Sharing of Big Data in Healthcare: Public Opinion, Trust, and Privacy Considerations for Health Informatics Researchers

Laura Moss[1,2], Martin Shaw[2], Ian Piper[2], Christopher Hawthorne[3] and John Kinsella[1]

[1] *Department of Anaesthesia, Pain & Critical Care, School of Medicine, University of Glasgow, Glasgow, U.K.*

[2] *Department of Clinical Physics, NHS Greater Glasgow & Clyde, Glasgow, U.K.*

[3] *Department of Neuroanaesthesia, NHS Greater Glasgow & Clyde, Glasgow, U.K.*

Keywords: Big Data, Privacy, Critical Care, Data Analysis, Trust, Data Security.

Abstract: Advances in technology has transformed clinical medicine; electronic patient records routinely store clinical notes, internet-enabled mobile apps support self-management of chronic diseases, point-of-care testing enables laboratory tests to be performed outside of hospital environments, patient treatment can be delivered over wide geographic areas and wireless sensor networks are able to collect and send physiological data. Increasingly, this technology leads to the development of large databases of sensitive electronic patient information. There is public interest into the secondary use of this data; many concerns are voiced about the involvement of private companies and the security and privacy of this data, but at the same time, these databases present a valuable source of clinical information which can drive health informatics and clinical research, leading to improved patient treatment. In this position paper, we argue that for health informatics projects to be successful, public concerns over the secondary use of patient data need to be addressed in the design and implementation of the technology and conduct of the research project.

1 INTRODUCTION

Healthcare is rapidly changing and advanced technology is enabling the collection of vast amounts of patient data. Consequently healthcare is experiencing a Big Data phenomenon. Whilst health informatics research is focused on the development of novel approaches to enable the intelligent analysis of this data, the use of electronic patient data to advance these approaches and the implementation of these technologies within real world healthcare environments raises many ethical questions and divides public opinion.

In this paper we explore some of the issues regarding the use of electronic patient data for secondary purposes, specifically trust, security and patient confidentiality. The paper is organised as follows: section 2 describes the prominence of Big Data within healthcare and challenges faced; section 3 discusses issues raised in the sharing of data, in particular the involvement of private companies and views of the general public; section 4 illustrates real-world issues faced whilst implementing a big data analysis platform into a healthcare environment; section 5 discusses considerations and opportunities for health informatics researchers; finally, section 6 concludes the discussion.

2 BACKGROUND

As technology advances, data is increasingly collected through a variety of mechanisms. It is estimated that 2.5 quintillion bytes of data is currently generated each day (IBM a, 2016). Big Data is a term which is now commonly used to describe such large and complex datasets. Advanced analytics are often applied to Big Data to extract meaning, insights and discovery of new knowledge. However, using these large datasets presents challenges due to the data's volume (many sources e.g. sensors, social networking), variety (many formats, e.g. videos, text) and velocity (speed at which data is produced and requirements for near-real time processing). Other attributes can include veracity (noisy, messy data), variability (meaning of the data can be constantly shifting) and fine-grained (Kitchen and McArdle, 2016).

Within healthcare, the increasing use of technology is creating large volumes of clinical data; in 2011, the global size of healthcare data was estimated to be 161 billion gigabytes (IBM b, 2016). This influx of data is from a variety of technical advances: e.g. enhanced clinical imaging, electronic medical records, and physiological sensors. Additionally, the growth

Moss L., Shaw M., Piper I., Hawthorne C. and Kinsella J.
Sharing of Big Data in Healthcare: Public Opinion, Trust, and Privacy Considerations for Health Informatics Researchers.
DOI: 10.5220/0006251504630468
In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*, pages 463-468
ISBN: 978-989-758-213-4

of wearable sensor devices (e.g. smart watches) is leading to patients themselves being able to generate large amounts of healthcare data.

