Data protection and privacy in research projects

Background material for the Human Brain Project citizen meetings
[data and time] 2016

[location]
# Table of Contents

1. The Europe-wide citizen meetings in the Human Brain Project ................................................................. 3
2. The Human Brain project .................................................................................................................................. 3
3. Data as a valuable resource for research ........................................................................................................... 3
4. EU regulations on data protection ...................................................................................................................... 4
5. Individual rights versus the common good .......................................................................................................... 5
   5.1 The value of privacy ..................................................................................................................................... 5
   5.2 Autonomy and the right to self-determination ............................................................................................... 6
   5.3 Solidarity and the common good .................................................................................................................. 6
6. Anonymity and technologies for anonymisation ................................................................................................. 7
   6.1 What is Anonymisation ............................................................................................................................... 7
   6.2 Anonymisation techniques .......................................................................................................................... 8
   6.3 The challenge of Anonymisation ................................................................................................................ 8
7. Research using personal health data in the HBP ................................................................................................. 8
8. Key points .......................................................................................................................................................... 9
9. Literature list ..................................................................................................................................................... 10
10. Glossary .......................................................................................................................................................... 11
1 The Europe-wide citizen meetings in the Human Brain Project

In February 2016, the Human Brain Project (HBP) will host citizen meetings in Austria, Bulgaria, Poland, Portugal, the Netherlands and Sweden. The aim of the HBP citizen meetings is to understand how EU citizens think about concepts such as privacy and data protection in relation to research projects. When you read this, it means you have been selected to participate in the citizen meeting in your country.

After the six citizen meetings, we will write up a report with your recommendations for how to develop a policy for privacy protection and data management. The results of the citizen meetings will therefore be able to directly influence policy development in the HBP project! The HBP citizen meetings are set up to provide you with an opportunity to reflect on issues related to privacy and data protection in research projects and to give your input on policy developments.

The present document is written to provide you with some background information, and for assisting you in your reflections before and during the citizen meetings. The document is divided into eight main sections. Section 1 and 2, serve as introduction and brief explanation of the HBP, section 3 gives an introduction to the concept of ‘data’, and explains why data is valuable for research, section 4 gives an overview of regulations at the European level, section 5 gives an overview of ethical issues related to data protection, privacy and research in personal information, section 6 explains what anonymisation is and about available techniques for anonymisation, section 7, gives examples of HBP research using health data for its research. Finally, in section 8, the document ends with a summary of the main points of the document. A glossary, containing the definitions of a number of key concepts, can be found at the end of this document.

2 The Human Brain project

The Human Brain Project (HBP) is a European initiative to come to a better understanding of the human brain, and to enable advances in neuroscience, medicine and future computing technologies. The vision of the HBP is to “gain profound insights into what makes us human, build revolutionary computing technologies and develop new treatments for brain disorders.” (The Human Brain Project, n.d.).

The HBP is one of two so-called flagship projects funded by the EU. It was launched in October 2013, and it is planned to run for a 10-year period. The project has a total budget of over 1 billion Euros, and it includes collaborators in more than 20 countries in Europe and beyond.

The HBP is building six ICT platforms for scientific research. In this background material we mainly talk about the Medical Informatics Platform (MIP). The MIP allows researcher to ask questions of personal health data stored in European hospitals. The HBP would like to know how people in Europe think about the use of their data in research.

For further information: http://www.humanbrainproject.eu

3 Data as a valuable resource for research

Data is a word we use to describe information or knowledge that is represented in such a way that it allows for storage, usage and processing. Data could be for example, your address, age, gender,
education, blood pressure, sexual orientation and so forth. In themselves, individual pieces of data might not say a lot about a person, a group or a country. Pieced together however, one starts to be able to make predictions on the basis of certain correlations between individual data points or data sets. Such correlations could be used to gain knowledge about the risk of disease in groups of individuals with certain behaviour. The more data one has to begin with, the more powerful one’s predictions will be.

Researchers are interested in health data, because it provides their research with a lot of power for prediction and pattern recognition. With more data, researchers hope to be able to gain new insights into disease, and they hope that such insights will contribute to better healthcare practices. The access and use of a person’s data is regulated via national law, and EU law and guidelines. In the following section, we will explain EU regulations and opinions on data protection.

