The Dark Side of Data Sharing: Considering the Ethicality of Data Brokering within the Context of Health

Bran Knowles
School of Computing and Communications
Lancaster University
Lancaster, UK
b.h.knowles1@lancaster.ac.uk

Roisin McNaney
School of Computing and Communications
Lancaster University
Lancaster, UK
r.mcnaney@lancaster.ac.uk

.

Abstract

In this paper we reflect on the ethical challenges that arise as a result of efforts to improve healthcare through analyses of large-scale datasets. We focus in particular on the importance of ensuring that the systems that manage sensitive data are both trustworthy and trusted, and whether/how this might reasonably (and ethically) be achieved. We consider the fragility of the concept of data ownership in the absence of trust—indeed, in the absence of trustworthiness—and the impact that this has on people’s perceptions of digital society. We reflect on case examples from our research; and the case of care.data and the public backlash that led to its closure in 2014. These cases highlight the fear and uncertainty that surround data sharing practices in general, and how this has caused challenges in the process of designing digital health systems.

Author Keywords

Data Sharing; Public Trust; Digital health; Consent; Ethical Practice

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Excerpt 1: On dossiers

“Steve Kroft: Are people putting this together and making dossiers?

Julie Brill: Absolutely.

Steve Kroft: With names attached to it? With personal identification?

Julie Brill: The dossiers are about individuals. That's the whole point of these dossiers. It is information that is individually identified to an individual or linked to an individual.

Steve Kroft: Do you think most people know this information is being collected?

Julie Brill: I think most people have no idea that it's being collected and sold and that it is personally identifiable about them, and that the information is in basically a profile of them” [11].

Introduction

In 2014 the English National Health Service (NHS), alongside the Health and Social Care Information Centre, launched the care.data program, a health data sharing service aiming “to securely bring together health and social care information from different healthcare settings, such as GP practices, hospitals and care homes, in order to see what’s working really well in the NHS—and what we could be doing better” [17].

Almost overnight, the program was driven into scandal, sensationalized by the press as a way for the NHS to “cash in on patient records” [5] by selling people’s personal data to pharmaceutical and insurance companies. What followed was a viral social media campaign resulting in 1,461,877 people [16] opting out of data sharing outside of their GP facilities. This meant that many patients were unable to access services which utilize personal data, such as automatic allocation for cancer screening based on case history and demographic information [19]. On a more systemic level, it caused nationwide challenges for the interoperability of NHS services (e.g. the flow of patient information from a GP practice to an A&E department), at the pinnacle of a newly envisioned ‘digital NHS’.

In this paper we consider provocative notions of data ownership [3], in a world where data collection and use practices are ‘hidden’ within legal terms and conditions [14]—where personal data can be bought and sold without our knowledge and then used to discriminate against us [11,18]—and what principles would need to underlie ethical, trusted models of ‘best practice’ around data sharing. We describe case studies from our own research that highlight how the blurring of boundaries around data use can have serious potential consequences for the adoption of technologies. As we hurtle into a fully digitized healthcare system, we call for considerations around these complex, multifaceted issues to become commonplace in HCI research focusing on health, lest we become unwittingly complicit in unethical practices.

Data: The Price we Pay

There is wisdom in the saying, “If you are getting something for free, you are the product” (paraphrased: [12]). The pervasiveness and ubiquity of the Internet means that it is easily mistaken as a `natural fact’, something that exists as part of the natural order of the world, rather than as a service being provided for free *for a reason*. As the CEO of the data broker company Epsilon said, “Consumers ought to understand that the Internet is an advertising medium” [11]. The profitability of the Internet is ensured through a shadowy industry that specializes in collating data on consumers so that they might be more effectively targeted to consume more (see excerpt 1). Under this model, we have tacitly consented to an arrangement where the price we pay for Internet services is the loss of control over both the production of data about ourselves and the sharing of that data [6]; i.e. it is the companies who collect the data that own that data (not the person the data is about). But the argument that this is an arrangement that one can enter into willingly is deeply flawed, considering that essential activities of daily life are mediated through Internet services that, in order to use them at all, users must agree upfront to these terms [23]. There is no choice here; not in the digital age; not yet, anyway.

