3-4-2015

Readability of Privacy Policies of Healthcare Websites

Tatiana Ermakova

Benjamin Fabian

Eleonora Babina

Readability of Privacy Policies of Healthcare Websites

Tatiana Ermakova[1], Benjamin Fabian[2], and Eleonora Babina[2]

[1] Department of Information and Communication Management, Technical University of Berlin, Berlin, Germany
tatiana.ermakova@tu-berlin.de
[2] Institute of Information Systems, Humboldt-Universität zu Berlin, Berlin, Germany
{bfabian,babinael}@wiwi.hu-berlin.de

Abstract. Health-related personal information is very privacy-sensitive. Online privacy policies inform Website users about the ways their personal information is gathered, processed and stored. In the light of increasing privacy concerns, privacy policies seem to be an important mechanism for increasing customer loyalty. However, in practice, consumers only rarely read privacy policies, possibly due to the common assumption that policies are hard to read. By designing and implementing an automated extraction and readability analysis toolset, we present the first study that provides empirical evidence on readability of over 5,000 privacy policies of health websites and over 1,000 privacy policies of top e-commerce sites. Our results confirm the difficulty of reading current privacy policies. We further show that health websites' policies are more readable than top e-commerce ones, but policies of non-commercial health websites are worse readable than commercial ones. Our study also provides a solid policy text corpus for further research.

Keywords: Privacy, Privacy Policies, Readability, Healthcare.

1 Introduction

Healthcare websites are currently popular among Internet users. Seven out of ten US Internet users admit having looked up online for health information in 2012 [1]. However, in spite of critique [2], most healthcare websites practice collecting, analyzing and sharing their consumers' information with interested third parties [3, 4]. Consumers increasingly worry about their online privacy owing to these practices [5]. Moreover, revealed medical information can be further misused, e.g., by healthcare product retailers sending prescription reminders and promotional letters for new healthcare products and by employers making their decisions [6-8].

Websites are supposed to use privacy policies to inform the customers about their practices of collecting and using their personal information. In the presence of increasing privacy concerns, privacy policies are essential [9]. Moreover, self-reported privacy statements with a strong guarantee of security are found even more effective than third-party seals regarding customers' willingness to disclose various types of personal information [10]. However, practice shows that Internet users only rarely

12th International Conference on Wirtschaftsinformatik,
March 4-6 2015, Osnabrück, Germany

Ermakova, T.; Fabian, B.; Babina, E. (2015): Readability of Privacy Policies of Healthcare Websites, in: Thomas. O.; Teuteberg, F. (Hrsg.): Proceedings der 12. Internationalen Tagung Wirtschaftsinformatik (WI 2015), Osnabrück, S. 1085-1099

read privacy policies [11-13] and prefer third-party seals and other alternatives to reading privacy policies [9]. One of the reasons privacy policies are often ignored could be their lacking readability [14-18]. Perceived comprehension of privacy notices was shown to positively impact people's reading of privacy policies and trust in the notice [9]. In [19-22], readability of a privacy policy was even found to be positively associated with user's trust towards the website.

In our study, we thoroughly examine the reading level of privacy statements of healthcare websites and therefore their efficiency to communicate their attitudes regarding consumers' privacy which influence the formation of consumers' behavior. Previous studies regarding readability of privacy policies are based on small sample sizes of 55 [14] and 75 [18]. Our work provides a sample of privacy policies that is two orders of magnitude larger than investigated in earlier work and is thus representative. This sample has been automatically extracted from health websites. Here, we designed and employed a sophisticated automated extraction and analysis toolset. To the best of our knowledge, such an approach has never been used in similar research projects. Further, since there is no established single metric for readability in linguistic literature, we use a broader number of readability measures to provide a more complete and reliable picture of the readability level of privacy policies of healthcare websites and explore them in a more detailed statistics summary. We further investigate the readability of privacy policies of 5,431 healthcare websites compared to 1,166 top e-commerce websites based on these well-established readability measures. Finally, we statistically analyze whether commercial and non-commercial healthcare websites differ in their readability of privacy policies.

The paper is structured as follows. We first provide a theoretical background on privacy and readability of privacy policies in general as well as in the healthcare context. Then, we present the method which we used to obtain our data set, including a software prototype for readability analysis. This is followed by the analysis and results. The paper concludes with a discussion of the key findings, limitations and directions for future research.

2 Theoretical Background

2.1 Privacy

There is no single definition or interpretation of the term "privacy" in the literature [23, 24] due to its broad meaning and diverse historical use. The global and open nature of the Internet put personal information at risk of being easily collected and misused [24, 25]. In the light of these changes, the IS research community defines privacy as individuals' ability to control for themselves when, how and to what extent information about them is communicated to others [26]. Various laws and regulations such as *Directive 95/46/EC of the European Parliament and of the Council* [27] and *Health Insurance Portability and Accountability Act* of 1996 (HIPAA) [28] attempt to protect individuals' information privacy. Nevertheless, according to the research conducted by the data privacy management company TRUSTe in mid-December 2013, Internet users are high concerned about their privacy on the Internet: 92% of US con-

sumers and 89% of British consumers admit worrying about their online privacy [5]. Interestingly, according to an earlier study conducted by Pew Research Center in September 2012 [1], while 72% of US Internet users say they looked online for health information in 2012, almost eight out of ten searches for medical information were started in the search engine such as Google, Bing or Yahoo. Another 13% of online health seekers who remembered their last time they went online for health information started their search at sites focusing on health information.

