Title:	Data Protection and Privacy

Authors:	Nik Rose, Arleen Salles, Josep Domingo-Ferrer, Tade Spranger
Contributors:	Markus Christen, Kathinka Evers, Michele Farisco, Lise Bitsch
Editor:	Bernd Stahl

Abstract:	
Keywords:	

Document Status

Version	Date	Comments
0.1	27.03.2016	Initial draft for comment to EAB and SP12
0.5	18.04.2016	Review by SP12 and EAB
0.5	18.04.2016	Review by Ethics Rapporteurs
0.8	June 2016	Review by SPs
1.0	27.06.2016	Adoption of revised version by SP12 and EAB

1 Introduction

A fuller understanding of the human brain, better diagnoses and treatment of brain disorders, as well as the development of new brain-like technologies are all goals of the Human Brain Project. Realizing these goals requires the collection, storage, curation, and analysis of data of various sorts over extended periods of time.

Securing privacy interests and advancing data protection measures are key concerns of the Human Brain Project. Their importance was recognized during the proposal development, taken up by the Ethics and Society Subproject (SP12) and reinforced by the Ethics Review in Jan 2015. The HBP needs to comply with national and European data protection legislation. But it is clear that the HBP must go beyond existing legal protections and show not only that it is ethically sensitive to privacy concerns even when such concerns fall outside regulatory frameworks, but also that it makes appropriate use of data and is able to identify and respond to new, unanticipated threats to privacy as they emerge.

This document expresses the opinion concerning data protection and privacy by those involved in the Ethics and Society section of the HBP. This includes the members of the sub-project on Ethics and Society (SP12), members of the Ethics Advisory Board and the Ethics Rapporteurs. We identify some of the main privacy-related concerns within HBP, articulate the basic ethical principles that should guide examination of the issues, and present a brief review of the history of data protection and regulation in Europe, focusing on the current state of such regulation. While aware that misuse of the information must be prevented, we are mindful that a form of privacy protection that would prohibit use of any medical or other records for research would stifle medical and scientific progress making it impossible to achieve expected benefits to health that are in the public interest. Therefore, we offer final recommendations that are intended to minimize potential risks while securing the public benefit anticipated from HBP research.

Finally, it is worth noting that there is a more general context to current debates on data protection and privacy. A variety of well publicised events have revealed the extent to which the security apparatuses of different national states acquire covert access to data stored on the internet and mine it in various ways in the course of their work. These revelations influence how citizens think about and how policymakers legislate data protection.

The structure of the the Opinion is as follows: it starts with a description of some of the key privacy challenges and concerns raised by the HBP. The Opinion then describes conceptual and empirical research on privacy and data protection undertaken in the context of the HBP. It outlines technical options and the regulatory environment within which the HBP operates. The Opinion concludes with a set of recommendations to the HBP.

2 Purpose and Audience

There is a rich literature on privacy in general and privacy in biomedical applications more specifically. This Opinion does not aim to review this broad literature. The purpose of the Opinion is to highlight the specific issues raised by the research activities within the HBP by drawing on the expertise of the various individuals involved in the HBP ethics and society work, including members of the Ethics Advisory Board (EAB), Ethics Rapporteurs and members of the Ethics and Society Sub-Project (SP12).

The audience of the Opinion is predominantly internal, i.e. the researchers and scientists as well as managers and decision makers within the HBP. The aim of the Opinion is to provide input into the research and management practice of the HBP to ensure that privacy and data protection are treated adequately. The Opinion will be made publicly available and may as a secondary purpose also contribute to broader debates about privacy in modern large-scale biomedical and other research.

3 Specific Privacy Issues and Concerns within HBP

Some of the concerns relating to privacy and data sharing in the Human Brain Project have much in common with those that affect other initiatives, notably biobanks, that collect personal biological, clinical, demographic, and/or lifestyle data on individuals for the purposes of biomedical research. On the basis of their examination of these issues as they relate to biobanks, Graham Laurie and colleagues usefully argue “that it is valuable to see privacy interests in four interrelated dimensions: (i) physical privacy, (ii) informational privacy, (iii) decisional privacy and (iv) proprietary privacy”, where physical privacy “relates to gathering and storing genetic samples and not tested them without consent”; informational privacy concerns “the possibility of misuse of information, not least the risk of discrimination”; decisional privacy “highlights the interest that biobank participants have in control or influence over what is done with (...) their data and sample”; and proprietary privacy concerns ownership of genetic samples and the control of identity as it relates to our genes” where “proprietary-type claims might be invoked in response to concerns that arise from new technologies such as data-mining and profiling”[1]. These privacy dimensions may at times overlap and at other times be in conflict. To illustrate the latter, allowing people to have control over what is done with their data (i.e. honoring their decisional privacy) might increase the risks of breaches of informational privacy, for de-anonymization might become easier to accomplish and thus more available for illegitimate uses.

