The Gathering Of User Data By National Medical Association Websites

K Masters

Citation

K Masters. *The Gathering Of User Data By National Medical Association Websites*. The Internet Journal of Medical Informatics. 2012 Volume 6 Number 2.

Abstract

Introduction: Medical informatics ethics is concerned with a range of issues such as confidentiality, privacy, security, informed consent, data sharing and doing no harm. Websites use analytics tools to gather data about visitors, or users. The purpose of this study is to examine national medical associations' websites in order to assess the extent to which they gather and share user data, and their adherence to ethical principles when doing so.Method: The websites of the 100 national medical associations listed as members of the World Medical Association (WMA) were investigated. Their analytics tools were identified by using a Firefox plugin, Ghostery, and further information about data usage and protection was obtained from the Ghostery database and from the analytics tools' data privacy policies.Results: Of the 100 associations, 80 websites could be accessed. Of these, 38 (47.5%) gathered user data, and only seven had any indication that user data were being gathered. Other ethical principles, such as consent, were even more frequently ignored. The most commonly used analytics tool was Google Analytics, used by 31 sites.Discussion and Conclusion: The extent to which informed consent is being ignored, data are being shared with third parties and retained for an unknown amount of time, and potential harm to users by an open exposure of data is of concern. Recommendations are put forward so that national medical associations' websites might follow practices that are in keeping with the principles of medical informatics ethics.

INTRODUCTION

Medical Informatics Ethics is the meeting point of medical practice, informatics and ethics [1]. As a result, issues of confidentiality, privacy, security, and informed consent that have been developed and refined by numerous medical codes such as the Nuremberg Code [2; 3], the Declaration of Helsinki (DoH) [4] and informatics codes, such as the US Association for Computing Machinery's Code of Ethics and Professional Conduct [5] and the Canadian Information Processing Society's Code of Ethics and Professional Conduct [6] are also found in medical informatics ethics codes. Examples of such codes are The American Medical Informatics Association's (AMIA) Code of Professional Conduct [7], the UK Council for Health Informatics Professions' UKCHIP Code of Conduct [8], and the European Parliament Directive (95/46/EC) of 1995 and later documents [9-13]. Among other things, these codes refer directly to honouring the rights of the individual, informed consent (including the right of withdrawal), respecting privacy and confidentiality (especially by protecting data), data sharing, and doing no harm.

These issues are frequently discussed in the light of patient or research subject information, usually with specific reference to electronic health records (EHRs) [14-17]. The area of concern in this study, however, is the unobtrusive and invisible tracking of users as they access national medical associations' websites.

Because of reports in the popular press [18; 19], many Internet users are now aware that search engines, such as Google, track their searches. What is less commonly-known, however, is that individual web sites use special software tools, known as web analytics tools, to track users' activities, and gather information about those users. There is no physical requirement for websites to obtain permission from users to gather this information, and it usually occurs without users' consent or even knowledge. There is also seldom any option to opt out of data collection.

Some web analytics tools (e.g. Piwik Analytics) are installed on the same web server as the website being visited by the user, and the owners of the website control the sharing of the gathered data. Others, however, are hosted by the analytics company. For example, Google Analytics, is hosted by

Google. This means that, in the very operation of Google Analytics, the data from the website are being shared with Google.

The type of information gathered falls within one of three categories:

--Anonymous information, including browser type, language, number of page views, date and time, and referring website.

--Pseudonymous information, including the user's Internet Protocol (IP) address. For a home user, an IP address can be used to identify the particular home, although not the specific computer in that home.

--Personally Identifiable Information (PII), such as a device number that uniquely identifies a computer or other device connected to the Internet. For computers, this would include the Media Access Control (MAC) address. In some instances, such as the home user, the US National Institute of Standards and Technology (NIST) considers IP addresses to be PII [20].

Most analytics tools gather a range of data, falling into more than one of these categories.

A Firefox add-on, Ghostery, allows users to identify various analytics tools that are gathering data about them as they browse Internet sites. (Ghostery also allows users to block most of these tools from gathering these data).

A cursory glance at The World Medical Association's (WMA) website (http://www.wma.net/en/10home/index.html) indicates that it uses such a tool (Piwik Analytics) to gather user data. Similarly, a brief examination of national medical associations' websites reveals that several of these sites also gather user data, and do not appear to inform the users of this data gathering. It may be that concepts of informed consent and data protection, so central to medical ethics and medical informatics ethics, are being ignored by national medical associations.