Many websites and web platforms offer information or services in exchange for users' personal data. They collect information about the users, their needs and preferences in order to strategically use or sell this data for promotional purposes. It appears that patients unwillingly [2] or unknowingly trade their right for information privacy in return for medical information [29], as well. A recent study by [3] discovers that 13 out of examined 20 popular websites specializing in healthcare applied one or more tracker elements. The investigation initiated with respect to 80 PHR (personal health records) and EHR (electronic health records) websites [4] similarly shows that many of them sell medical information, use external software to scan the webpage content and track user behavior for targeted advertising. Medical information can be used by healthcare product retailers for sending prescription reminders and promotional letters for new healthcare products and by companies for basing their employment decisions on. Among potential harms, there are also impaired reputations [6-8] and ruined personal credit history. Recently, about 4.5 million patients were put at heightened risk of identity theft as their medical records, including their names, social security numbers, physical addresses, birthdays and telephone numbers were stolen by criminals who broke into computers of Community Health Systems operating 206 hospitals across the United States [30].

Almost ninety percent of both US and British Internet users admit avoiding companies that do not protect their privacy [5], what is in line with the IS research community repeatedly concluding that Internet users' privacy concerns are one of the most formidable barriers to people intending use of e-services [7, 31].

2.2 Privacy Policies

Websites usually use privacy policies to communicate their privacy attitude to their customers [13, 32]. US Legal Dictionary defines a privacy policy in the online context as "*a statement that declares a firm's or website's policy on collecting and releasing information about a visitor*" [33].

Nowadays, policy content can be represented in many conceivable formats. A vast majority of websites utilize a natural language format. Another text-based format known as the *Platform for Privacy Preferences* (P3P) protocol was developed by the World Wide Web (WWW) Consortium in order to "*enable Websites to express their privacy practices in a standard format that can be retrieved automatically and interpreted easily by user agents*" [34]. However, P3P was often criticized for being too complex and confusing for an average user and the P3P working group was closed in 2006 [34]. An approach for semi-automated privacy policy feature extraction is presented in [35], leveraging natural language processing and crowdsourcing. Since or-

ganizations usually have conflicting goals of creating notices that are easy to understand but also complete and legally compliant, the concept of *Multilayered Notices* was introduced [36]. In attempts to facilitate the overview over privacy policy content, researchers also created a number of graphical solutions. For example, the *Privacy Nutrition Label* developed by [37] works with P3P policies and displays the content in a graphical matrix which shows the types of data collected and the usages of this data. Similarly, the solution *"KnowPrivacy"* suggested by [38] is an icon-based policy summary. Nonetheless, [39] found evidence that presenting content of a privacy policy as a grid with the help of icons and colors did not improve comprehension in comparison to natural language policies.

Even in the presence of significant online privacy concerns [5], empirical evidence reveals that privacy policies are only consulted in 26% of cases where a policy is available [13]. Similarly, in a privacy study on Facebook, 77% of respondents claimed not to have read the privacy policy [11]. Several factors might explain the fact why privacy policies are not read. According to the US national phone survey conducted by the Annenberg Public Policy Center in 2005, most Internet users falsely believe that the presence of a privacy policy on a website means the site will not share their personal information with other websites and companies [40]. Similarly, the presence of a privacy statement on a website was found to positively influence users' beliefs about their privacy being better protected on the website [41], on the amount of personal information they were eager to disclose to the website [42], their trust towards the website [13], and willingness to purchase on the website [13, 43]. However, another explanation of Internet users ignoring privacy policies could be the psychological process of habituation observed by [12] in the context of consumers' desensitization to certain privacy concerns due to the increased number of reported data breaches. Analogically, it can be supposed that consumers get used to the fact that privacy policies are long, confusing and poorly written and therefore decide not to read them in the future. Studies show that privacy policies often use specific terminology unknown to a common user [16]. The average privacy policy requires two years of college level education to get comprehended [14]. Even the most readable policies are found to be too difficult to read [17]. The language of online privacy policies was found to rather downplay privacy issues and mitigate questionable practices [32]. As recently observed, many (even highly rated) mobile health app privacy policies do not even focus on the app [15]. Based on [44], Internet users possibly base their decision of either to ignore a privacy policy or not on a personal benefit/loss analysis, comparing the reduction of the information asymmetries and perceived costs of reading privacy policies, e.g., time cost.

2.3 Readability

Readability of privacy policies appears to be essential in people's decisions whether to read them or not. In general, when texts exceed the reading level of readers, they usually stop reading [9, 45]. Moreover, several studies tested and revealed a positive association between readability and user's trust [9, 19-22].

Readability can be defined as "*the ease of understanding or comprehension due to the style of writing*" [46]. One can also observe an interaction of text and reader variables in determining readability [47]. For the reader, these variables include the reader's knowledge, reading skills, interest and motivation; for the text, they are content, design, organization and style [48]. The assessment of how readable a text is can be performed either by employing a readability test on readers (e.g., [9, 18-22]) or by counting language elements in the text (e.g., [14, 15, 17]) [49].

Table 1. Overview over readability metrics
(ACW = Average number of characters per word; ALW = Average number of letters per word; ASL = Average sentence length; ASW = Average number of syllables per word; LW = Number of words with more than six characters; PDW = Percentage of words not on Dale-Chall list of 3,000 common words; SYW = Number of words with three or more syllables)