In this section, we discuss the various aspects of the HBP that raise privacy and data protection issues. Although informational privacy appears to be more obviously relevant within HBP, some of the concerns intersect with other privacy dimensions as well.

[1] Laurie, G., et al., (2010) Managing access to biobanks: How can we reconcile individual privacy and public interests in genetic research? Medical Law International. 10(4): 315-337.

One major area where concerns about data protection arise in the HBP is within the Medical Informatics Platform (MIP, Sub-Project 8 in the Ramp-Up Phase). The MIP aims to federate clinical data, including genetics and imaging, currently locked in hospital and research records and files with a view to identifying biological signatures of diseases. The MIP is sensitive to the questions this may raise both in terms of consent for the use of clinical data for such research and for data protection of the information contained in the records. To address this, the following approach is used: the medical data is left with the hospitals, and HBP researchers work only with de-identified, aggregated, and anonymised copies of such data. If effective, with this approach none of the previously identified privacy interests would be breached. We discuss a somewhat different moral issue raised by it later in this section.

While the MIP raises some of the most visible data protection issues due to the use of patient data, it is clear that data protection is a more general issue across the HBP. For example, within the Theoretical Neuroscience Platform (SP4) there is some concern that the use of EEGs and fMRIs to model signals in networks to find out what is normal may raise some "physical" and "informational" privacy issues: if a model is fitted to a specific individual, it could allow identification and thus compromise privacy interests. This means that releasing the data would minimally require the data giver's consent, with provenance meta-data and with usage tracking. (It is worth noting that if a threshold resolution were achievable the scan would no longer uniquely identify a person and the data, if anonymized could be released without such consent). While theoretically possible, however, the risk of re-identification in this kind of case seems low when we take into account the efforts and tools required to achieve it.

Another privacy concern (associated with informational interests) is raised by work within SP7 (High Performance Computing): The goal of this platform is not to create new data sets but rather to provide access to them. Related to this, a number of privacy related concerns have been brought to SP12. For instance, the increased power of supercomputers might be used to deduce identity from the available data either by defeating anonymization techniques or by linking large datasets. What this would mean and what the implications for privacy would be are still uncertain. Another concern is related to the implications of the information provided by visualization techniques in cases when the data might show abnormalities in a brain scan that could uniquely identify a patient. While it is important and encouraging that the researchers involved are aware of these possibilities (particularly considering that one common fear is that data will be used in ways that could harm people), these are just theoretical possibilities and they do not appear to be ethically urgent at this time.

Additional privacy related concerns are raised by researchers in the Brain Simulation Platform (SP6). Although SP6 is a consumer of data primarily from SP5, there is a possibility that a model might reveal information that is medically significant (incidental findings). If so, the issue (widely discussed in the context of genomic medicine) becomes whether and how to inform affected patients.

In addition to the above mentioned privacy concerns (physical, informational, decisional and proprietary), it is possible to identify a fifth privacy-related dimension within HBP:

“legality”.[2] This concern is particularly evident in some platforms. For example, MIP’s principles of operation are to access data in response to queries, having ensured that de-identification and anonymization are undertaken by the hospitals that hold the records prior to making them available on dedicated servers, and then re-identification is further guarded against by aggregating the data. A Bayesian algorithm that respects privacy is used to ensure that data from HBP queries are anonymous. This process ensures sufficient signal in the data to perform meaningful analysis. However, because a subset of the hospitals contributing to this platform have a policy that requires that patients should be able to request details of the purposes for which their data have been used, it may be necessary to maintain a table linking patient codes and identifiers. Such a table would be held by local hospital data controllers and would not be accessible to HBP staff or to users of the MIP. This means that researchers using the MIP would not be able to use the code to trace individual patients. However, unless HBP researchers created a logfile for all instances of data use of the anonymized data set, patients would not be able to find out what has been done with their data. This raises a number of issues. First, how long should such a logfile go back in time? E.g. should a patient whose data has been gathered 10 years ago still be in a position to find out what has been done with it 5 years ago? Second, the creation of logfiles would in itself jeopardize anonymization. Indeed, hospitals that uphold the aforementioned policy could not consider the data as anonymous for the purposes of data protection law. Discussions are in progress with hospital administrations to find a suitable solution before the MIP comes online. Third, such a logfile would create risks to security and anonymisation in its own right.