In the light of these associations' presumed adherence to medically ethical practices, there are several questions that need to be addressed about this data gathering. Which national medical association sites gather user data? What data are gathered, where are they stored, how are they used, and with whom are they shared? Do the sites inform their users that they are being tracked? Do the sites obtain consent from their users to obtain these data? Do they offer an option to opt out of having their data gathered and/or shared with third parties?

This study attempts to answer these questions, with a view to examining the practice of data gathering by national medical association websites within the context of medical informatics ethics.

METHOD

During April and May 2012, the websites of the 100 national medical associations listed as members of the WMA were visited. Where no web address (URL) for the medical association was listed, or where the given URL was incorrect, a Google search was used to try to find the association's site. For each association's site found, the following operations were performed:

Using the Firefox plugin, Ghostery (Ver 2.7.2 updated 2 April 2012), the analytics tools on the site's home page were identified. (Pages other than the home page were ignored). If a website had multiple languages, and used the same analytics tool across languages, then the analytics tool was identified only once.

Information about the analytics tools most commonly used by the medical associations was obtained from the Ghostery database, and from the data protection policies of the various tools.

Associations' pages were investigated to find any notice informing the user of the use of data-gathering or analytics tools. These included pages labelled "Privacy," "Terms of use" or "Disclaimer." In addition, the "About Us" (or equivalent) page was investigated.

Where pages were in a language not spoken by the author, Google's translate feature was used to identify pages.

Data were placed into a Microsoft Excel spread sheet, and descriptive statistical analyses were performed.

Ethics approval for the study was granted by the SQU College of Medicine & Health Sciences Medical Research Committee and Ethics Committee (MREC#543).

RESULTS

Of the 100 national medical associations, the websites of 80 could be found. This was taken as the sample size.

Of the 80 sites, 38 (47.5%) used analytics tools to gather user data. Of these 38 sites, 24 (63.2%) used one analytics

tool, nine (23.7%) used two analytics tools, four (10.5%) used three analytics tools, and one (2.6%) used four analytics tools.

Of the 38 sites using analytics tools, the most commonly used tool was Google Analytics, used by 31 (81.6%) of the sites. Other analytics tools used were Facebook Social Plugins, Twitter Badge, and WebTrends (three sites each), and AddThis, Facebook Connect, and Twitter Button (two sites each). A further 13 analytics tools were used by only one site each.

Of the 38 sites using analytics tools, only seven (18.4%) had any information regarding the gathering of user data. Of these seven:

Four had a privacy page indicating that they gathered data, and indicated that these data may be shared with third parties. Of these four, two provided an email address that the user could contact to object to the collection and sharing of personal information. It is unlikely, however, that this could apply to anonymous information (because, being anonymous, one's information could not easily be identified to be removed).

One had a privacy page indicating that it gathered data, but did not indicate that the data were gathered by an analytics tool that shared the information with third parties.

One indicated that the gathered data would not be shared with third parties, but used DoubleClick, Google Analytics, and WebTrends, all of which are third parties, and may also share their data with other third parties.

One had information about the concept of using cookies to gather data, but no specific information about data gathered and or shared by the site.

Unlike the other menu options, the Privacy statement was usually located through a link at the bottom of the page, in plain text, and in a font smaller than the font reserved for menu and other headings on the page. In no instance was there an "Opt out" option, such as a simple button to be pressed that would automatically stop a user's data from being gathered or shared. In other words, although some sites "informed," the requirement of "consent" was not met.

The information gathered on the most commonly used (i.e. by more than two sites) analytics tools is presented in Table I. This information includes the type of data collected, and the analytics tools' policies on security and data sharing.

Figure 1

Name	Data Collected	Data Sharing	Data Retention
Google Analytics [21; 22]	Anonymous: ad serving domains, browser type, demographics, language settings, page views, time/date. Pseudonymous: IP address.	Anonymous data are shared with third parties.	Undisclosed.
Facebook Social Plugins [23; 24]	Anonymous: browser type, location, page views. Pseudonymous: IP address, actions taken.	Data are shared with third parties.	Data are deleted from backup storage after 90 days.
Twitter Badge [25; 26]	Anonymous: ad clicks, browser type, mobile carrier, page views, referring URLs, time/date. Pseudonymous: IP address, search queries, tweets. PII: device and application IDs.	Anonymous and aggregate data are shared with third parties.	Log data are deleted or separated from PII after 18 months.
Web Trends [27; 28]	Anonymous: browser type, exit pages, ISP, operating system, page views, referring URL, time/date. Pseudonymous: IP address.	Anonymous and aggregate data are shared with third parties. Pseudonymous and PII data are shared with third party service providers.	Undisclosed.
AddThis [29; 30]	Anonymous: browser type, page views, referring URL, traffic statistics, time/date. Pseudonymous: IP address. PII: unique device ID.	Aggregate data are shared with third parties.	Undisclosed.
Facebook Connect [24; 31]	Anonymous: browser type, location, page views. Pseudonymous: IP address, actions taken.	Data are shared with third parties.	Data are deleted from backup storage after 90 days.
Twitter Button [26; 32]	Anonymous: ad clicks, browser type, mobile carrier, page views, referring URLs, time/date. Pseudonymous: IP address, search queries, tweets. PII: device and application IDs.	Anonymous and aggregate data are shared with third parties.	Log data are deleted or separated from PII after 18 months.