Choice is a common theme within discussions of ethics—informed consent implies (requires!) choice, for example. What is less often discussed in the context of ethics, however, is *trust,* but especially *trustworthiness.* As the examples below are meant to illustrate, there are several key tenets of trustworthy systems which, when absent, impede what we might call informed consent. It is our position that making data sharing trustworthy (i.e. deserving of people's trust) in the context of digital health is essential for realizing the potential of big data, and that it is essential for making data sharing ethical; but also that doing so means actively attending to the broader context in which unethical data sharing practices have been permitted to emerge. We cannot simply ignore the dark side of data sharing, i.e. data brokering, because a) it affects public perceptions of the trustworthiness of systems, and b) data brokers are not obligated to respect individual's privacy when it comes to health data donated in the name of science any more than any other type of personal data.

Fear and Mistrust in Data Sharing Practices: Case study examples

If you can't trust the council, then who?
Plans to realize the vision of a fully digitized society are ostensibly problematized by the 'older adult' demographic who are generally more resistant to digital technologies, in part due to fear of risks. As part of a series of focus groups to better understand older adults' resistance to digital technologies, questionable practices around data sharing emerged as a common contributor amongst participants.

Almost proving the point that the Internet is an advertising medium, participants spoke with passionate annoyance about the number of phone calls they received from companies trying to sell them things based on data they had bought from trusted Internet service providers. What would seem an especially egregious example of a betrayal of trust (if proved true) was one participant's assertion that data leakage was originating from their local authority. It is worth noting that this betrayal occurs in the context of a massive push for online-only government services, which have forced many digital holdouts to acquiesce to using online government services [1]. Responding to this accusation, another participant asked the interviewer, "Don't most organizations do that now? I thought that was quite common." (It is more than quite common; it is the dominant business model [2]!) When asked whether they would stop using the council's online services if they knew they were sharing their data, the response was, "I suppose so. But I don't know how you go about doing it. That's the point."

This exchange highlights several missing ingredients in a recipe for trustworthy data sharing. Firstly, it is quite possible that the council is wrongly accused here; but there is no *visibility* [7] of the council's data sharing practices—and no *traceability* of where that data ultimately goes when you send it to the council. In lieu of visibility and/or traceability, these practices have to be actively deduced, which can easily lead to mistrust even when agents are worthy of trust. Secondly, there is no *accountability* [7]: if the council knew that you could see what they were doing with their data, they would experience a certain amount of pressure to act in an ethically responsible way that would preclude selling your data to companies who wanted to target consumers. Thirdly, there is no plausible *recourse* for breaches of trust. Accountability requires some leverage to punish bad behaviour, and that leverage

Excerpt 2: What is sold about you

"Steve Kroft: You can buy from any number of data brokers, by malady, the lists of individuals in America who are afflicted with a particular disease or condition… Alcoholism?

Tim Sparapani: Yes. Absolutely.

Steve Kroft: Depression?

Tim Sparapani: Certainly.

Steve Kroft: Psychiatric problems?

Tim Sparapani: No question.

Steve Kroft: History of genetic problems?

Tim Sparapani: Yes. Cancer, heart disease, you name it, down to the most rare and, and most unexpected maladies.

Steve Kroft: Sexual orientation?

Tim Sparapani: Of course" [11].

would normally be a choice to use/not use a service. But in the words of one participant, "They've got you."

Maintaining Data Privacy Across Agencies

This mistrust around data sharing practices has a direct consequence to the ways that we, as HCI researchers, think about how to design technologies that have the potential to collect and monitor data about people's health. With the emergence of 'hot topics' in HCI around big data, and the quantified self movement which sees us collecting data about every element of our daily lives, there are exciting opportunities to change the way we understand human living. However, we must remain mindful of the ways in which data collected for the wider good, with the intent of helping, might be used maliciously.

As previously noted, our participants are considering these issues and they are concerned over how their data might be abused. In a second study exploring the design of rehabilitative exergames with People with Parkinson's [15], participants raised explicit concerns around how data about their movements, symptoms and potential improvements in the pathway of their physical rehabilitation might be used against them: "Your identity would need to be protected because there would be concerns about that information becoming available to the benefits agencies."