Similarly, for the purposes of data protection legislation, the data controllers of individual hospitals are responsible for anonymised patient data held in their own hospital repositories. The data controller for the overall MIP and for metadata and provenance files will be the partner responsible for the MIP. Thus, despite the adoption of stringent technological measures to guard against re-identification of subjects, the MIP does have the legal status of a data controller, with all the obligations that this entails.

One further aspect worth highlighting is the complexity of data flows within the HBP. Several SPs are designated data creators (e.g. SP1, 2, 3). Other SPs are data processors, aggregators and users. Much of their work aims to use data for the purpose of modelling (SP5, SP6). Other SPs then make use of the data and improved understanding of brain functions (SP8, SP9, SP10). It is also important to underline that the HBP explicitly sets out to use additional data sources, some located in the EU and thus subject to EU legislation, others outside the EU. Data flows can therefore become highly complex and might therefore raise non-obvious privacy-related issues.

A final issue worth discussing here is that of informed consent. Data protection processes often rely on informed consent. In the context of neuroscientific research this can raise significant concerns, e.g. in the cases of patients with mental illnesses, neurologically severely impaired patients with brain lesions or brain problems or in the case of children. This Opinion works on the assumption that valid informed consent can be obtained following established procedures which are typically authorised by appropriate bodies,

[2] Rose N Aicardi C Reinsborough M (2015) The HBP foresight lab report on future medicine.

such as local Research Ethics Committees. Where consent is required but cannot be gained personal data cannot be used for research purposes.

4 Ethical Considerations and Basic Principles

While the use of human data is an integral and indispensable part of HBP research, the implications of such use extend beyond the scientific realm. The need to comply with the relevant regulation is uncontroversial (for examination of Privacy Regulations see section 5). However, compliance with current regulations does not exhaust the ethical issues raised. First, because current oversight might not fully protect people from associated privacy related risks. Second, because, though necessary, legislation is not sufficient to make people more ethically sensitive to privacy violations and more aware of the importance of respecting privacy and the need to meet the duties they might have towards those who provide the data.

Promoting ethical sensitivity and awareness in the context of the HBP requires the identification of the principles and of the relevant ethical considerations that must underlie research. International legal and ethical documents agree in that research with human beings, their data or tissue, should be carried out in a way that reflects basic ethical principles.[3] The principle of respect for persons entails the recognition that the moral status of people does not depend on variables such as class, education or individual achievement. Respect for persons requires acknowledging people's autonomy, i.e. their capacity to make decisions and act on the basis of those decisions, and their integrity, i.e. the inviolability of their bodily and psychological self.[4] The principle of beneficence calls for securing people's well-being, minimizing harms while at the same time maximizing societal benefits. A third, widely recognized, principle is the principle of justice, related to allocation of burdens and benefits of research. These principles are generally used to justify measures -such as data protection and confidentiality- intended to secure privacy.

In the preceding section we noted that there are a number of overlapping dimensions of privacy that can be used to map concerns within the HBP. This is not surprising, for legal scholars and philosophers have framed privacy in various ways (Nissenbaum 2010; Solove 2006; Solove 2002[5]). Some have seen it as essentially associated with personhood and identity; others as related to the capacity to be autonomous; and yet others to the protection of an intimate space. One of the dominant conceptions of privacy in the medical and research contexts in particular relates it to control over information about

[3] For example, the National Commission for the Protection of Human Subjects (1979). The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Nuffield Council on Bioethics (2015) The Collection, linking and use of data in biomedical research and health care: ethical issues.

[4] UNESCO. The Principle of Respect for Human Vulnerability and Personal Integrity. (Paris: United Nations Educational, Scientific and Cultural Organization, 2013); Fjellstrom R. Respect for persons, respect for integrity. Medicine, Health Care and Philosophy 8:2 (2005): 231-242.

[5] Nissenbaum H. (2010) Privacy in Context. Stanford Law Books; Solove D. (2002), Conceptualizing Privacy, California Law Review, 90:1087; Solove, D. (2002) A Taxonomy of Privacy. University of Pennsylvania Law Review, 154: 477-564

oneself, to the extent that others can access or use this information only with the consent of the individual to which that information refers. These different theories recognise that privacy is valuable – either instrumentally (as a means to attain a desirable end such as wellbeing, autonomy, or close relationships) or intrinsically (as an essential component of human dignity itself)[6] – and that breaches of privacy are harmful.