DISCUSSION

This study has shown that, of the 80 national medical associations that are members of the World Medical Association and have accessible websites, 31 (38.8%) gather data about the users of their websites. Only seven sites give any indication that some data are being gathered, only four indicate that these data may be shared with third parties, and only two have a mechanism for users to object to the collection of their data. Even in these cases, the effectiveness of this objection is unclear, and there is no simple opt -out option for users.

There are several areas of concern raised by these results.

INFORMED CONSENT

The issue of informed consent is central to medical ethics, and also to medical informatics ethics. So crucial is it, that it is found in every ethics code that deals with doctor-patient and researcher-research subject relationship and in informatics ethics codes [2-9; 11; 12; 33]. Given this prominence (and apart from speculation), it is difficult to

explain why so few national medical associations have informed their visitors that their data are being gathered. Technical reasons cannot be used as an excuse, as the information needs only to be delivered as a simple text webpage, linked to by a Privacy heading.

Similarly, the consent of the user, and the ability to change his / her mind is central to medical ethics. Just as users should be able to opt out of receiving advertising [34], so they should be able to opt out of being tracked. Although creating a single "Opt out" button on the Privacy page would require some programming, it is a relatively trivial exercise.

DATA SHARING

Also inherent in medical ethics codes is the concern about third parties' accessing data. While it is recognised that data may be shared (once informed consent has been given), full disclosure about the nature of the sharing is required. Even if the third party does not meet the strict definitions of a Trusted Third Party (TTP) [35] or Attested Trusted Third Party (ATTP) [15], there is a duty to ensure that the third party follows ethics guidelines on the storage and further sharing of the data.

Again, the technicalities of informing users of this data sharing involves a simple web page. It also, however, requires that the information is accurate, and not misleading, as is the case in one of the sites.

DATA RETENTION

Explicit in most medical, medical informatics codes, and informatics principles outside medicine, is the realisation that, once the purpose of any research has been met, any data that have been collected should be destroyed [14; 20; 34]. While four of the analytics engines listed in Table I state explicitly that they remove data, three do not. Particularly worrying, given that it is so widely used, is that Google Analytics does not give this information.

HONESTY AND IGNORANCE

The fact that one site indicates that the gathered data are not shared with third parties, but then uses tools that, by their very nature, are third parties, is worrying. Given the stature of the national medical association, it is unlikely that this was done dishonestly. It is more likely that the association employs technical staff who perform work that many would consider standard, but that the technical staff are unaware of the medical ethics policy of not sharing data. This does not excuse the medical organisation entirely, as the ethical responsibility for safeguarding data cannot simply be transferred to technical personnel. It is incumbent on national medical associations to ensure that their technical staff are well-aware of ethical issues involved.

WHERE IS THE HARM?

Is this a storm in a teacup? After all, much of these data are anonymous or pseudonymous, and are not as explicit as the data found in EHRs.

Firstly, this question avoids the central issue: the collection of any data requires informed consent and the option to opt out. In addition, this informing should be obvious, and the opt-out should be as simple as possible, in order to be the equivalent of a research subject simply saying "No."

Secondly, when anonymous and pseudonymous data are collated, and cross-referenced, they then become information, and can be used to identify a specific person [20; 36-38] in a process known as "de-anonymizing" or "reidentification." [36; 37] After all, an important reason for collecting data is to perform marketing targeted at specific users or households. The fact that IP addresses are routinely used to identify persons who download illegal music and films is testament to the power of knowing that simple piece of information. In addition, Google's own Privacy Policy Document [22] describes in some detail how a MAC address or IP address can be combined with anonymous data sent by partners (in this case, the national medical association's pages) and email addresses to uniquely identify an individual. From there on, tracking that individual's activity on the Internet is a trivial affair.

Readers should note that Google's Privacy Policy goes on to say that "Our Privacy Policy applies to all of the services offered by Google Inc. and its affiliates, including services offered on other sites (such as our advertising services), but excludes services that have separate privacy policies that do not incorporate this Privacy Policy." [22] In other words, Google's Privacy Policy appears to apply to all other partners, except where it does not. This rather ironic approach does not appear to meet basic requirements of data sharing in medical informatics ethics.