Metric	*Formula*	*Score mapping*
FRES	FRES = 206.835 - 1.015 x ASL - 84.6 x ASW	90 - 100 = Very easy = 4th grade; 80 - 90 = Easy = 5th grade; 70 - 80 = Fairly easy = 6th grade; 60 - 70 = Standard = 7th to 8th grade; 50 - 60 = Fairly difficult = Some high school; 30 - 50 = Difficult = High school or some college; 0 - 30 = Very difficult = College graduate
LIX	LIX = ASL + 100 x LW / words	0-24 = Very easy; 25- = Easy; 35-44 = Standard; 45-54 = Difficult 55 and above = Very difficult
NDC	NDC = 0.1579 x PDW + 0.0496 x ASL + 3.6365	4.9 and below = 4th and lower reading grade; 5.0 to 5.9 = 5th – 6th reading grade; 6.0 to 6. = 7th - 8th reading grade; 7.0 to 7.9 = 9th – 10th reading grade; 8.0 to 8.9 = 11th - 12th reading grade; 9.0 to 9.9 = 13th - 15th reading grade (college); 10 and above = 16th and upper reading grade (college graduate)
FKG	FKG = 0.39 x ASL + 11.8 x ASW – 15.59	US reading grade level
RIX	RIX = LW / words	7.2 and above = College; 6.2 and above = 12th reading grade; 5.3 and above = 11th reading grade; 4.5 and above = 10th reading grade; 3.7 and above = 9th reading grade; 3.0 and above = 8th reading grade; 2.4 and above = 7th reading grade; 1.8 and above = 6th reading grade; 1.3 and above = 5th reading grade; 0.8 and above = 4th reading grade; 0.5 and above = 3th reading grade; 0.2 and above = 2th reading grade; Below 0.2 = 1th reading grade
SMOG	SMOG = square root of (SYW x 30 / sentences) + 3	0 - 6 = Low-literate; 7 = Junior high school; 8 = Junior high school; 9 = Some high school; 10 = Some high school; 11 = Some high school; 12 = High school graduate; 13 - 15 = Some college; 16 = University degree; 17 - 18 = Post-graduate studies; 19+ = Post-graduate degree
CLI	CLI = 5,89 x characters/words - 30 x sentences/words - 15,8	US reading grade level
GFI	GFI = 0,4 x ASL + 100 x SYW/words	The index indicates how many formal educational years are required by readers to comprehend the text.
ARI	ARI = 0.5 x ASL + 4.71 x ALW - 21.43	US reading grade level

Readability formulas started to appear in the 1920s, and by 1973 more than 200 different readability formulas with different language variables (such as number of syllables, words and sentences) were developed. Since no singe metric can be considered superior for assessing readability, we base on the most established formulas such as *Flesch Readability Ease Score* (FRES) [50], *Laesbarhedsindex* (LIX) [51], *New Dale*

Chall Score (NDC), *Flesh-Kincaid Grade Level* (FKG) [50], *Readability Index* (RIX) [51], *Simple Measure of Gobbledygook* (SMOG), *Coleman-Liau Index* (CLI), *Gunning Fog Index* (GFI), *Automated Readability Index* (ARI) [50] and *Fry Readability Graph* (Fry) [50, 52] (see Table 1). The Fry Readability Graph works as follows: in the passage of 100 words the average number of sentences and the average number of words are calculated and then plotted in the Fry graph, and the zone where the two coordinates meet presents the grade level (see Fig. 2).

3 Method

We extracted a list of health-related websites from the popular health directory *dmoz.org* (*DMOZ*) [53] sorted by category and added to a database. The DMOZ health directory is free and easy to extract sites from. Another option was Amazon's *Alexa* Web Service, which allows the retrieval of URLs along with a category (but with costs) [54]. Crawling *Alexa.com* to extract the list of health websites was not possible due to their terms of use. A further alternative, the Google Search Service API, only allows approximately 100 requests per day, and, most importantly, would have resulted in a not categorized listing. Besides DMOZ health-related websites' list, we separately added another two smaller lists to our database: the top e-commerce (.com) websites from Alexa and a set of mobile health application privacy policies converted from CSV and PDF formats which were kindly provided by [15]. A visual overview of the project components and processes is given in Fig. 1.

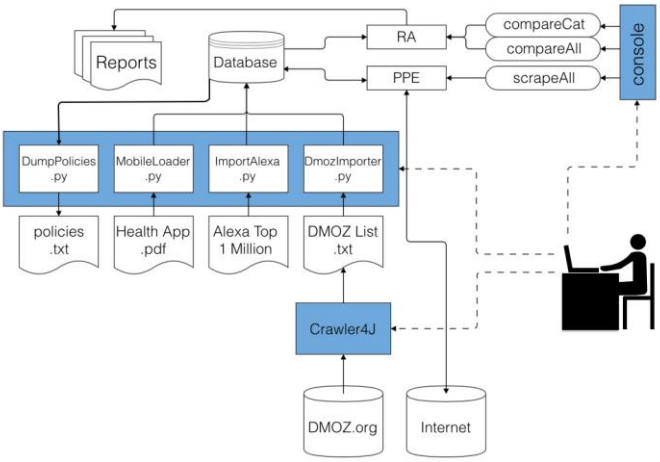

Fig. 1. Privacy policy analyzing software
(RA = Readability Analyzer, PPE = Privacy-Policy Extractor)

Our web crawler was mainly based on the open source Java library *Crawler4J* [55]. We applied crawling ethics developed by [56] that recommend that a crawler application should avoid flooding servers with too many requests and insert delays between

requests. To accelerate the crawling process, we adopted the *Pattern.Asynchronous* library for running parallel crawlers, each responsible for between 50-100 different sites. To convert the crawled policies from HTML to plain text for further analysis, we developed a Python-based *Privacy-Policy Extractor* (PPE), which mainly relies on the *boilerpipe* library's HTML parser and the Unicode data library to properly handle possible Unicode characters. Based on a simple Bayes classifier, a candidate text was deemed to be a privacy policy, if any of the following strings were found: "*privacy policy*", "*policy statements*", "*policy statement*", "*privacy*" in combination with "*cookies*". Finally, our RA component retrieved the refined and cleaned policies from the database and used readability measures to assess them based on the Python *Natural Language Toolkit* (NLTK) [57].