Image by Jørgen Stamp - Creative Commons Attribution 2.5 Denmark

4 EU regulations on data protection

In the process of creating, storing and using data there are three different roles: Data subject, data controller, and the data processor. The data subject is for example the individual about whom the data is collected. The data controller is the party that stores and controls the data. The data controller is legally responsible for any breaches of data security or harms coming from use of the data they control. The data processor is the party that may receive and use data from the data controller. (European Parliament, 1995; Stationery Office, 1998) It is worth noting, that EU data regulation is presently undergoing change.

The relevant EU regulations on data protection make sure that no person or institution is allowed to hold or process your personal data, unless you have given permission for it or if it is required by other
laws (European Parliament, 1995, 2015). However, this protection does not apply when personal data is considered to have been anonymised. Anonymisation is considered successful, when it would take more effort and resources to retrieve individual data subjects than what can be reasonably expected.

For example, if re-identifying a data subject from a dataset only requires a normal home computer, and some simple software, data cannot be considered anonymised. However, if it would require an office full of scientists, and one of the fastest supercomputers in the world to perform complex calculations for several weeks, the amount of effort is probably beyond what can be expected from an attacker; especially when the dataset contains relatively insensitive data.

For research purposes, it is worth noting, that researchers also need to have their research approved by local research ethics committees before they can carry out their research. In addition to legal protections of data, there are also moral grounds for protecting health data used in research. In the next section we go through the key arguments.

5  Individual rights versus the common good

Data is being created around us all the time, and the variety of data that exists about most individuals is extensive. This data does not only range from weight, diagnoses and age in a clinical record, there exists a lot of data about other people in other places, for example all your emails, or the people you have called the last year, where you went to school, who your employer is and everything that relates to you on social media. Because some of this data can be very sensitive and private the protection of this data, that was just discussed, is considered part of a person’s right to privacy. This type of privacy is called data privacy.

In this section, we introduce general ethical issues related to research that uses personal data. The first section develops the argument from the starting point of individual rights, while the second section starts from considerations of societal benefits.

5.1  The value of privacy

Apart from considering privacy important in itself, an important common perspective is that privacy is important because it protects data subjects from potentially negative consequences of other people having and using their data. An important aspect of such negative consequences is unlawful discrimination, for example when you are not accepted for a job because you have a high risk at developing dementia at a young age, when you are not allowed to take out insurance because you suffer from mental instabilities, or when you can’t get a loan because of your religion (Rose, 2015; The Danish Council of Ethics, 2015).

Naturally, not all types of information about a person are equally sensitive. However, information about a person’s health, and other data used and produced in research, are typically considered sensitive information. This is a type of information that can affect a person’s ability to change the course of their life, to get a job or to form new relations. The EU also considers data about race, ethnic origin, political, religious –or philosophical beliefs, or information about memberships of unions, health or sexual orientation as sensitive information that requires special protection.
5.2 *Autonomy and the right to self-determination*

In medical research, an individual’s right to make one’s own life choices is secured through informed consent. This right is considered to be important because it allows people live their lives based on their own values. For example, people might have different ideas about the types of research they would like to support. Following this line of thinking, individuals should have the opportunity to decide what types of research their data is used for. Furthermore, participation in research is not always without risks, while the outcome of the research may not necessarily benefit the participants.

The general idea of informed consent is that *every time* a researcher wishes to use data from individual persons for research the researcher has to inform the test subject of all the relevant details of a research project, and then ask for the individual’s permission to use their data for the research. At present, this is exactly how informed consent is structured. This means that scientists need their data subjects to sign an informed consent form for every separate study they do, which is generally regarded as quite a hassle; limiting the amount of available data and slowing down research.

Because of the potential great value of data-driven research, alternatives to informed consent are being conceptualised, that should maintain an individual’s right to self-determination, but increase the amount of available data. One alternative is broad consent, where an individual does not have to give consent to every individual research, but to all research taking place within certain predefined scientific areas. Another example is dynamic consent, where the capabilities of modern IT technologies could be used for researchers to ask a digital and convenient form of consent for individual research.

However, beware that as explained in section 4, researchers do not need to get informed consent from individuals about the use of their data, if the data has been anonymised.