Because of the value many people and many cultures give to privacy and its connection to related cultural values such as autonomy and well-being, considerable attention has been given to the issue of how to protect it while at the same time recognizing that the value of privacy might sometimes be trumped by other considerations, such as people's interest in the social benefits to be gained from research. In the research context, protection of privacy has typically been associated with the notions of informed consent and anonymization, which are at their core supported by the principle of respect for persons.

At the simplest level, the term "informed consent" refers to the idea that individuals have property rights over their data, and that only their explicit permission legitimises the collection, use and disclosure of such data. In order to give consent and thus exercise their right to self determination, people must have access to the relevant information. But it is not just an issue of self determination: informed consent can also be justified by the principle of beneficence, on the grounds that it protects data givers and their interest in promoting the wellbeing of others. While important, however, as information technology becomes more powerful the efficacy of informed consent in offering sufficient privacy protection in contexts such as the HBP has been called into question (Solove 2013; Christen et al 2016[7]).

First, in general, there are values that will always exist in some tension within research, particularly the value of individual data ownership – based on the principle of respect for persons – and the principle of beneficence that in the research context calls for maximising scientific quality and the public good. In the case of HBP, medical and clinical data have been gathered in the course of treatment, often within a publicly funded health care system, with overriding responsibilities to the protection and improvement of public or population health. If the aggregation and analysis of data could be the basis of clinical advances or other developments that would improve the health of fellow citizens, then the principle of the commitment of medical and healthcare personnel to the improvement of the health of all, and not just of each, might well conflict with, and perhaps sometimes override, the principle of autonomous control of individual data by a specific data subject, especially if it can be shown that no harm would flow to that data subject by the use of their data in this way.

Second, biomedical research has been transformed by the application of IT and the dominance of Big Data. Ideally, both the medic/researcher and the data subject can know and understand how and for what purposes their data might be used in the future. But 'big data,' which in this context means the aggregation of very large data sets from multiple

[6] Rachels, J (1975) Why Privacy is important. Philosophy and Public Affairs 4(4): 323-333

[7] Solove D (2013) Privacy Self-Management and the Consent Dilemma. 126 Harvard Law Review 1880; Christen M et al (2016) On the Compatibility of Big Data Driven Research and Informed Consent- the Example of the Human Brain Project. In Ethics of Biomedical Big Data, eds Floridi & Mittelstadt.

sources – such as GP and hospital records, biobanks, repositories of genetic information, data from clinical and pharmaceutical trials, information from longitudinal studies, demographic and social data and much else – relies on data sharing to an unprecedented extent. Handling data on this scale involves multiple procedures for capture, storage, transfer, aggregation, and curation of the data, and complex analytics of search, data mining, using and developing algorithms often themselves produced by machine learning technologies. Much of these data may have been gathered years ago, when big data analytics was not on the horizon, and so no appropriate information could be given to those consenting concerning future uses of the data. Further, even when gathering the data, often researchers do not know all the research questions to which the data is to be directed. This means that, setting aside very general descriptions of purpose, it is quite likely that they will not be able to explain aspects of future research projects, anticipate the potential results, and discuss possible downstream applications now and in the future.

Considering the above, some have suggested 'open' or 'broad consent' by the data subject: subjects agree that data will be widely shared by the research community and used in the future for the public benefit. However, this raises a number of issues, from the general conceptual issue of whether open consent is truly informed to the fact that, particularly in research projects such as HBP, this kind of consent might result in unanticipated harms even if the subjects privacy is not violated. Some of the potential harm may result from a violation of the contextual integrity of the data (i.e. the expected flow of information within the specific context), whereas other harms may result from violating important values that a person has. To illustrate the former, suppose that the anonymized data of a person contributes to research that looks for interrelations between brain health status and the probability of fraud when taking out a private liability insurance. This may lead to a change in underwriting policy that indirectly harms the person whose data was a small puzzle piece contributing to this policy change. An example of the latter would be the following: suppose that down the line the anonymized data of a person allowed the development of a prenatal test for certain brain diseases or psychiatric disorders thus opening up the spectrum for the screening and termination of fetuses thought likely to be affected – but the person that contributed the data may be a strict opponent of termination of pregnancies, either in general or on these grounds. This type of harm, however, is probably not best addressed at the level of individual informed consent but by adequate legal safeguards on the societal level preventing, e.g., unjustified discrimination in case of insurances or certain screening procedures in prenatal diagnosis.

In the context of biobanking, some have proposed an alternative to open consent, what they call the "dynamic consent model" which requires increased data-giver participation (Kaye et al 2015).[8] Dynamic consent uses IT "to satisfy the legal and regulatory requirements for research consent," to engage and communicate with participants who are considered partners in the research process (Kaye et al 2012).[9] However, and despite its potential for further promoting the autonomy of data givers by keeping them continuously

[8] Kaye J. et al. (2015) Dynamic consent: a patient interface for twenty first century research networks. European journal of Human Genetics 23:141-146.