4 Analysis and Results

4.1 Healthcare Privacy Policies

Following the approach presented above, we retrieved a set of 5,234 unique DMOZ health websites' privacy policies together with their DMOZ categories and added the 197 mobile health apps' policies [15] to the database as the category "Mobile". Our final sample consisted of 5,431 privacy policies covering various healthcare areas which involved Medicine, Conditions and Diseases, Animal, Mental Health, Alternative, Public Health and Safety, Mobile, Addictions, Pharmacy, Nursing, Reproductive Health, Professions, Dentistry, Senior Health, and others. Additionally, another set of 1166 privacy policies of Alexa top e-commerce websites was collected. For the purpose of analysis, we imported the resulting reports into the R environment for statistical computing.

4.2 Readability of Healthcare Privacy Policies

To gain insights into readability of healthcare websites, we analyzed different readability metrics for the privacy policies in the health directory in terms of their summary statistics. Table 2 shows that the length of the privacy policies of healthcare websites ranges from 10 words till over 25,000 words. On average, they are slightly over 1,000 words long. Interestingly, 75% of the policies are about 300 words over the mean length.

As indicated by the FRES and LIX scores, the readability of the privacy policies in the health directory ranges from very easy (max. FRES = 125.70, min. LIX = 20.00) to very difficult (min. FRES = -53.71, max. LIX = 141.70) and is, on average, difficult to comprehend (mean FRES = 39.68, mean LIX = 54.02). They require an educational level of high school or some college to be read with ease, based on the Flesch Readability Ease Score (mean FRES = 39.68) and the Simple Measure of Gobbledygook (mean SMOG = 14.00). College, or 13th till 15th reading grade, is also a mean educational level to comprehend them according to the New Dale Chall Score (mean NDC = 8.93). The Flesh-Kincaid Grade Level, Coleman-Liau Index, Automated

Readability Index suggested the average US reading grade level of 13.36 (sd = 2.41), 12.99 (sd = 1.65), 14.07 (sd = 2.89) study years for reading the health directory' policies, respectively. The Readability Index (mean RIX = 7.13) maps their average readability to the 12th reading grade. The Gunning Fog Index suggests that the privacy policies of healthcare website could be easily read if an average of 16.33 formal educational years were completed. Furthermore, we constructed the Fry graph for the privacy polices of health websites, as Fig. 2 shows. It illustrates that the majority of policies lie in the grade levels 12 to 15 and above requiring college level readability.

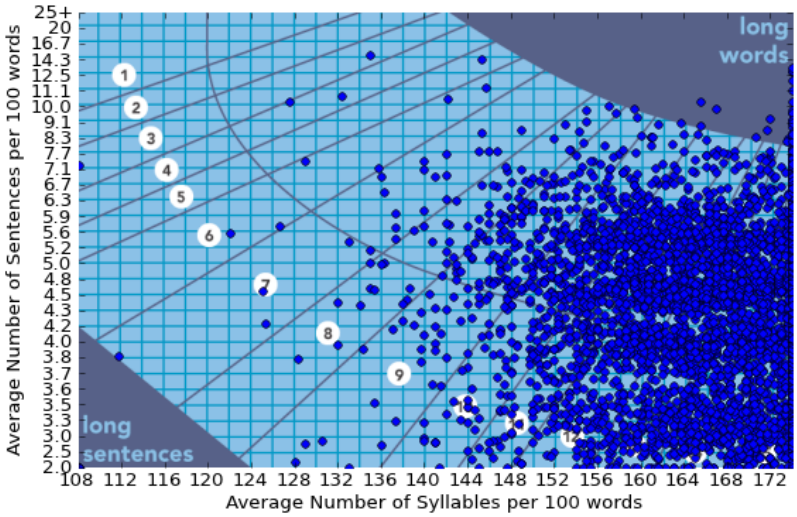

Fig. 2. Fry readability graph

4.3 Readability of Healthcare versus Top Commercial Privacy Policies

We also calculated the readability measures for top commercial website as benchmarks (see Table 2). Healthcare websites appear to provide smaller privacy policies in terms of word length when compared to e-commerce (1029 words vs. 2009 words on average). On average, other readability metrics values of the health sample are also better than those of the e-commerce scores; however, they are very close: FRES (39.68 vs. 37.49), FKG (13.36 vs. 14.09), CLI (12.99 vs. 13.18), GFI (16.33 vs. 16.77), NDC (8.93 vs. 9.14), ARI (14.07 vs. 15.00), RIX (7.13 vs. 7.73), LIX (54.02 vs. 55.99), and SMOG (14.00 vs. 14.23). These results can be followed through the violin plots in Fig. 3, which are a combination of a box plot and a kernel density plot. It appears from this visualization that better readable privacy policies are slightly more likely to be encountered on healthcare websites.

Table 2. Readability scores for 1,166 top e-commerce websites (white) vs. 5,431 healthcare websites (grey)
(Min. = Minimum, Qu. = Quartile, Max. = Maximum, Sd = Standard Deviation)