At this stage, it important to remember that the gathering of such data does not discriminate, and data from all population groups, including teenagers and younger children, are being gathered. While one might argue that these groups are not likely to visit national medical associations' pages, this is little comfort to the parents of those children who do visit those sites. Just as commercial enterprises are encouraged to

be wary of gathering such information [34], so national medical associations should also desist from this.

RECOMMENDATIONS FOR NATIONAL MEDICAL ASSOCIATIONS' WEBSITES

The United States' "Do not Track" legislation, aimed primarily at protecting children, is currently being prepared. National medical associations, however, guided by principles of medical ethics and medical informatics ethics, do not have to wait for legislation to do the right thing. These associations should adhere to the FTC's principle of "Privacy by Design," [34] and ensure that their users are protected.

Given the information outlined in the paper thus far, the author would like to make the following recommendations for national medical associations that wish to collect information on the users accessing their web sites:

Inform the user: Every national medical association site should have a simple page informing the user of:

In fairness to users, and to avoid being accused of doing only the bare minimum, the Privacy link to this information page should be displayed as prominently as any main subject heading in the site's menu system.

The information page should have "Opt out" buttons that allow the users to opt out of having their data gathered and shared. If the individual associations do not have access to the technical expertise to create such buttons, these could be developed by the WMA, and the code embedded into the page, in much the same way that they routinely embed buttons for Facebook, Twitter and other social networking sites.

Assuming informed consent has been granted, data should be shared with third parties only if those third parties undertake to be bound by the same ethical rules that govern the national medical association.

National medical associations should use only analytics tools that allow them to control the retention of data, or tools that indicate how long the data will be retained. This information must be given to the user.

National medical associations should ensure that their technical staff (or, in the case of outsourcing, technical support companies) are aware of the medical ethics policies on gathering, sharing and retaining data. The associations should run periodic checks to ensure that the standards of these ethics policies are being maintained.

AREAS OF FURTHER STUDY

This study has raised further questions that require attention in future research. A starting point should be a deeper exploration of the associations' non-adherence to basic principles of informed consent. One might wish to speculate on these reasons, but a survey of the website owners would allow for deeper insight.

In addition, the broader context could also be investigated. This study has focused on national medical associations. What are the policies of smaller, regional medical associations?

CONCLUSION

This study has examined the websites of national medical associations across the world, with a view to answering some questions about their user data gathering and sharing procedures. The study has found that, for the most part, national medical associations' practices fall short of standard requirements in medical ethics and medical informatics ethics codes. User data are routinely gathered and shared with third parties without informed consent or any reasonable possibility of opting out. There is little control over the retention of data beyond reasonable periods of time; there is an ignorance of the fact that anonymous and pseudonymous can be re-organised to identify users.

Based on these results, this paper proposes steps to be taken by national medical associations, so that their data gathering and sharing procedures may be in line with principles of medical ethics and medical informatics ethics.

References

1. Masters K: Health Informatics Ethics (5th Ed.). Hoyt B, Yoshihashi A, Bailey N, editors. Lulu.com, 2012. p 195-215.
2. Nuernberg Military Tribunals: Trials of war criminals before the Nuernberg Military Tribunals under Control Council Law No. 10, Volume I: "The Medical Case." Washington: US Government Printing Office, 1949.
3. Nuernberg Military Tribunals: Trials of war criminals before the Nuernberg Military Tribunals under Control Council Law No. 10, Volume II: "The Medical Case" and "The Milch Case." Washington: US Government Printing Office, 1949.
4. World Medical Association: World Medical Association Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects; 2008; http://www.wma.net/en/30publications/10policies/b3/index.html(Accessed 31/07/2012).
5. Association for Computing Machinery: Code of Ethics and Professional Conduct; 2011; http://www.acm.org/about/code-of-ethics> (Accessed 31/07/2012).
6. Canadian Information Processing Society: Code of Ethics