The potential clinical benefit of using this data is well established. Analysis of healthcare data can drive improvement in many areas, including: clinical and organisation processes, optimisation of treatments, and reduction of healthcare costs through intervention at an earlier point and more proactive and targeted care. Patient data can be used to predict clinical risk, targeting resources where they are needed most and identifying problems that would benefit from early intervention. Technologies that aid clinical decision-making and help clinicians to manage the exponential growth in medical knowledge offer substantial opportunities to reduce variation, improve the quality of care, and possibly lower costs (Jaspers et al., 2011), (Fillmore et al., 2013).

Whilst a thriving health informatics community is focused on the development of novel approaches and tools to enable the intelligent and sophisticated analysis and use of healthcare data, the transfer of these technologies and ideas to real world clinical environments faces many challenges. Healthcare providers often do not have access themselves to data science expertise, required technology infrastructure and funding to do this themselves. In many cases, to enable the potential for big data analysis to improve healthcare, partnerships have to be formed between healthcare providers, research organisations and private companies. High profile examples include DeepMind and IBM Watson technology. IBM's Watson is being used within oncology in US and Canadian hospitals to assess tumours (IBM c, 2016) and DeepMind is using patient data from UK hospitals to develop diagnostic tools in areas such as acute kidney injury (DeepMind a, 2016) and ophthalmology (DeepMind b, 2016).

3 TRUST, SECURITY AND PRIVACY

Trust and confidentiality between a clinical provider and patient is not new: it is central to the practice of healthcare and has been focused on since Hippocrates. Whilst the concept of patient confidentiality has endured as an ideal throughout history, the precise nature of it has changed with the sociohistoric context (for a detailed review see (A.Ferguson, 2012)). In the digital age, patient confidentiality is often framed within the context of electronic patient records and the potential involvement of third parties. Whilst the involvement of private organisations and other research organisations can resolve many practical issues for healthcare providers, it often involves the transfer of sensitive patient data to these organisations for research development and this can raise many questions. For example:

- Should electronic patient data be used for secondary research purposes?
- Who does the electronic patient data belong to?
- Who should be able to use the electronic patient data? Public sector healthcare organisations (e.g. NHS) and/or commercial companies?
- Who is collecting the data and where is it stored?
- What safeguards should be put in place to protect patient confidentiality?
- How do patients and relatives feel about the collection and use of this data?

Most countries have some form of regulation or processes which have been legally put in place to resolve some of the above issues. Additionally, institutions often have internal procedures and practices in place to protect patient data. However, public perception of the adequacy of these frameworks can be divided and concerns raised about whether private companies, in particular, can be trusted with patient data. For example, the public sector can be perceived as more trustworthy than profit-making organisations when using patient data (Focus Groups et al., 2013). Recent publicity in cases such as DeepMind's arrangement with the Royal Free Hospital, London, has been controversial and highlighted the requirement for more robust safeguards to be put in place to ensure patient data is adequately secured (New Scientist, 2016). Additionally, there are constant reports in the media about data leaks and unsuccessful, large-scale, health I.T projects (Presser et al., 2015).

It also raises issues about whether patient consent and public awareness of healthcare data sharing is adequate. A number of studies have identified that there is low awareness by the general public of electronic patient record systems and how and why healthcare data might be used (BMA, 2016) (Riordan et al., 2015). Riordan et al (Riordan et al., 2015) found that most people would prefer to opt-in before their identifiable records were used and half of participants would share their de-identified records under implicit consent. A recent consultation with the general public in the UK on this issue identified that they had little confidence in the safeguards put in place to protect data. Additionally, it was felt that there was a lack of accountability within the system, and malicious use of data by private companies (e.g. pharmaceutical and insurance) was a concern for many people (BMA, 2016). Lessons learnt from the failure of large scale patient data sharing projects show the requirements for clear communication to the public, easy to understand consent rules and strong oversight and communication regarding distribution and use of patient data (Presser et al., 2015). For a more detailed systematic review of the literature regarding public responses to sharing of health data see (Aitken et al., 2016b).