5.3 *Solidarity and the common good*

The arguments in the previous sections have mainly argued from the point of view of individual rights. However, one could also take the common good as a starting point. Research, like for example research...
on brain function, aims to contribute to a better understanding of mental diseases. A better understanding of mental health conditions could contribute to the development of better treatment options and medications. Following this line of thinking, research would benefit society in general. Therefore, the argument is that if we as citizens in a society would like to benefit from the outcomes of research, then we should also contribute our data to this research. The question is then, how much one can be expected to contribute, and what kind of risks one can ask an individual to run for the common good.

Finally, research is not only motivated by the common good. Research is also motivated by financial concerns. Both public and private actors see research as a way of being seen as important in society, and a way to be better than their competition. The development of new medication takes place in large multinational companies, who stand to eventually profit from their inventions. One could therefore argue that, both public and private actors have a social responsibility if they use publically donated data to gain advantages.

6 Anonymity and technologies for anonymisation

From reading the previous sections we have learned that: Personal health data is important for scientific progress, that as long as data is anonymised and the research has approval from a research ethics committee, researchers may use available data for research and that informed consent is not necessary when data is anonymised. In this section we will explain what anonymisation is, and what some of the techniques for anonymisation actually are.

6.1 What is Anonymisation

When items such as for example blood group, genetic information, gender, race, age, weight, brain scans and psychological diagnoses are being organised into a collection, it can be called a person’s medical record. Medical researchers often try to gather a number of individual’s clinical records and put them together into a dataset.

Some of the information in the medical records might be personal identifiers, for example a name or social security number. One might think that in order to make a dataset anonymous, it would be enough to replace these personal identifiers with a code or to remove them. However, even with the best anonymisation techniques available, it could in theory still be possible to de-anonymise a dataset. In the future this theoretical risk could increase with increasing computer power. (Data Protection Working Party, 2014, p. 10).

One example of de-anonymisation of an anonymised dataset is the Netflix Prize example. Netflix released an anonymised dataset with movie ratings of 500,000 users, as part of a competition back in 2006. They had removed all personal identifiers from the dataset thinking that the data had been
properly anonymised, but researchers at the University of Texas showed that 99% of all users in the dataset could be uniquely identified based on six movie ratings and another open-access dataset (Narayanan & Shmatikov, 2008).

6.2 Anonymisation techniques

In order to create anonymised datasets from non-anonymous datasets, and to protect it against the de-anonymisation methods, the data processor has a variety of different anonymisation techniques at its disposal. These techniques can roughly be divided into two different anonymisation approaches: Randomisation and generalisation. We will not go into the mechanisms of these different techniques and their individual strengths and weaknesses, we will only emphasise that: Presently, all anonymisation techniques are theoretically vulnerable to de-anonymisation.

6.3 The challenge of Anonymisation

The problem of creating anonymity by using these anonymisation techniques is that, apart from completely deleting a dataset, there is no absolute anonymity. The challenge for researchers is to find the right balance between anonymisation and usefulness of a dataset. The goal is to apply enough anonymisation for the data to qualify as anonymous without making the dataset useless for research.

7 Research using personal health data in the HBP

The research of the HBP has plans to use existing health data. This is data already collected and stored on the servers of European hospitals. The data could for example be brain scans, previously received therapies, mental evaluation, genetic information, etc.

One of the six online platforms that the HBP is building is the Medical Informatics Platform (MIP). The MIP gives researchers access to datasets stored and collected on European hospitals. However, it is important to understand, that no personal data will be transferred from the hospitals to the MIP. Instead, the hospitals will create anonymised datasets, and send the anonymised datasets to the MIP. The MIP subsequently combines the incoming anonymised datasets into one anonymous dataset that is given to the researchers. This process is also referred to as ‘data federation’.
Researchers will get access to the anonymised datasets by asking ‘questions’ about the content of the datasets on the hospital servers. The MIP is designed so that only some questions are allowed. Questions that could lead to answers that are not in line with the requirement of anonymity, for example if the answer to a particular question reveals too much personally identifying information about an individual.

8 Key points

- **Your medical data** can be very valuable for medical research: The more complete and detailed research data is, the bigger the chance is that scientists come to better understanding of existing diseases, and discover new therapies for existing diseases.

- **Informed consent** (informing test subjects and asking for permission) is needed for scientists to work with personal data.
  - This is not necessary if the data does not allow the data subjects to be identified (when it is anonymised).
  - Data is considered anonymised if the amount of effort and resources that would be needed to re-identify the data subjects exceeds what can be reasonably expected.