For participants who were below retirement age, there was a fear that the government, with little regard for the fluctuating nature of Parkinson's, might consider them fit enough to return to work. Others feared that they might have their supported physical therapy removed if they were seen to be progressing with the game.

What this example demonstrates most perspicuously is the ethical problems that arise from a lack of *intelligibility* [13]. How individuals were assessed was left to the imagination, rather than made clear through the system design. Lacking essential information about how data is being used precludes individuals acting in ways that protect themselves from harm. And this, too, was precisely the flaw with care.data: "We cannot have patients not understanding what they are opting out of and how that will affect them." [4]

The frightening truth about re-identification

The Caldicott Review [4], which ultimately led to the closure of the care.data program, stressed the need for effective measures for ensuring that anonymised data remains anonymous, i.e. that individuals could not be re-identified. The parliamentary announcement on the matter stated, "I can confirm today that the Government is supportive of the introduction of stronger criminal sanctions against those who use anonymised data to re-identify individuals" [8].

Whether or not re-identification of individuals is made criminal, the fact is that it is surprisingly easy to do; hence, legal action is unenforceable. Re-identification occurs when shared identifiers (e.g. ZIP code, date of birth, gender) are used to link anonymous datasets with non-anonymous ones [20]. As one example, [22] (from 1998) found that one could use the ZIP code, birth date and gender of individuals from a voter registration list to gain access to information on patients' diagnoses, medical procedures and medications through hospital discharge data.

The threat of re-identification is even more prevalent as the enticements for self-disclosure multiply and technologies such as facial recognition improve. In a study by Gross (2005) [9], the author found that 61% of profile images used on a university website are of sufficient quality to make it possible to re-identify an individual on a de-identified site (e.g. a dating site) using facial recognition software. Gross & Acquisti (2005) also found [10] that the data that is freely self-disclosed on social networking sites, such as hometown and birthday, can be used in combination with "social engineering" to make an accurate prediction of the last four digits of a person's social security number, creating a "substantial" risk for identity theft.

What this example shows is that there is no *predictability* in the context of data sharing: we have no way of knowing what new data sets might be used in what new combinations to compromise identities; or indeed what technologies might emerge to enable extraction of compromising information. Data brokers trade on their ability to make these kinds of connections between previously disconnected datasets, to reveal valuable information about particular individuals that the individual intended to be de-identified (see excerpt 2).

Clearly, the fact that re-identified data is used to create dossiers that enable discrimination against the targeted individual is not ethical. What is important to bear in mind, then, is that it is not enough for those handling the data to act "ethically"; they must also enact security protocol that ensures that even de-identified data cannot be hacked or leaked and then linked with publicly available identity disclosures. And for the HCI practitioner, this illustrates the need for continual evolution in response to emerging (i.e. not anticipatable) threats to individuals whose data is used.

Excerpt 3: Care.data and "trust"
From the Caldicott Review [4]:

This report focuses particularly on two aspects of people's trust. Firstly, it looks at whether data security is good enough. Are there adequate systems in place to prevent people's confidential information falling into the wrong hands? Can those systems be made strong enough to protect against known and potential dangers without being so restrictive that information cannot be shared appropriately among staff providing care? Secondly, the report looks at the basis upon which information is shared. Do people understand who will have legitimate access to their personal confidential data? When is the individual's specific consent required? When can people consent to or opt out from information being used and when may this be overruled? Are the current arrangements protecting people's confidentiality adequately upheld, and do they allow for appropriate information sharing to benefit patients, service users and the entire health and care system?

Conclusions

One of the Caldicott Review's [4] key findings (see excerpt 3) was that insufficient *privacy* made care.data untrustworthy, and hence ethical. Undoubtedly, one of the essential features of trustworthy systems is data privacy. Here, we have tried to expand the concept of trustworthiness to other, less often discussed elements of trustworthy systems to better understand how to deliver on the promises of big data in an ethical way. Our intention is not to imply that ethical data sharing in the health context cannot be done. Rather, when so much discussion around so called "data for good" [21] (in health and other contexts) focuses on questions of 'how to get people to share their data', we have tried to show that a prerequisite of gaining public trust is ensuring that systems that share data are *deserving* of trust. This means, among other things, designing with attention to promoting visibility, traceability, accountability, and intelligibility; and with an awareness of how the lacking of these features in other intertwining systems affects the degree to which any given system may be deemed "ethical."