To investigate the opinions of the general public further, in a small study we asked the audience at a science festival their thoughts on the topic of secondary use of patient data (Kinsella et al., 2017). Questions covered included: whether the participants were aware of the potential of using their medical data for secondary research purposes; whether patient data should be used for research purposes and how likely would they be themselves to share their own personal data for research; if they trust clinicians with their data; and their opinions on the role of private/commercial companies in supporting and/or carrying out research on their own medical data. 39 out of 41 adults responded to the survey (from which we have full results for 37 adults). Table 1 shows results from the survey. The vast majority of respondents felt that their medical data should be used for research purposes and would be happy to share their data. This is in keeping with a number of other studies (as summarised in (Aitken et al., 2016b). Additionally, most respondents trusted clinicians, but when it came to private companies, the response was mixed. This difference in perceived trustworthiness between clinicians and private companies has been found previously (e.g.(Aitken et al., 2016a),(BMA, 2016)).

We also looked for any age divide in the participant's responses. We asked which age category participants fell into (21 and under, 22-34, 35-44, 45-54, 55-64, 65 and over) and three questions regarding the involvement of private companies: 1) Would you trust a private company to do research with your medical data? 2) Would you trust a private company to do research with anonymous medical data? 3) Would you be comfortable with a private company providing the support to medical researchers to enable them to do medical research? Tables 2 and 3 display the results of this analysis. In both age groups, more people had positive responses (agreed or strongly agreed) than negative ones (disagree or strongly disagree). Although younger people are often thought to be more confident with technology and data sharing, no sizable difference between the two age groups was found in this study. A larger sample size would be required to show statistical significance.

Public opinion and perception of the use of healthcare data can be divided and it is clear from these studies and the opinions of other researchers (e.g. (van Staa et al., 2016)), that for health informatics projects involving transfer and analysis of patient data by third parties to be a success, the trust of the general public needs to be earned and respected by all involved.

4 CHART-ADAPT CASE STUDY

To illustrate some of the issues which may need to be considered in collaborative health informatics research projects, in this section we discuss some of the actions which the CHART-ADAPT project instigated to try and overcome data sharing concerns (CHART-ADAPT, 2016).

The CHART-ADAPT platform allows the fast analysis of high and low frequency data collected from a critical care unit; enabling the creation and assessment of novel, closed loop, diagnostic or therapeutic models and algorithms. Routinely recorded patient data is automatically transferred from the electronic patient record system in the critical care unit to a high performance computing platform implementing a Spark (Apache a, 2016), Scala (Scala, 2016) and Hadoop (Apache b, 2016) technology stack. Complex physiological algorithms are then applied to the data to derive clinically useful variables which are returned back into the clinical environment and integrated with the existing electronic patient record system.

A lack of the required technical infrastructure within the hospital to process the patient data within clinically meaningful timescales meant it was essential for the healthcare provider and academic researchers to form a collaborative team which included commercial partners to provide the required high performance computing infrastructure.

Due to the nature of the collaboration and the requirement to transfer patient data, the project team was aware of the need to maintain patient confidentiality, the governance of patient data, and the need to gain the confidence of patients and unit staff. Several concerns were identified: 1) failure of the de-identification software and subsequent transfer of identifiable patient data outside the healthcare provider's network, 2) secure handling of the anonymised patient data by the commercial partner, 3) correct re-identification of patient data when it reentered the healthcare provider's network, and 4) public perception of a commercial partner supporting the patient data analysis. These concerns were considered from the start of the project and all the collaborating organisations worked together to integrate the following activities into the project plan:

Table 1: General Public Opinions on Secondary Healthcare Use.

