International Journal of Law and Psychiatry 60 (2018) 40–44

Contents lists available at ScienceDirect

International Journal of Law and Psychiatry

Privacy in Big Data psychiatric and behavioural research: A multiple-case study

M. Mostert *, B.M. Koomen, J.J.M. van Delden, A.L. Bredenoord

Department of Medical Humanities, Julius Center for Health and Primary Care, University Medical Center, Utrecht, the Netherlands

ARTICLE INFO

Article history:
Received 31 January 2018
Received in revised form 3 July 2018
Accepted 6 July 2018
Available online xxxx

Keywords:
Big Data
Psychiatry
Research
Privacy
Qualitative research

ABSTRACT

In Big Data health research, concerns have risen about privacy and data protection. While the ethical and legal discussion about these issues is ongoing, so is research practice. The aim of this qualitative case study is to gain more insight into how these concerns are currently dealt with in practice. For this multiple-case study, the YOUth cohort, a longitudinal cohort focusing on psychosocial development, and Big Data Psychiatry, a pilot study in Big Data analytics on psychiatric health data, were selected. A broad range of relevant documents were collected and semi-structured interviews with stakeholders were conducted. Data were coded, studied and divided into themes during an iterative analytical process. Three themes emerged: abandoning anonymisation, reconfiguring participant control, and the search for guidance and expertise. Overall, the findings show that it takes considerable effort to take privacy and data protection norms into account in a Big Data health research initiative, especially when individual participant level data need to be linked or enriched. By embracing the complexity of the law in an early phase, setbacks could be prevented, the existing flexibility within the law could be utilised, and systems or organisations could be designed and constructed to take relevant rules into account. Our paper illustrates that a close collaboration of experts with different backgrounds within the initiative may be necessary to be able to successfully navigate this process.

1. Introduction

Big Data is finding its way into health research. Some believe that this will provide unprecedented opportunities for psychiatry (Monteith et al., 2015). A broad range of issues, however, need to be dealt with. One of the key areas of concern in Big Data health research is related to privacy and data protection (Mittelstadt and Floridi, 2016), especially when psychiatric or other sensitive health-related data are collected, re-used, linked and analysed.

The rise of such data-intensive health research initiatives has sparked a lively debate about how the use of data should be governed by principles and rules, especially during the adoption of the General Data Protection Regulation (GDPR) in the EU (Mostert et al., 2016; Ploem et al., 2013; Sethi, 2015). Although this debate on normative issues is ongoing, researchers and other stakeholders already need to deal with challenges related to privacy and data protection on a daily basis. They cannot wait until the normative framework is sufficiently crystallized. They are confronted with a level of normative complexity and uncertainty which could have a negative impact, both on achieving scientific goals and on the protection of relevant rights and interests. In the UK, for example, a study has shown that the confusing nature of the regulatory landscape resulted in a culture of caution and (overly) conservative approaches to data sharing (Sethi and Laurie, 2013).

Against this background, some health research initiatives have attempted to engage with and utilise the potential of Big Data, while at the same time ensuring privacy and data protection. To our knowledge, no qualitative research has been published about how this challenge is dealt with by relevant stakeholders in the specific context of such groundbreaking initiatives. By mapping the relevant challenges faced and solutions sought by those involved in the organisation of such initiatives, valuable lessons can be learned. In this qualitative case study, we analyse two real-world examples of data-intensive psychiatric and/or behavioural research. The study is designed to provide insight into challenges related to privacy and data protection in data-intensive health research, and aims to contribute to a better understanding of how rules and interests can be taken into account in a specific initiative or context.

2. Methods

A qualitative multiple-case study has been conducted. The case study is a commonly used empirical research methodology, which allows the researcher to investigate a phenomenon in depth and within its real-world context (Baxter and Jack, 2008; Yin, 2014). Information

* Corresponding author at: Julius Center, room no. STR 5.133, University Medical Center Utrecht, 3508 GA Utrecht, The Netherlands.
E-mail address: m.mostert-2@umcutrecht.nl (M. Mostert).

https://doi.org/10.1016/j.ijlp.2018.07.002
0160-2527/

was gathered about the Big Data Psychiatry pilot project (hereafter: BDP) and the YOUth cohort (hereafter: YOUth). This multiple-case study has been evaluated and exempted from further ethical scrutiny by the Research Ethics Committee of the University Medical Center Utrecht. Explicit informed consent has been obtained from all respondents and the management of both initiatives.

2.1. Case selection and background

The cases have been selected because of their approaches to different aspects of Big Data research. BDP employs a Big Data approach to its analytical methods, in particular for aiding in hypothesis generation. In YOUth, another aspect of Big Data is reflected in its comprehensive data collection, which is continuously being supplemented and updated. Although no clear and widely accepted definition of Big Data exists, such innovative ways in which data are analysed or captured are considered to be core building blocks of a Big Data approach (Mayer-Schönberger and Ingelsson, 2018).