Metric	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*	*Sd*
Words	8	843	1613	2009	2690	18480	1688.23
Words	10	435	762	1029	1315	25780	988.75
FRES	-35.67	32.22	38.33	37.49	43.61	79.26	10.87
FRES	-53.71	33.50	40.22	39.68	46.06	125.70	10.74
FKG	5.67	12.62	13.91	14.09	15.40	31.17	2.48
FKG	2.94	11.89	13.21	13.36	14.69	43.16	2.41
CLI	6.84	12.33	13.05	13.18	13.75	25.21	1.71
CLI	-10.15	12.09	12.94	12.99	13.89	37.19	1.65
GFI	9.29	15.23	16.60	16.77	18.17	35.11	2.59
GFI	8.01	14.73	16.08	16.33	17.66	56.68	2.64
NDC	6.01	8.48	8.87	9.14	9.31	19.66	1.69
NDC	5.62	8.31	8.84	8.93	9.38	23.87	1.13
ARI	4.70	13.20	14.73	15.00	16.68	36.68	3.00
ARI	1.04	12.36	13.85	14.07	15.61	54.15	2.89
RIX	2.13	6.46	7.53	7.73	8.75	23.00	1.97
RIX	0.06	5.95	6.93	7.13	8.09	36.00	1.91
LIX	29.18	51.83	55.60	55.99	59.72	101.00	6.73
LIX	20.00	49.92	53.64	54.02	57.86	141.70	6.85
SMOG	9.32	13.29	14.16	14.23	15.16	23.64	1.61
SMOG	4.72	12.96	13.86	14.00	14.94	31.98	1.71

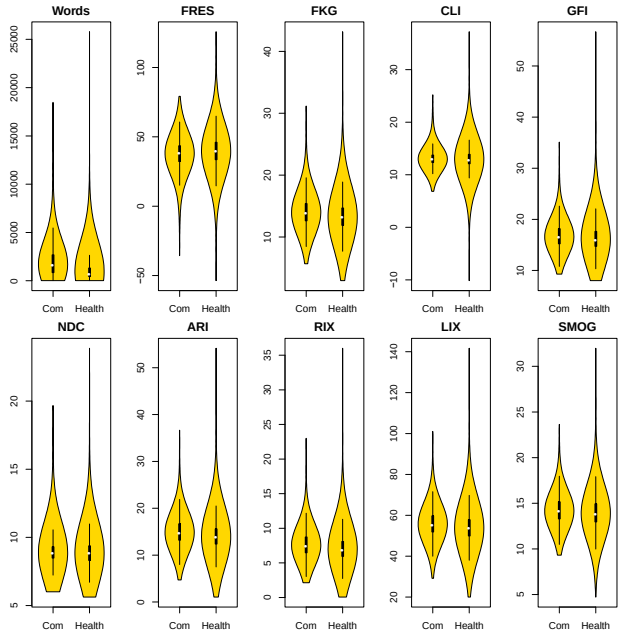

Fig. 3. Readability of top e-commerce (Com) vs. healthcare websites' (Health) privacy policies

According to Shapiro-Wilk test of normality, readability metrics of both healthcare (words: test static W = 0.7067; FRES: W = 0.9729; FKG: W = 0.9415; CLI: W = 0.9156; GFI: W = 0.9198; NDC: W = 0.7906; ARI: W = 0.9238; RIX: W = 0.923; LIX: W = 0.9482; SMOG: W = 0.9648, with p-value < 2.2e-16) and top e-commercial websites (words: W = 0.7881; FRES: W = 0.9241; FKG: W = 0.9412; CLI: W = 0.8254; GFI: W = 0.9283; NDC: W = 0.5171; ARI: W = 0.9325; RIX: W = 0.919; LIX: W = 0.9489; SMOG: W = 0.9573, with p-value < 2.2e-16) are not normally distributed. So we further applied Mann-Whitney-Wilcoxon tests with continuity correction to compare readability metrics for healthcare and top e-commercial privacy. The Mann-Whitney-Wilcoxon test does not require that the compared groups follow a normal distribution. Our results reveal that healthcare and top e-commercial privacy policies differ in terms of words (test statistic W = 4594598, p-value < 2.2e-16), FRES (W = 2788970, p-value = 1.614e-10), FKG (W = 3765321, p-value < 2.2e-16), GFI (W = 3543111, p-value = 1.699e-10), ARI (W = 3803187, p-value < 2.2e-16), RIX (W = 3798039, p-value < 2.2e-16), LIX (W = 3746232, p-value < 2.2e-16), SMOG (W = 3479611, p-value = 1.095e-07) at a significance level of 1%, and in terms NDC (W = 3311923, p-value = 0.01357) and CLI (W = 3284184, p-value = 0.04569) at a significance level of 5%.

We conclude that policies of healthcare websites are significantly shorter and generally provide significantly higher readability of their privacy policies than top commercial websites.

4.4 Readability of Commercial versus Non-Commercial Healthcare Websites

In order to test possible differences in privacy policies between commercial and non-commercial health websites, we selected two groups from the original data set: 2,723 privacy policies representing commercial websites (2,577 .com and 146 co.uk) and 2,030 coming from non-commercial websites (81 .gov, 151 .edu and 1,798 .org). While the use of .gov and .edu is restricted to governmental and educational entities, .org can be purchased by anyone. However, the .org domain is recommended for non-commercial organizations, such as NPOs, communities and philanthropic projects [61]. Therefore, the assumption can be made that .gov, .edu and .org websites mainly pursue non-commercial interests.