[9] Kaye J Curren L Anderson N et al (2012) From patients to partners: participant-centric initiatives in biomedical research. *Nature Reviews Genetics* 13: 371-376.

informed, even if technically possible, dynamic consent might be psychologically too demanding. Just imagine a person that regularly gets updates and is constantly asked to re-consent to the use of her data in a number of possible ways (both non trivial and trivial). Will people be able to make this kind of decisions on a regular basis? on what grounds? will they be discouraged from participating in research if they think that they will be in repeated interaction with researchers? In light of these and other concerns, it has been argued that even if potentially useful to sharpen consent strategies, a broad consent "combined with ethics review and an active information strategy is a more sustainable solution" (Steinsbekk et al 2013).[10]

Regarding anonymization, it has been argued that complete anonymisation or depersonalisation of data with sufficient preservation of analytical utility in a context of accumulation and matching of many (big) data sources over time is extremely difficult i, and most experts accept that absolute security, confidentiality, and secrecy of personal data cannot be assured by technological means, however advanced. However, the risk of de-anonymization should always be seen in the context of the effort required by it and the legal consequences of attempts to de-anonymize as well as the risk that de-anonymization imposes on the de-anonymized person. When the process of de-anonymization is hard enough, there is no reasonable attack scenario for the average person.

Some of the shortcomings of both anonymization and informed consent could be overcome by the development of trust. Indeed, some have argued cogently that most of those who give or withhold their informed consent, in relation both to medical procedures and to research, do not do so on the basis of a careful analysis of the consent forms, but on the basis of their trust, or lack of it, in the doctors and researchers with whom they are interacting, or on the basis of their general level of trust in the institutions or the political system within which they are located. Thus, while, HBP researchers might be committed to securing privacy while advancing societal benefits, such commitment alone will not ensure that people will trust that they behave ethically. Trust can be fostered by transparency. Thus, to the above mentioned principles (i.e. respect for persons, beneficence and justice) we can add the principle of transparency, which carries with it the implication of moral accountability. Transparency embodies honesty and good communication and thus involves the responsibility for a meaningful account of how decisions regarding the data are made. In turn, it fosters scientific integrity and best practice, fundamental values in science.

5 The Public's View of Privacy and the HBP[11]

Six citizen meetings were carried out in order to understand how the European public view issues of privacy and data protection in relation to research projects and the HBP[12]. The

[10] Steinsbekk K Myska B Solberg B (2013) Broad consent versus dynamic consent in biobank research: Is passive participation and ethical problem? *European Journal of Human Genetics* 21: 897-902.

[11] The analysis presented in this section should be read as preliminary. At the time of writing we are still working on the data analysis.

[12] The citizen meeting were based on the 'interview meeting' methodology developed by the Danish Board of Technology Foundation. The method combines questionnaires with group interviews. A

meetings took place in February 2016, and covered Austria, Bulgaria, Poland, Portugal, The Netherlands and Sweden[13]. At each meeting approximately 30 citizens were present[14]. In this section we report on the views of the participants in relation to: privacy, consent and anonymisation, access and use of personal data, and best practices in relation to data use in research projects.

The public's views on privacy mainly cluster under what we categorise as decisional understandings of privacy. Across all meetings, the majority of the citizens discussed privacy as the opportunity to choose what data about them is shared with third parties. The citizens understood private data to cover a wide range of types: from sports activities, bar visits, content of their correspondence with others, smoking status and friends to political and religious views as well as information on their health. The Portuguese meeting results showed a tendency for participants to both value their ability to decide on sharing of their data with third parties in combination with transparency on how their data is used.

When it came to the issue of consent and anonymisation, the picture from the meetings is less clear. One thing that is clear though, is that the citizens did not immediately think anonymisation, as a stand-alone-solution, to be adequate protection of their data. However what they would like in addition is less clear. The answers from the questionnaires cluster on options that either introduce the involvement of ethics committees to ensure adequate protection[15], or a system that would let individual data providers agree to every use of their data (even in anonymised form). The citizen’s view of data protection therefore seems to go further than what current legislation for the use of personal data in research projects require. When asked about their main concerns in relation to the use of their data, the citizens mainly pointed to worries that their data would be used for financial gain instead of scientific progress.The citizens from Portugal and Austria were also concerned with their data being used against them, while the Bulgarian participants were split between worries over financial gains from their data and concerns about where their data might end up.