The first case, BDP, aims to explore the potential of Big Data analytics in gaining new insights in the complex psychiatric phenotype. The ultimate goal in BDP is to develop a Big Data analytics instrument that will support health care professionals in their daily practice, for instance by predicting the chance of side effects of medication on the basis of individual patient profiles (Scheepers et al., 2018). A relatively limited set of databases, related to a group of psychiatric patients in Utrecht, was used in the pilot phase of BDP. As a proof of concept, the Cross Industry Standard Process for Interactive Data Mining (CRISP-IDM) was performed on these databases. This resulted in a number of hypotheses and findings, including those related to the themes of aggression during hospitalisation and the effects of medication (Menger et al., 2016). Four working groups have been formed in BDP, and one of these working groups is committed to exploring the theme of privacy and confidentiality. This multi-disciplinary working group focuses on how to safeguard the privacy of participants in the pilot phase and the future programme.

The second case, YOUth, is a longitudinal cohort. YOUth aims to explain why some children develop well and others fail to thrive in society by examining how neurocognitive development mediates the influence of biological, child-related and environmental determinants on behavioural development. The cohort study focuses on psychosocial development, ranging from normal development to deviant behaviour and psychiatric disorders. In order to do so, a great variety of health-related data are continuously collected. These data vary from an array of behavioural and cognitive test results to data about environmental, general child and biological factors (including results from EEG and MRI examinations). The YOUth data being collected will also be linked to other data sources for a broad range of future studies, all in the field of behavioural and psychiatric research.

2.2. Data collection

During our data collection phase, both factual information and the views of different stakeholders from the two cases were collected. The factual information includes internal reports of meetings and discussions, research protocols and other documentation, files related to the application for ethical approval, and text on public websites. Our data collection in YOUth took place between February and April 2017, and in BDP between November 2015 and January 2016. The stakeholders were selected on the basis of their variation in backgrounds and involvement in dealing with privacy and data protection related issues related to the cases. Among the stakeholders, the following areas of expertise or backgrounds are represented: management, lead researcher, research staff, privacy and health law, information technology, consultancy, data management, and patient representation. We conducted 14 semi-structured qualitative interviews in total to collect the views of the stakeholders in both cases. The stakeholders were asked questions related to the challenges they experienced regarding privacy and data protection, and how these challenges were dealt with or should be dealt with according to their views.

2.3. Data analysis

After collecting data, our research group developed codes and identified themes. The full transcripts and other relevant collected data were coded using NVivo. Mostert and Koomen coded the gathered data. Mostert and/or Bredenoord read the coded data and checked the codes for consistency. During the process of analysis, the codes were adjusted through constant comparison across the transcripts and other relevant data and through discussion within the research group. After reaching consensus on the coding, the themes mentioned below were identified by analysing the data. All interviews were conducted in Dutch and the quotes in the results section have been translated idiomatically. The results were presented to respondents to be checked for accuracy.

3. Results

During the process of analysis, it became clear that all respondents encountered challenges or issues related to privacy and data protection. After analysis of the interviews and the other information, three main themes emerged: abandoning anonymisation; reconfiguring participant control, and; the search for guidance and expertise.

3.1. Abandoning anonymisation

The first theme concerns the move away from anonymisation as a strategy to prevent the applicability of data protection law. During the first meetings of the working group on privacy and confidentiality in BDP, some of the respondents adjusted their view on what data could be regarded as anonymous. In this phase, the importance of distinguishing between pseudonymous and anonymous data became clear, but the difficulties in making this distinction were also acknowledged:

"(..) the difference between anonymous and pseudonymous data is hard to understand by layman, and it turned out that it is incredibly difficult for jurists to explain what this difference is. Only after this difference has been made clear, you are able to proceed (..)."(R1BDP).

Afterwards, it became clear to all respondents in BDP that irreversible anonymisation according to the standards as set out in the forthcoming GDPR would severely limit the use of data. Another way to proceed had to be found. BDP chose to integrate a Trusted Third Party (TTP) in the data warehouse architecture of BDP. A TTP aims to facilitate the data linkage process on behalf of multiple data holders in a secure way. Only data that are relevant to a certain research question are extracted from local data sources by the TTP. Afterwards, the different personal data sources are linked by the TTP and a unique pseudonym is assigned to the linked data to prevent future data linkage or enrichment on the individual participant level. The TTP was not considered to be a viable solution in YOUth as it would hinder a permanent enrichment of the cohort with external data sources:

"Or a sort of Trusted Third Party, that is always complicated… because than you need to link data for every single research question and that is a barrier to this kind of cohorts. (..) sometimes I just want to enrich my whole dataset (..)." (R11YOUth).

Furthermore, respondents in both cases regarded de-identifying or pseudonymising data as a challenge, especially when it pertained to unstructured or rich data sources, such as open text fields or imaging data. One of the respondents emphasised the difficulties in de-identifying such data as follows:

"Once you start working with big data, (..) *you could potentially link data sources to enrich the profile of people in such a way that identification may become very easy. (..). With a limited number of variables you could*