Table 3. Readability of 2,723 healthcare commercial websites' privacy policies (white) vs. 2,030 healthcare non-commercial websites' privacy policies (grey)
(Min. = Minimum, Qu. = Quartile, Max. = Maximum, Sd = Standard Deviation)

Metric	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*	*Sd*
Words	18	421	745	1011	1306	10580	921.91
Words	10	442.2	757	1005	1292	10460	864.58
FRES	-20.52	33.99	40.54	39.93	45.75	89.09	10.33
FRES	-53.71	32.60	39.53	38.89	45.66	85.86	10.52
FKG	3.24	11.81	13.12	13.25	14.62	38.31	2.32
FKG	2.94	12.06	13.32	13.55	14.80	37.15	2.36
CLI	5.93	12.13	12.90	12.96	13.76	20.38	1.51

CLI	6.81	12.14	13.05	13.10	14.02	22.02	1.53
GFI	8.07	14.58	15.96	16.15	17.54	41.87	2.49
GFI	8.07	14.94	16.22	16.57	17.89	43.37	2.6
NDC	5.77	8.28	8.75	8.83	9.26	19.37	0.97
NDC	5.62	8.39	8.94	9.03	9.58	19.96	1.05
ARI	3.77	12.25	13.76	13.91	15.50	45.21	2.75
ARI	3.57	12.55	13.99	14.30	15.77	42.22	2.82
RIX	1.83	5.85	6.85	7.02	8.00	25.24	1.82
RIX	1.00	6.05	7.00	7.28	8.20	36.00	1.97
LIX	29.58	49.68	53.25	53.63	57.56	117.60	6.53
LIX	20.00	50.35	54.02	54.50	58.16	120.60	6.77
SMOG	8.48	12.87	13.80	13.89	14.80	23.64	1.62
SMOG	8.48	13.12	13.98	14.18	15.09	31.98	1.73

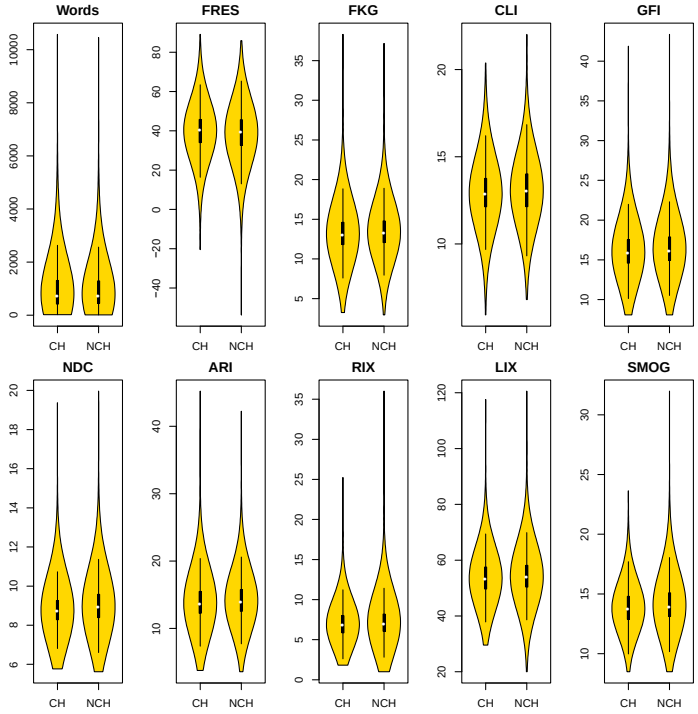

Fig. 4. Readability of commercial (CH) vs. non-commercial healthcare (NCH) privacy policies

According to Shapiro-Wilk test of normality, readability metrics of both commercial (words: W = 0.7834; FRES: W = 0.9789; FKG: W = 0.9543; CLI: W = 0.9755; GFI: W = 0.9537; NDC: W = 0.8746; ARI: W = 0.945; RIX: W = 0.9401; LIX: W = 0.9649; SMOG: W = 0.9808, with p-value < 2.2e-16) and non-commercial healthcare websites (words: W = 0.8092; FRES: W = 0.9742; FKG: W = 0.9566; GFI: W = 0.9461; NDC: W = 0.8862; ARI: W = 0.9522; RIX: W = 0.8991; LIX: W = 0.9611; SMOG: W = 0.9531 with p-value < 2.2e-16; CLI: W = 0.9881 with p-value < 6.363e-

12) are also not normally distributed. As before, we further applied Mann-Whitney-Wilcoxon tests with continuity correction to compare readability metrics for commercial and non-commercial healthcare websites' privacy policies. Our results reveal that commercial and non-commercial healthcare websites' privacy policies differ significantly in terms of FRES (W = 2763822, p-value = 0.000947), FKG (W = 2432050, p-value = 4.194e-05), CLI (W = 2452722, p-value = 0.0002777), GFI (W = 2372778, p-value = 5.995e-08), NDC (W = 2279648, p-value = 6.535e-14), ARI (W = 2411799, p-value = 5.407e-06), RIX (W = 2428523, p-value = 2.977e-05), LIX (W = 2428427, p-value = 2.949e-05), SMOG (W = 2374000, p-value = 6.978e-08) at a significance level of 1%. However, they did not differ significantly in terms of words (W = 2591334, p-value = 0.5874).

We conclude that commercial and non-commercial healthcare websites have similarly long privacy policies; however, commercial healthcare websites display more readable privacy policies.

5 Conclusion, Implications and Further Research

In the present work, we investigated the readability of a large and representative number of privacy policies of healthcare websites in general and in groups, as well as in comparison to top commercial websites. Privacy policies in the healthcare domain are difficult to read, what is consistent with prior research. They contain a mean of slightly more than 1,000 words. On average, a reader is expected to be educated at the college level, to have the 13th reading grade level or be 16 years formally educated. Healthcare websites provide shorter and in general more readable privacy policies than top e-commerce websites. Commercial and non-commercial healthcare websites have identically long privacy policies, although the policies of commercial healthcare websites are more readable.

Through integrating diverse metrics with different language variables and approaches and reporting several summary statistics, our work enables in-depth readability analysis and better comparison potential for future research studies in this field. For practice, our results imply that in terms of their readability, privacy statements of current healthcare websites do not appropriately communicate their attitude regarding consumers' privacy on the website and do not positively influence the formation of consumers' behavior. Healthcare websites' providers, especially those working on a non-commercial basis, should make serious efforts to rewrite these statements. In particular, improving privacy policies should be a concern to non-commercial healthcare but also top e-commerce website providers.