The discussion from the Polish and Bulgarian citizen meetings opened up the discussion about consent a bit more. Participants in particularly Poland seemed open to imagining a system of dynamic consent, which would allow them to agree use of their data from case to case. However, they did also comment that such an approach to consent might be experienced as ‘burdensome’ in practice. As a solution they pointed to the opportunity of providing a broader consent to types of research. The Bulgarian participants found it hard to imagine consent as anything but a formality in which citizens do not have much choice, particularly with regards to medical procedures, where consent was presented by them as a precondition for treatment.

meeting lasts for approximately 3 hours. For more information visit: http://www.tekno.dk/article/citizen-meetings-in-the-human-brain-project/?lang=en

[13] The present text is based on the citizen meetings in the Austria, Bulgaria, Poland, Portugal and the Netherlands. At the time of writing we had not yet analysed the Swedish questionnaires.

[14] Citizens were selected to form a representative sample across age, gender, education and socio-economic backgrounds.

[15] The preference for involvement of ethics committees was slightly higher in Bulgaria and Portugal.

It might come as a surprise then, that the majority of the citizens thought both public and private organisations may use their data. Access and use, was however tied up with the condition that such organisations would be strictly controlled for living up to anonymisation standards, and include the involvement of ethics committees to oversee procedures. The citizen's views on use of their data linked up with a decisional understanding of privacy. They agreed that their anonymised data may be used, but only in research projects that they have agreed to. The Portuguese and Bulgarian answers were split between allowing use for research projects they had agreed to, and for allowing use of anonymised data for any research project deemed appropriate by researchers.

Across the meetings the participants expressed a desire for information about data use in research projects. The majority of the citizens also indicated that they did not know where to find information about the use of personal data in research. Transparency on procedures and use, and the involvement of private companies were often mentioned in the citizens' recommendation for improvement in management and use of personal data in research projects.

6 Privacy models and anonymization techniques

A number of technologies have been developed and advocated for data protection and privacy. These include privacy models and anonymization methods, which aim at transforming data in such a way that they cannot be traced back to the individual data subjects to whom they refer, that is, such that subjects cannot be re-identified. Anonymized data should not be confused with de-identified data: de-identification merely refers to removing explicit *identifiers* from the data, but this may not be enough to prevent re-identification (e.g. a 17-year old widow is likely to be re-identifiable, even if her record has been de-identified by removing her name and passport number). Anonymization goes beyond removing identifiers and perturbs or reduces the detail of *quasi-identifiers* (attributes that are not direct identifiers in isolation but that together may identify the subject, such as civil status, age and gender in the widow example). Current EU & US data protection laws do not apply to fully anonymized data. However, the forthcoming EU General Data Protection Regulation still applies to data protected in ways weaker than anonymization, like pseudonymization (replace identifiers by pseudonyms) or the above mentioned de-identification.

Regarding protection of medical data, the U.S. regulations distinguish three categories: identified patient data sets, limited data sets, and anonymized data sets. Identified data sets (that is, fully original data sets containing patients' identifiers) can only be released for research if broad informed consent from all patients has been obtained, which may be impractical. Limited data sets are those where 16 designated attributes have been removed; furthermore, users of limited data sets must sign a data use contract. Anonymized data sets improve data utility without decreasing protection with respect to limited data sets. They can be obtained in two accepted ways: either by applying the so-called safe harbor rules (which basically consist in removing or reducing the detail of 18 designated types of identifiers or quasi-identifiers) or by expert determination (by applying more sophisticated anonymization methods). See an example application of safe harbour and expert anonymization in Sanchez D, Martinez S and Domingo-Ferrer J (2016)[16].

Attributes can be classified in several categories depending on their privacy disclosure potential: *identifiers* and *quasi-identifiers* mentioned above, plus *confidential* (a.k.a. *sensitive*) *attributes* reporting sensitive information on the subject (diagnosis, salary, religion, etc.), and *non-confidential attributes* reporting non-sensitive information.

There are two general approaches to obtaining anonymized data. **Privacy-first** anonymization, favored by the computer science community, uses one or several anonymization methods to enforce a **privacy model** (like k-anonymity, t-closeness or ε-differential privacy); while privacy-first yields *ex ante* privacy guarantees, it often results in anonymized data with poor data utility/linkability. **Utility-first** anonymization, favored by the official statistics community and by most data controllers, tries to preserve utility as much as possible and operates on a trial-and-error basis: an anonymization method is first applied with "mild" privacy parameters, then the disclosure risk is measured and, if it is too high, the method is applied again with more stringent privacy parameters; the process goes on until the risk is brought down to acceptable levels. This iterative process to reduce the risk *ex post* most likely sacrifices some utility too, but it aims at the barely minimum utility sacrifice.

