim or health coverage

    - Extremely concerned
    - Concerned
    - Somewhat concerned
    - Not concerned

17. Have you adjusted privacy settings on a social media account?

    - Yes
    - No
    - Not sure

18. Have you deleted or stopped using a social media account because of privacy concerns?
    - Yes
    - No
    - Not sure
19. Have you refused to provide an organization or business with your personal information because of privacy concerns?
    - Yes
    - No
    - Not sure
20. Have you stopped doing business with a company that experienced a privacy breach?
    - Yes
    - No
    - Not sure
21. Have you raised a privacy concern with a company or organization?
    - Yes
    - No
    - Not sure
22. How often, if at all, do you read privacy policies, notices or pop-ups when using mobile applications or conducting transactions online?
    - Always
    - Sometimes
    - Never
23. What is your primary reason for not always reading privacy notices? *Please specify if choosing "Other"*
    - They are too long

- They contain too much legal jargon
- You don't care
- Other

24. Have you or someone you know been impacted by a privacy breach?
    - Yes
    - No
    - I don't know
25. There are a growing number of news reports of sensitive personal information being lost, stolen or made public. Has this had a major effect, moderate effect, minor effect or no effect at all on your willingness to share personal information with organizations?
    - Major effect
    - Moderate effect
    - Minor effect
    - No effect at all
26. Why do you want to participate in this study?
27. Please provide us with a Canadian mailing address. *We need this information to validate that you are a resident of Canada and to send you the gift card.*

6.2 Popular Websites Considered

The following is the list of the top 50 heath websites retrieved using SimilarWeb[1] with the filters health and worldwide on 18 May 2023, ordered by popularity (across then down).

[1] similarweb.com, accessed 2023-05-18.

6.2 Popular Websites Considered

nih.gov	healthline.com	mayoclinic.org
webmd.com	medicalnewstoday.com	cvs.com
clevelandclinic.org	cdc.gov	medlineplus.gov
nhs.uk	msdmanuals.com	walgreens.com
sportkp.ru	altibbi.com	1mg.com
doctolib.fr	menshealth.com	alodokter.com
tuasaude.com	aarp.org	womenshealthmag.com
vinmec.com	halodoc.com	medonet.pl
doctoralia.com.br	drugs.com	vidal.ru
activebeat.com	verywellhealth.com	athenahealth.com
abczdrowie.pl	my-personaltrainer.it	psychologytoday.com
who.int	apteka.ru	goodrx.com
webteb.com	mscoldness.com	babycenter.com
fitbit.com	health.clevelandclinic.org	hellosehat.com
vnimanie.pro	uworld.com	everydayhealth.com
medscape.com	myfitnesspal.com	hopkinsmedicine.org
eatthis.com	rlsnet.ru	