and Professional Conduct; 2007; http://www.cips.ca/?q=system/files/CIPS_COE_final_2007.pdf(Accessed 31/07/2012), 2007.
7. American Medical Informatics Association (AMIA): AMIA Code of Ethics; 2012; http://www.amia.org/about-amia/ethics(Accessed 31/07/2012).
8. UK Council for Health Informatics Professions: UKCHIP Code of Conduct; 2012; http://www.ukchip.org/?q=page/UKCHIP-Code-Conduct(Accessed 31/07/2012).
9. European Parliament: Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995, Brussels, 1995.
10. European Parliament: Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006, Brussels, 2006.
11. European Parliament: Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009, Brussels, 2009.
12. European Commission: Article 29 Data Protection Working Party. Opinion 04/2012 on Cookie Consent Exemption. Brussels: European Commission, 2012.
13. De Lusignan S, Chan T, Theaom A, Dhoul N: The roles of policy and professionalism in the protection of processed clinical data: A literature review. Int. J. Med. Inf.; 2007; 76:261-8.
14. Agrawal R, Johnson C: Securing electronic health records without impeding the flow of information. Int. J. Med. Inf.; 2007; 76:471-9.
15. Haas S, Wohlgemuth S, Echizen I, Sonehara N, Müller G: Aspects of privacy for electronic health records. Int. J. Med. Inf.; 2011; 80:e26-31.
16. Kluge E-HW: Informed consent and the security of the electronic health record (EHR): Some policy considerations. Int. J. Med. Inf.; 2004; 73:229-34.
17. Sadan B: Patient data confidentiality and patient rights. Int. J. Med. Inf.; 2001; 62:41-9.
18. Kang C: Google Announces Privacy Changes Across Products; Users Can't Opt Out. The Washington Post; 2012; http://www.washingtonpost.com/business/economy/google-tracks-consumers-across-products-users-cant-opt-out/2012/01/24/gIQArgJHOQ_story.html(Accessed 26/05/2012) Washington.
19. Vaknin S: Five Ways Google's Unified Privacy Policy Affects You. C | Net; 2012; http://howto.cnet.com/8301-11310_39-57388626-285/five-ways-googles-unified-privacy-policy-affects-you/(Accessed 26/05/2012).
20. McCallister E, Grance T, Scarfone K: NSIT Special Publication 800-122: Guide to Protecting the Confidentiality of Personally Identifiable Information (PII). Gaithersburg, MD: [US] National Institute of Standards and Technology, US Department of Commerce, 2010.
21. Ghostery: About Google Analytics; 2012; http://www.ghostery.com/apps/google_analytics(Accessed 22/05/2012).
22. Google: Google Analytics: Safeguarding Your Data; 2012; http://www.google.com/intl/en/analytics/privacyoverview.html(Accessed 25/05/2012).
23. Ghostery: About Facebook [Social Plugins]; 2012; http://www.ghostery.com/apps/facebook_social_plugins(Accessed 22/05/2012).
24. Facebook: Data Use Policy; 2012; http://www.facebook.com/policy.php(Accessed 25/05/2012).
25. Ghostery: About Twitter [Badge]; 2012; http://www.ghostery.com/apps/twitter_badge (Accessed 22/05/2012).
26. Twitter: Twitter Privacy Policy; 2012; http://twitter.com/privacy (Accessed 25/05/2012).
27. Ghostery: About WebTrends; 2012; http://www.ghostery.com/apps/webtrends (Accessed 22/05/2012).
28. Webtrends: Privacy Policy; 2011; http://www.webtrends.com/AboutWebtrends/PrivacyPolicy.aspx (Accessed 25/05/2012).
29. Ghostery: About AddThis; 2012; http://www.ghostery.com/apps/addthis (Accessed 22/05/2012).
30. AddThis: Privacy & Data Practices; 2012; http://www.addthis.com/privacy (Accessed 25/05/2012).
31. Ghostery: About Facebook [Connect]; 2012; http://www.ghostery.com/apps/facebook_connect (Accessed 22/05/2012).
32. Ghostery: About Twitter [Button]; 2012; http://www.ghostery.com/apps/twitter_button (Accessed 22/05/2012).
33. The White House. Consumer Data Privacy in a networked world: a framework for protecting privacy and promoting innovation in the global digital economy. Washington, 2012.
34. [US] Federal Trade Commission: Protecting Consumer Privacy in an Era of Rapid Change. FTC, 2012.
35. Lekkas D, Gritzalis S, Katsikas S: Quality assured trusted third parties for deploying secure internet-based healthcare applications. Int. J. Med. Inf.; 2002; 65:79-96.
36. Narayanan A, Shmatikov V: De-anonymizing Social Networks. 30th IEEE Symposium on Security and Privacy. Oakland, California; 2009.
37. Ohm P: Broken promises of privacy: Responding to the surprising failure of anonymizataion. UCLA L. Rev.; 2010:1701-77.
38. Sweeney L: Computational Disclosure Control: A Primer on Data Privacy Protection [dissertation]. Massachusetts Institute of Technology, 2001.

Author Information

Ken Masters
Assistant Professor: Medical Informatics, Medical Education & Informatics Unit, College of Medicine & Health Sciences , Sultan Qaboos University