- **A challenge in using anonymised data** is to keep as much useful data in the dataset while providing enough anonymisation. The more anonymisation is applied, the less detailed and valuable a dataset becomes.

- An alternative solution to this problem would be to **introduce different mechanisms of informed consent** that should make getting access to data with informed consent easier for researchers.
  - This would mean that data does not necessarily need to be anonymised.
  - Examples of alternatives are broad consent and dynamic consent.

- **The Human Brain Project** is a large scale research project where **anonymised data is being used**, among other things, to improve how we deal with diseases of the brain.
  - In order to do this it is building the Medical Informatics Platform (MIP)
  - The MIP collects anonymised data from several European Hospitals and allows scientists to ask questions about the data.

- Morally there is a conflict between **the common good and societal benefits** and individual freedom and self-determination when it comes to using anonymised datasets without consent.
9 Literature list


## 10 Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>A property that describes the data subject (e.g. weight, age, etc.)</td>
</tr>
<tr>
<td>Broad Consent</td>
<td>A type of informed consent that is not case specific, but that provides informed consent for a range of research goals</td>
</tr>
<tr>
<td>Data controller</td>
<td>The party that stores and protects data</td>
</tr>
<tr>
<td>Data donor</td>
<td>A person that voluntarily provides data to research</td>
</tr>
<tr>
<td>Data federation</td>
<td>The process of gathering anonymised datasets from hospitals to the MIP</td>
</tr>
<tr>
<td>Data privacy</td>
<td>Privacy of data and information</td>
</tr>
<tr>
<td>Data processor</td>
<td>The party that uses the data</td>
</tr>
<tr>
<td>Data subject</td>
<td>The individual to whom personal data relates</td>
</tr>
<tr>
<td>Dataset</td>
<td>A collection of data</td>
</tr>
<tr>
<td>De-anonymisation</td>
<td>The process of identifying natural persons from anonymised datasets</td>
</tr>
<tr>
<td>Dynamic Consent</td>
<td>A type of informed consent in which modern communication technologies are being used to ask data subjects for their consent</td>
</tr>
<tr>
<td>EU Directive</td>
<td>A Legal instrument of the European Union that sets a binding goal for member states, how this goal is reached is up to the member states</td>
</tr>
<tr>
<td>Generalisation</td>
<td>A category of anonymisation techniques that increases the scale of the values in the dataset and subsequently creates identical records in a dataset</td>
</tr>
<tr>
<td>ICT platform</td>
<td>A computer system that serves as a foundation for digital activities</td>
</tr>
<tr>
<td>Inference</td>
<td>Estimating with significant certainty what the unknown value of an attribute is</td>
</tr>
<tr>
<td>Informed consent</td>
<td>A written declaration of understanding of the risks and details involved in research or therapy an individual signs upon receiving treatment or taking part in research</td>
</tr>
<tr>
<td>Linking</td>
<td>Identifying that two records in a dataset belong to the same individual or group of individuals</td>
</tr>
<tr>
<td>MIP</td>
<td>Medical Informatics Platform</td>
</tr>
<tr>
<td>Natural Person</td>
<td>A real human being, not an organisation</td>
</tr>
<tr>
<td>Neurorobotics</td>
<td>A field of research that tries to implement knowledge of the brain in robots</td>
</tr>
<tr>
<td>Neuroscience</td>
<td>A field of science that investigates how the brain works</td>
</tr>
<tr>
<td>Personal identifier</td>
<td>A value of an attribute that leads to immediate identification (e.g. name, social security number)</td>
</tr>
<tr>
<td>Randomisation</td>
<td>A category of anonymisation techniques that add random noise to the dataset</td>
</tr>
<tr>
<td>Recitals</td>
<td>A numbered list of 'reasons' for the contents of a legislation, that precedes the actual provisions.</td>
</tr>
<tr>
<td>Record</td>
<td>A collection of attributes of one person</td>
</tr>
<tr>
<td>Sensitive information</td>
<td>Information that describes characteristics of personality</td>
</tr>
<tr>
<td>Singling-out</td>
<td>Identifying a (possibly unknown) individual from a dataset</td>
</tr>
</tbody>
</table>