Question	**Yes**	**No**	
Are you aware that medical data could be used for research?	30 (81.2%)	7 (18.9%)	
Question	**Strongly agree/agree**	**Undecided**	**Disagree/Strongly disagree**
Medical data should be used for research	33 (89.2%)	4 (10.8%)	0
Would you be happy to share your healthcare data?	31 (83.8%)	4 (10.8%)	2 (5.4%)
Do you trust clinicians with your healthcare data?	27 (73%)	7 (18.9%)	3 (8.1%)
Do you trust private companies to use your medical data for research purposes?	8 (21.6%)	20 (54%)	9 (24.3%)

Table 2: Responses < 35 years old, Total = 21 respondents.

Question	**Strongly Agree**	**Agree**	**Undecided**	**Disagree**	**Strongly Disagree**
1	1 (4.8%)	7 (33.3%)	10 (47.6%)	2 (9.5%)	1 (4.8%)
2	4 (19%)	8 (38.1%)	7 (33.3%)	1 (4.8%)	1 (4.8%)
3	5 (23.8%)	10 (47.6%)	5 (23.8%)	0 (0%)	1 (4.8%)

Table 3: Responses >= 35 years old, Total = 16 respondents

Question	**Strongly Agree**	**Agree**	**Undecided**	**Disagree**	**Strongly Disagree**
1	1 (6.25%)	3 (18.8%)	8 (50%)	2 (12.5%)	2 (12.5%)
2	2 (12.5%)	7 (43.8%)	5 (31.3%)	2 (12.5%)	0 (0%)
3	4 (25%)	5 (31.3%)	6 (37.5%)	1 (6.25%)	0 (0%)

- Regulatory approval (beyond minimum requirements) was acquired for the transfer of patient data (e.g. NHS Research Ethics, Caldicott Guardian approval)
- The project developed software to automatically anonymise the patient data before it left the healthcare environment. A rigorous testing plan was followed and repeated at regular intervals to ensure confidentiality was maintained.
- The commercial partner responsible for technical support of the data analysis (Aridhia) developed an Information Governance Strategy for the project which made explicit the data handling and security procedures put in place. Close communication was also maintained between the personnel responsible for Information Governance in both organisations (healthcare and commercial).
- Public engagement initiatives were implemented. For example, posters and leaflets were made available in the unit, staff were briefed and updated on project progress, and a public event was hosted to discuss patient data sharing within critical care.
- Attendance at relevant academic and healthcare events was scheduled into the project. This gave the team the opportunity to discuss the platform and gather feedback which was fed into the development of the project.

Although some activities were time-consuming and beyond the usual scope of a research project, it was beneficial not only for development of the platform, but also to make sure, to aid future acceptance of the technology, that we took the clinical staff and general public with the project, rather than exclude them and present the technology as a fait accompli.

5 OPPORTUNITIES FOR HEALTH INFORMATICS RESEARCH

Whilst public opinion on trust, security and privacy of patient data needs be carefully considered in research projects, there is also an opportunity for the health informatics community to develop tools and technologies to address these concerns. Below are some suggestions (although this is not exclusive) and comments on how the health informatics community may be able to contribute:

- **Communication of Health Informatics Projects** - There is a need to develop clear, concise, up-todate summaries of health informatics projects to aid transparency, in particular regarding the use of patient data. Researchers should consider how they will engage the public when designing and implementing the health informatics project.

- **Dynamic Consent** - Current mechanisms of informed consent for patient data sharing are static, paper-based and designed around legal frameworks and regulation. They are also specific to individual research studies and have to be repeated for subsequent studies. There is a growing awareness that this is inadequate and future policies are moving towards a more sophisticated form of consent (e.g. the proposed EU General Data Protection Regulation (GDPR, 2016)). Dynamic consent provides patients with a personalised interface, placing them in control of how their healthcare data is used; data sharing preferences can be stated and often they can view how their data is being used by researchers (Kaye et al., 2015), (Williams et al., 2015). Once consent has been specified by patients, new tools and technologies are required which enable their preferences to be dynamically and automatically applied across multiple clinical databases and research studies.