In our follow-up research, we are going to take more factors into account such as design, content and organization [48]. In reality, organizational elements such as bullet-points are used to visualize the content and help general comprehension. Second, policies in other languages are to be crawled and analyzed. Third, content analyses of the current data sets are promising in the future. The content of policies concerning sensitive topics [62] (e.g., categories Addictions, Mental Health) are to be compared to less sensitive categories. Moreover, certain archetypes of privacy policies can be

derived from researching similarities and common patterns using machine learning and text mining. In general, we are planning to investigate the ways to make privacy statement more comprehensive and of better value to websites' consumers, and to work out improvement guidelines and recommend exemplary privacy policies. [1]

References

1. Pew Research Center: Health Online, http://www.pewinternet.org/files/old-media//Files/Reports/PIP_HealthOnline.pdf (Accessed: 16.11.14) (2013)
2. Schwartz, P.M.: Internet Privacy and the State. Connecticut Law Review 32, 815-860 (1999)
3. Huesch, M. D.: Privacy Threats when Seeking Online Health Information. JAMA Internal Medicine 173, 1838-1840 (2013)
4. Kaletsch, A., Sunyaev, A.: Privacy Engineering: Personal Health Records in Cloud Computing Environments. In: ICIS (2011)
5. TRUSTe, http://www.truste.com/(Accessed: 16.11.14)
6. Rohm, A.J., Milne, G.R.: Just What the Doctor Ordered – The Role of Information Sensitivity and Trust in Reducing Medical Information Privacy Concern. Journal of Business Research 57, 1000-1011 (2004)
7. Bansal, G., Zahedi, F., Gefen, D.: The Impact of Personal Dispositions on Information Sensitivity, Privacy Concern and Trust in Disclosing Health Information Online. Decision Support Systems 49, 138-150 (2010)
8. Laric, M.V., Pitta, D.A., Katsanis, L.P.: Consumer Concerns for Healthcare Information Privacy: A Comparison of U.S. and Canadian Perspectives. Research in Healthcare Financial Management 12, 93-111 (2009)
9. Milne, G.R., Culnan, M.J.: Strategies for Reducing Online Privacy Risks: Why Consumers Read (or don't Read) Privacy Notices. Journal of Interactive Marketing 18, 15-29 (2004)
10. Peterson, D.; Meinert, D.; Criswell II, J.; Crossland, M.: Consumer Trust: Privacy Policies and Third-Party Seals. Journal of Small Business and Enterprise Development 14, 654-669 (2007)
11. Acquisti, A., Gross, R: Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. Privacy Enhancing Technologies 36-58 (2006)
12. Acquisti, A.: The Economics of Personal Data and the Economics of Privacy. In: OECD Joint WPISP-WPIE Roundtable 1 (2010)
13. Jensen, C., Potts, C., Jensen, C.: Privacy Practices of Internet Users: Self-Reports versus Observed Behavior. International Journal of Human-Computer Studies 63, 203-227 (2005)
14. Graber, M.A., Donna M. D Alessandro, Johnson-West, J.: Reading Level of Privacy Policies on Internet Health Web Sites. Journal of Family Practice 51, 642-642 (2002)
15. Sunyaev, A., Dehling, T., Taylor, P.L., Mandl, K.D.: Availability and Quality of Mobile Health App Privacy Policies. Journal of the American Medical Informatics Association (2014)
16. Antón, A I., Bertino, E., Li, N., Yu, T.: A Roadmap for Comprehensive Online Privacy Policy Management. Communications of the ACM 50, 109-116 (2007)
17. McDonald, A.M., Cranor L.F.: The Cost of Reading Privacy Policies. ISJLP 4, 543 (2008)

[1] The work presented in this paper was performed in part to support the TRESOR research project [63].

18. McDonald, A.M., Reeder, R.W., Kelley, P.G., Cranor, L.F.: A Comparative Study of Online Privacy Policies and Formats. Privacy Enhancing Technologies, 37-55 (2009)
19. Ermakova, T., Baumann, A., Fabian, B., Krasnova, H.: Privacy Policies and Users' Trust: Does Readability Matter? In: AMCIS (2014)
20. Sultan, F., Urban, G. L., Shankar, V., Bart, I. Y.: Determinants and Role of Trust in e-Business: A Large Scale Empirical Study. MIT Sloan School of Management, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=380404 (Accessed: 16.11.2014) (2002)
21. Bansal, G.; Zahedi, F.; Gefen, D.: The Moderating Influence of Privacy Concern on the Efficacy of Privacy Assurance Mechanisms for Building Trust: A Multiple-Context Investigation. In: ICIS (2008)
22. Bansal, G.; Zahedi, F.; Gefen, D.: Efficacy of Privacy Assurance Mechanisms in the Context of Disclosing Health Information Online. In: AMCIS (2008)
23. Dinev, T., Xu, H., Smith, H. J., Hart, P.: Information Privacy and Correlates: An Empirical Attempt to Bridge and Distinguish Privacy-Related Concepts. European Journal of Information Systems 22, 295-316 (2013)
24. Smith, H.J., Dinev, T., Xu, H.: Information Privacy Research: an Interdisciplinary Review. MIS Quarterly 35, 989-1016 (2011)
25. Gutwirth, S.: Privacy and the Information Age. Rowman & Littlefield (2002)
26. Westin, A. F.: Privacy and Freedom. Atheneum (1967)
27. European Parliament and Council: Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, Official Journal 281, 31–50 (1995)
28. U.S. Department of Health & Human Services: Health Insurance Portability and Accountability Act of 1996 (HIPAA), http://www.hhs.gov/ocr/privacy/ (Accessed: 08.12.14) (1996)
29. Clarke, R.: Introduction to Dataveillance and Information Privacy, and Definitions of Terms, http://www.rogerclarke.com/DV/Intro.html (Accessed: 12.08.14) (1999)
30. CNNMoney: Hackers … stole data on 4.5 million patients, http://money.cnn.com/2014/08/18/technology/security/hospital-chs-hack/index.html?iid=SF_T_Lead (Accessed: 16.11.14) (2014)
31. Dinev, T., Hart, P.: Internet Privacy Concerns and Social Awareness as Determinants of Intention to Transact. International Journal of E-Commerce 10, 7-29 (2006)
32. Pollach, I.: What's Wrong with Online Privacy Policies? Communications of the ACM 50, 103-108 (2007)
33. BusinessDictionary: Definition of Privacy Policy, http://www.businessdictionary.com/definition/privacy-policy.html (Accessed: 16.11.2014)
34. W3C, http://www.w3.org/P3P/ (Accessed: 16.11.14)
35. Sadeh, N., Acquisti, A., Breaux, T.D., Cranor, L.F., McDonald, A.M., Reidenberg, J., Smith, N.A., Liu, F., Russell, C., Schaub, F., Wilson, S., Graves, J.T., Leon, P.G., Ramanath, R., Rao, A.: Towards Usable Privacy Policies: Semi-Automatically Extracting Data Practices From Websites' Privacy Policies. In: SOUPS (2014)
36. Hunteon Privacy Blog. 10 Steps to Multilayered Privacy Notice, https://www.huntonprivacyblog.com/wp-content/files/2012/07/Centre-10-Steps-to-Multilayered-Privacy-Notice.pdf (Accessed: 16.11.14)
37. Kelley, P.G., Bresee, J., Cranor, L.F., Reeder, L.W.: A Nutrition Label for Privacy. In: 5th Symposium on Usable Privacy and Security (2009)
38. KnowPrivacy, http://www.knowprivacy.org/ (2009)