References

1. Asthana, A. & McVeigh, T. (2010). Government servces to be online only. The Guardian [online].20 November 2010. Available: https://www.theguardian.com/society/2010/nov/20/government-services-online-only
2. Boutin, P. (2016). The secretive world of selling data about you. Newsweek Europe [online]. 30 May 2016. Available: http://tinyurl.com/htpk5tq
3. Brown, B., Weilenmann, A., McMillan, D., & Lampinen, A. (2016, May). Five provocations for ethical hci research. In Proc. of the CHI'16 (pp. 852-863). ACM.

4. Caldicott, F. (2016). National Data Guardian for Health and Care: Review of Data Security, Consent and Opt-Outs. June 2016. Available: http://tinyurl.com/gqcndef

5. Chapman, J. & Dolan, A. (2014). Daily Mail [online]. Cashing in on patient records to be banned: But you'll still have to opt out to keep private details off database. 1 March 2014. Available: http://tinyurl.com/gqxgnv7

6. Ehrenberg, B. (2014). The Guardian [online]. How much is your personal data worth? 22 April 2014. Available: http://tinyurl.com/gn7z6f2

7. Erickson, T., & Kellogg, W. A. (2000). Social translucence: an approach to designing systems that support social processes. ACM transactions on computer-human interaction (TOCHI), 7(1), 59-83.

8. Freeman, G. (2016). The Care Quality Commission and National Data Guardian for Health and Care's Independent Reviews into Data Security, Consent and Opt-Outs:Written statement - HCWS62 [online]. 6 July 2016. Available: http://tinyurl.com/zssskq3

9. Gross, R. (2004). Re-identifying facial images. Technical report, Carnegie Mellon University, Institute for Software Research International.

10. Gross. R., & Acquisti, A. (2005). Information revelation and privacy in online social networks. In Proceedings of the 2005 ACM workshop on Privacy in the electronic society (WPES '05). ACM, New York, NY, USA, 71-80.

11. Kroft, S. (2014). The Data Brokers: Selling your personal information. CBS News [online]. 24 August 2014. Available: http://tinyurl.com/kgautlc

12. Lanier. J. (2013). Who Owns the Future? Simon & Schuster: NY

13. Lim, B. Y. (2010). Improving trust in context-aware applications with intelligibility. In Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing-Adjunct (pp. 477-480). ACM.

14. Luger, E., & Rodden, T. (2013). Terms of agreement: Rethinking consent for pervasive computing. Interacting with Computers, iws017

15. McNaney, R. et al. (2015). Designing for and with People with Parkinson's: A Focus on Exergaming. In Pro CHI '15. ACM (p.501-510).

16. NHS Digital. (2016) Care Information Choices. May 2016. Available: http://tinyurl.com/ju7rlxf

17. NHS England. (2016). The care.data programme. Available: http://tinyurl.com/jalgcva

18. O'Neil, Cathy. "Weapons of Math Destruction." How Big Data Increases Inequality and Threatens Democracy, New York: Crown (2016).

19. Ramesh, R. (2015). NHS disregards patient's requests to opt out of sharing medical records. The Guardian [online]. 22 January 2015. Available: http://tinyurl.com/h25f2c2

20. Samarati, P. and Sweeney, L. (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and cell suppression. Technical report, SRI International, 1998.

21. Skatova, A. & Goulding, J. (2015). Donate your data – how your digital footprint can be used for the public good. The Conversation [online]. Available: http://tinyurl.com/jbm35cf

22. Sweeney, L.(2002). k-Anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570.

23. Tolmie, P., Crabtree, A., Rodden, T., Colley, J., & Luger, E. (2016). "This has to be the cats": Personal Data Legibility in Networked Sensing Systems. In Proc of CSCW'16 (pp. 491-502). ACM.