- **Safe Havens** - To control how electronic patient data is used by researchers, many healthcare providers are making it accessible through Safe Havens (i.e. it doesn't leave an authorised environment) (Caldicott, 2016). Safe Havens pull together data from multiple healthcare sources and links made between the datasets whilst maintaining patient confidentiality. Safe Havens require a suite of software tools to: ensure security of the centrally stored data (e.g. defend against cyber attacks), enforce data access controls, and audit the use of the patient data. Whilst basic tools have been implemented, there is still potential for more sophisticated software to support these activities.

- **De-identification of Patient Data** - Generally, there is public support for the sharing of de-identified data for research purposes. National and international guidelines specify methods for de-identification and can include the removal or generalisation of certain attributes. Experts can also be asked to identify attributes with an associated risk leading to patient identification. As removal of data can lead to a lack of quality of the dataset overall, there is a balance to be struck between usability and patient confidentiality. This is a non-trivial optimisation problem which computing and artificial intelligence fields are well placed to contribute towards workable solutions.

- **Re-identification of Patient Data** - Even when patient data has been de-identified, there is still a possibility that it can be re-identified through the use of other, publicly available, datasets. This is likely to be a growing concern, especially with initiatives to make more data available and machine readable (e.g. Semantic Web). Some solutions to reduce the chances of this happening include: removal of high risk variables from a dataset (e.g. features which are available in multiple documents and publicly available); and generalisation of patient data into 'bins' of data (e.g. values are generalised over 5 patients). Again, computing and artificial intelligence fields are well placed to develop tools which enable the automatic identification of high risk attributes.

6 CONCLUSION

The health informatics community has an important role to play in the development of novel technology and algorithms to enable advances in clinical knowledge and the quality of patient care. This type of research requires access to sufficient volumes of patient data which raises important issues by the general public regarding ethics, trust and security of patient data, especially if private companies are involved in the research activities. Our position is that, despite these concerns, it is necessary for private companies, research institutions and healthcare providers to work together to successfully transition technology projects from research to real-world environments. However, it is vital that patient confidentiality is maintained during all stages of development. There is a role for policy makers to ensure that existing legislation and procedures are adequate for a fast moving technology industry and that there is clear accountability. Additionally, there needs to be greater public engagement on health informatics projects and open communication regarding the potential use of their data.

ACKNOWLEDGEMENTS

The CHART-ADAPT project is a collaboration between the University of Glasgow, Aridhia, NHS Greater Glasgow & Clyde, and Philips Healthcare. It has been co-funded by Innovate UK (ref:102113). Approval for the CHART-ADAPT work was granted by the West of Scotland Research Ethics Committee (REC ref: 15/WS/0222) and local Caldicott Guardian approval has been provided. Approval for the survey (section 4) was provided by MVLS, University of

Glasgow. The CHART-ADAPT project we would like to acknowledge the staff and patients of the Neurointensive care unit, Neurosciences Institute, Glasgow.

REFERENCES

A.Ferguson (2012). The evolution of confidentiality in the united kingdom and the west. *AMA Journal of Ethics*, 14(9):738–742.

Aitken, M., Cunningham-Burley, S., and Pagliari, C. (2016a). Moving from trust to trustworthiness: Experiences of public engagement in the scottish health informatics programme. *Science and Public Policy*, 43(5):713–723.

Aitken, M., de St Jorre, J., Pagliari, C., Jepson, R., and Cunningham-Burley, S. (2016b). Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Medical Ethics*, 17(73).

Apache a (2016). Apache Spark. https://spark.apache.org/. Accessed: Nov 2016.

Apache b (2016). Apache Hadoop. https://hadoop.apache.org/. Accessed: Nov 2016.

BMA (2016). Secondary Uses of Data, Public Workshop. https://www.bma.org.uk/collective-voice/policy-and-research/ethics/secondary-uses-of-data/public-workshop. Accessed: Nov 2016.