39. Reeder, R.W., Kelley, P.G., McDonald, A.M., Cranor, L.F.: A User Study of the Expandable Grid Applied to P3P Privacy Policy Visualization. In: WPES (2008)
40. Feldman, L., Turow, J., Meltzer, K.: Open to Exploitation: American Shoppers Online and Offline. Annenberg Public Policy Center (2005)
41. Li, H., Sarathy, R., Xu, H.: The Role of Affect and Cognition on Online Consumers' Decision to Disclose Personal Information to Unfamiliar Online Vendors. Decision Support Systems 51, 434-445 (2011)
42. Hui, K.L., Teo, H.H., Lee, S.Y.T.: The Value of Privacy Assurance: an Exploratory Field Experiment. MIS Quarterly 31, 19-33 (2007)
43. Tsai, J.Y., Egelman, S., Cranor, L., Acquisti, A.: The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study. Information Systems Research 22, 254-268 (2011)
44. Acquisti, A., Grossklags, J.: Privacy and Rationality in Individual Decision Making. IEEE Security & Privacy 2, 24-30 (2005)
45. Egelman, S., Tsai, J., Cranor, L.F., Acquisti, A.: Timing is Everything?: the Effects of Timing and Placement of Online Privacy Indicators. In: International Conference on Human Factors in Computing Systems (2009)
46. Klare, G.R..: Measurement of Readability. Iowa St. (1963)
47. Harris, T.L., Hodges, R.E. (Eds.): The Literacy Dictionary: The Vocabulary of Reading and Writing. International Reading Association, Newark (1995)
48. DuBay, W.H.: Smart Language: Readers, Readability, and the Grading of Text. BookSurge Publishing (2007)
49. Klare, G.R.: Assessing Readability. Reading Research Quarterly, 62-102 (1974)
50. Shedlosky-Shoemaker, R., Sturm A.C., Saleem, M., Kelly, K.M.: Tools for Assessing Readability and Quality of Health-Related Web Sites, Journal of Genetic Counseling 18, 49-59 (2008)
51. Anderson, J.: Lix and Rix: Variations on a Little-Known Readability Index. Journal of Reading, 490-496 (1983)
52. Fry, E.: A Readability Formula that Saves Time. Journal of Reading 11, 513-578 (1968)
53. DMOZ, http://dmoz.org (Accessed: 16.11.14)
54. AWS, http://aws.amazon.com/de/awis/ (Accessed: 16.11.14)
55. Crawler4J. Google Code, http://code.google.com/p/crawler4j (Accessed: 16.11.14)
56. Thelwall, M., Stuart, D.: Web Crawling Ethics Revisited: Cost, Privacy, and Denial of Service. Journal of the American Society for Information Science and Technology 57, 1771-1779 (2006)
57. Natural Language Toolkit, http://www.nltk.org (Accessed: 16.11.14)
58. Fitzsimmons, P.R., Michael, B.D., Hulley, J.L., Scott, G.O.: A Readability Assessment of Online Parkinson's Disease Information. The Journal of the Royal College of Physicians of Edinburgh 40, 292-296 (2010)
59. Hedman, A.S.: Using the SMOG Formula to Revise a Health-Related Document. American Journal of Health Education 39, 61-64 (2008)
60. Ley, P., Florio, T.: The Use of Readability Formulas in Health Care. Psychology, Health & Medicine 1, 7-28 (1996)
61. ORG Domain, http://pir.org/domains/org-domain/ (Accessed: 16.11.14)
62. Dickson-Swift, V., James, E.L., Liamputtong, P.: What is Sensitive Research? Undertaking Sensitive Research in the Health and Social Sciences: Managing Boundaries, Emotions and Risks, 1-10 (2008)
63. TRESOR, http://www.cloud-tresor.com/ (Accessed: 08.12.14)