We briefly review the main privacy models in use:

- **k-anonymity**. A data set is said to satisfy k-anonymity if each combination of values of the quasi-identifier attributes in it is shared by at least *k* records. k-Anonymity can be enforced using anonymization methods such as generalization, suppression or microaggregation. k-Anonymity transforms original records so that they are indistinguishable within a group of k in the anonymized data set; however, it may happen that the confidential attribute values within a group are too similar, which would lead to attribute disclosure. Fixes are the **l-diversity** and **t-closeness** extensions of k-anonymity.
- **ε-differential** privacy attempts to ensure that, when a statistical query is made on a data set, the query results be quite independent of the presence or absence of any specific record in the data set. In this way, the privacy of any subject is safe when returning the query response. Noise addition is the usual way to enforce this privacy model, which usually causes a large utility loss.

The main anonymization methods used in the privacy-first or utility-first approaches fall into the following categories:

- *Masking*. A modified version X' of the original data set X is generated. Masking can be perturbative, if X' is a perturbed version of X, or non-perturbative, if X' is

[16] Sanchez D, Martinez S and Domingo-Ferrer J (2016) Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata", *Science*, 351(6279):1274. See also supplementary materials to the previous paper at: Sanchez D, Martinez S and Domingo-Ferrer (2015), Supplementary materials for "How to Avoid Reidentification with Proper Anonymization", http://arxiv.org/ftp/arxiv/papers/1511/1511.05957.pdf

obtained from partial suppressions or reductions of detail in X. While perturbative masking can offer more detail, non-perturbative masking yields truthful data. Example perturbative masking methods include noise addition, microaggregation, swapping and rank swapping, post-randomization, etc. Examples of non-perturbative masking include sampling, generalization, top and bottom coding, local suppression, etc.

- *Synthesis*. A synthetic or simulated data set X' is generated that preserves some preselected properties of X. One can choose between fully synthetic data (where all data records are simulated), partially synthetic data (where only certain values in some records are simulated), or hybrid data (where the anonymized data set is a mixture of the anonymized data set and a fully synthetic data set).

For further background on privacy models and on data anonymization techniques see, respectively, Domingo-Ferrer et al.[17] and Hundepool et al.[18]. References to the seminal papers of each model and technique can be found there.

7 The Current State of EU Data Protection Law and Regulation

Recently, concerns about data protection have become highly salient in Europe (and elsewhere), and they have major implications for data federation in the HBP. There are two related issues raised in this context – legality (i.e., compliance with legal norms and provisions at EU level and at country level) and trustworthiness.

In 2012, the European Commission proposed a comprehensive reform of the EU's Data Protection Directive 95/46/EC from 1995[19], not only because the legislation had been implemented differently in different member states leading to fragmentation and additional bureaucracy, but also because technological progress had changed the way that data was collected and accessed.

While the initial draft legislation generally required specific and explicit consent for the use and storage of personal data, various exemptions regarding medical and health-related research were made. Insofar as certain criteria were met, personal data could be processed for medical and epidemiological research without specific consent from each individual. Firstly, the data needed to be "pseudo-anonymized" which requires the masking of the individual's identity to protect their privacy. Secondly, the research needed to be subject to strong ethical and governance safeguards, approved (for example) by a competent and qualified research ethics committee.

[17] Domingo-Ferrer J, Sanchez D and Soria-Comas J (2016) *Database Anonymization: Privacy Models, Data Utility and Microaggregation-based Inter-model Connections*, Morgan & Claypool.

[18] Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Schulte Nordholt E, Spicer K and De Wolf PP (2012) *Statistical Disclosure Control*, Wiley.

[19] Proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data, COM (2012) 11 final, 25.01.2012.

However, in the wake of the revelations about the mining of electronic data by the US National Security Agency, the draft was amended by LIBE, the Civil Liberties, Justice and Home Affairs Committee of the Parliament. The new draft legislation, being debated at the time of writing[20], prohibits the use of such personal medical data without specific consent by each 'data subject' for each particular use of the data.