Caldicott (2016). Information: To share or not to share? The Information Governance Review. https://www.gov.uk/government/uploads/system/uploads /attachment_data/file/192572/2900774 _InfoGovernance_accv2.pdf. Accessed: Nov 2016.

CHART-ADAPT (2016). CHART-ADAPT. http://www.chartadapt.org. Accessed: Nov 2016.

DeepMind a (2016). DeepMind Acute Kidney Injury. Royal Free London. Google DeepMind: Q&A. https://www.royalfree.nhs.uk/news-media/news/google-deepmind-qa/. Accessed: Nov 2016.

DeepMind b (2016). DeepMind Moorfields Eye Hospital. Moorfields announces research partnership. http://www.moorfields.nhs.uk/news/moorfields-announces-research-partnership. Accessed: Nov 2016.

Fillmore, C., Braye, C., and Kawamoto, K. (2013). Systematic review of clinical decision support interventions with potential for inpatient cost reduction. *BMC Med Inform Decis Mak*, 13(135).

Focus Groups, Stevenson, F., Lloyd, N., Harrington, L., and Wallace, P. (2013). Use of electronic patient records for research: views of patients and staff in general practice. *Fam Pract*, 30(2):227–23.

GDPR (2016). GDPR: Regulation (EU) 2016/679. http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf. Accessed: Nov 2016.

IBM a (2016). IBM Big Data. Extracting business value from the 4 V's of big data. http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data. Accessed: Nov 2016.

IBM b (2016). The 4 V's of big data. http://www.ibmbigdatahub.com/infographic/four-vs-big-data. Accessed: Nov 2016.

IBM c (2016). IBM's Watson supercomputer to speed up cancer care. http://www.bbc.co.uk/news/technology-32607688. Accessed: Nov 2016.

Jaspers, M., Smeulers, M., Vermeulen, H., and Peute, L. (2011). Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *J Am Med Inform Assoc*, 18(3):327–34.

Kaye, J., Whitley, E., Lund, D., Morrison, M., Teare, H., and Melham, K. (2015). Dynamic consent: a patient interface for twenty-first century research networks. *Eur J Hum Genet*, 23(2):141–6.

Kinsella, J., Hawthorne, C., Shaw, M., Piper, I., Healthcare, P., Aridhia, and L.Moss (2017). Public perception of the collection and use of critical care patient data beyond treatment: a pilot study. In *Proceedings of the Society of Critical Care Medicine Congress (SCCM)*. SCCM.

Kitchen, R. and McArdle, G. (2016). What makes big data, big data? exploring the ontological characteristics of 26 datasets. *Big Data & Society*, Jan-June 2016(3):1–10.

New Scientist (2016). Revealed: Google AI has access to huge haul of NHS patient data. New Scientist 2016 Apr 29. https://www.newscientist.com/article/\2086454-revealed-google-\ai-has-access-to-\huge-haul-of-nhs-patient-data/. Accessed: Nov 2016.

Presser, L., Hruskova, M., Rowbottom, H., and Kancir, J. (2015). Care.data and access to uk health records: patient privacy and public trust. *Technology Science*, 2015081103(Aug 11).

Riordan, F., Papoutsi, C., Reed, J., Marston, C., Bell, D., and Majeed, A. (2015). Patient and public attitudes towards informed consent models and levels of awareness of electronic health records in the uk. *Int J Med Inform*, 84(4):237–347.

Scala (2016). Scala Programming Language. http://www.scala-lang.org/. Accessed: Nov 2016.

van Staa, T.-P., Goldacre, B., Buchan, I., and Smeeth, L. (2016). Big health data: the need to earn public trust. *BMJ*, 354:i3636.

Williams, H., Spencer, K., Sanders, K., Lund, D., Whitley, E., Kaye, J., and Dixon, W. (2015). Dynamic consent: A possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. *JMIR Med Inform*, 3(1).