This amended draft is currently strongly contested by a large number of medical and scientific research organizations across Europe, on the grounds that it would seriously damage medical research. In the case of the HBP, such a modification of the legislation could make it exceptionally difficult to federate and mine data as proposed by the MIP, where broad consent for the research use of their data has not been obtained from the patients concerned,or from their families or guardians when the patients are deceased or otherwise unable to provide such consent. While the MIP procedures described above aim to provide complete anonymity, and such fully anonymized data falls outside the remit of the Data Protection regulations, as we have seen, some argue that full and complete anonymization with sufficient analytical utility preservation and in a context of matching and accumulation over time of many (big) data sources is extremely difficult, and hence may question the claims made for the anonymization procedure adopted by the MIP.

For instance, even in a cultural environment of strong trust such as the UK National Health Service, the care.data programme in the UK[21] failed to gain the trust of those whose data it would use. The care.data proposal to allow personal data from general practitioners and hospitals to be aggregated in electronic form and mined for the purposes of medical research was mired in a storm of controversy. Although entirely legal under current UK legislation, there was a distinct lack of adequate consultation with the population whose data was to be shared for the purposes of research.

A further set of legal issues arises due to the international and cross-jurisdictional nature of the HBP. The Data Protection Regulation will harmonise data protection in all EU Member States, but its adherence is not guaranteed in associated countries (e.g. Switzerland or Israel) and even less so in other third countries. Further open questions arise with regards to the Safe Harbour principle that is currently being re-developed by various European bodies following the ECJ's ruling that the Safe Harbour agreements were not safe.[22]

8 Recommendations

The description of conceptual, sociological, legal and technical aspects of data protection and privacy shows that this is a complex and emerging field. Across the HBP there are numerous activities that touch on and influence data protection.

[20] Annex of Item Note 5455/16 of the Council of the European Union, 28.01.2016.

[21] http://www.england.nhs.uk/ourwork/tsd/care-data/

[22] http://www.europarl.europa.eu/news/en/news-room/20151015IPR97903/Safe-Harbour-ruling-MEPs-called-for-clarity-and-effective-protection see also http://europa.eu/rapid/press-release_IP-16-216_en.htm

It should be clear that data protection and privacy is not a problem that can be “solved” by following specific instructions or algorithms. Instead, it needs to remain a topic that is raised in the various stages of technology and infrastructure development and that needs to be debated in an ongoing discussion. Only continuous reflection and awareness of the ethical issues at stake will ensure that recent developments are appropriately considered in the HBP. The recommendations below aim at facilitating such a discussion and promoting structures and processes that promote them. Substantive solutions (e.g. anonymisation) will then be reviewed in these discussions.

Thus, considering the discussion of privacy and data protection within HBP, through this Opinion the Ethics and Society SP and EAB recommend the following to relevant decision makers in the HBP:

1. Create a coherent approach to data governance that covers all aspects of research, including data generated, data imported, and data exported by the HBP. Data protection should be one component of this data governance structure. It is likely that this will be achieved by:
    a. Appointing a person who takes responsibility for privacy and data protection across the HBP. This person should be a senior leader and member of the Scientific and Infrastructure Board of the HBP.
    b. Setting up a Data Governance Committee for the HBP comprised of representatives of all stakeholders involved in data collection and processing, and representatives of patient groups and of the general public to review privacy and data protection processes.
    c. Establishing a regular Privacy Impact Assessment for the HBP, and a Research Audit structure that can identify, authorise and audit all users of the MIP.
    d. Ensuring principles of data stewardship which will include finding ways of informing participants in a simple way of how their data has contributed to the public good. This may be achieved public engagement and dissemination programme for the results of the HBP
2. As a general rule, adopt a privacy model when anonymizing data in view of releasing them for secondary use. A privacy model specifies a precise privacy guarantee that can be explained to any interested party. The choice of the model and the model parameters to be used depends on the specific data release (what are its attributes; to what extent some of these attributes are available in external databases containing identifiers like census rolls, social networks, etc.; whether the release is one-shot, longitudinal or continuous in time, etc.). Hence, a specific disclosure protection analysis is needed for each data set to decide on suitable methods and parameters.
3. Encourage the use of systems development methodologies that are geared towards data protection, e.g. value-sensitive design or privacy by design.
4. Explore the potential of ICT tools for managing privacy and data protection related issues in order to achieve a more practical and sustainable consent process.

5. Explore the possibility and potential of broad consent, in particular with a view to the European Data Protection Regulation.
6. Consider the importance of promoting trust and transparency: for example, explore the possibility of having summaries of research proposals accessing and using human data publicly posted.
7. Develop data protection processes that are resilient to technical failure, e.g. by backing up anonymisation technologies with appropriate terms of service prohibiting re-identification of personal data.
8. Ensure regular reviews to evaluate the extent to which technological developments open new and unforeseen possibilities for re-identifying and de-anonymising data.