The Anonymisation of Research Data — A Pyric Victory for Privacy that Should Not Be Pushed Too Hard by the EU Data Protection Framework?
Quinn, Paul

Published in:
European Journal of Health Law

Publication date:
2017

Citation for published version (APA):
The Anonymisation of Research Data — A Pyric Victory for Privacy that Should Not Be Pushed Too Hard by the EU Data Protection Framework?

Paul Quinn
Vrije Universiteit Brussel, Brussels, Belgium

Abstract

Personal health data is essential to many forms of scientific research. Such data may come from a large variety of sources including electronic health records (EHRs), datasets used for previous research and from data linked to biobanks. European data protection law recognises that in addition to using consent as a legal basis for the processing of personal health data for scientific research, such data may be used without consent where it is in the ‘public interest’. Despite the existence of such a legal option, ethics bodies in a number of states have shown reticence to utilise it, often pushing researchers into either obtaining consent or anonymising the data in question. Whilst the latter option may be appealing from a legal point of view, if carried out properly, the result may be that the research value of the data is reduced or even destroyed.

Keywords

scientific research – anonymisation – data protection – consent – public interest – privacy

Introduction

The use of pre-existing personal health data in scientific research is becoming more common. Important sources such as electronic health records (EHRs) and datasets used in previous research offer large quantities of data without the need to engage in the primary collection of data. The use of such sources not only reduces the costs of conducting research on issues associated with
human health and the manifestation of disease, but also allows the combination of differing types of data, opening up new research possibilities (possibilities that are ever expanding given increases in computing power and the availability of new processing techniques). This paper will discuss the legal environment in which such activities take place in Europe. This includes a focus on the main grounds for the processing of personal health data for research processes identified in Directive 95/46/EC, i.e., explicit consent and the possibility of ‘processing in the public interest’. Despite the availability of this latter option, various bodies (such as ethics committees) in many Member States have been reticent to allow research using personal data without consent, even where the conditions for utilizing the public interest option would apparently exist. This has often driven researchers to a choice between two often unappealing options: (i) gaining explicit consent or, (ii) the anonymisation of the data concerned. The former may often be onerous or impractical given the vast amount of data subjects that may be involved and also the diversity of sources from which they are drawn. The latter option may therefore appear attractive from a legal perspective given that it means that the use of health data will no longer be subject to the demands of data protection law. From the perspective of those involved in research however, the use of anonymisation may reduce or even destroy the potential research value of the data, rendering it something of an illusory ‘option’.

This paper aims to explore these issues in order to explain why anonymisation may often present little real alternative to obtaining consent and why therefore the option of ‘processing in the public interest’ remains important. Section 1 will discuss why the use of pre-existing health data is ever more in demand for scientific research. Section 2 will explore why such uses of personal health data give rise to concerns for individual privacy. Section 3 will discuss the current legal possibilities for the processing of health data for research under Directive 95/46/EC. Section 4 will discuss the reticence of some ethics bodies to utilize the public interest option (which the European Parliament is also currently seeking to weaken in the negotiations surrounding the


impending data protection regulation) and the preference for a choice between consent and anonymisation. Sections 5 and 6 will describe why anonymisation is often not in reality an attractive prospect for datasets to be used in research given its potential to harm the research value of the data in question, something which the Article 29 Working Party seems to underline in its recent opinion on anonymisation.

1 A Decoupling of Physical Experimentation and the Collection and Use of Research Data

The nature of the personal data that is used in research related to human health is changing. Historically such data would have been collected directly from physical experimentation that was performed on a select group of individuals. Given that experiments on individuals could only be carried out under the strictest of ethical standards and under the provision of informed consent, data concerning particular individuals would also be collected under such conditions. This meant that the collection and use of medical data for research usually occurred alongside the provision of informed consent by the data subject, providing a legitimacy that also covered the use of the data in question. This ‘legitimacy linkage’ resulted from the increasingly strict application of the principle of informed consent in the post-war decades for activities that could be seen as violating the physical integrity of individuals. Such developments can be traced back to events such as the Nuremburg trials and the Declaration of Helsinki.

Technological developments in the late twentieth and early twenty-first century have facilitated forms of research that are not subject to the traditional relationship described above. This change is the result of several developments. Two of the most important have been the rise in computing power and connectivity between groups, organizations or institutions that may be controlling various data sources. These developments have worked in tandem.

---

5 *Supra* note 1, p. 395.
With regards to the former, processing power has increased enormously allowing computers to undertake complex forms of analysis that were not previously possible. In addition, the ubiquitous nature of the internet and related forms of connectivity have allowed data to be shared across borders between researchers situated far from each other. The availability of such computing power and the access to large pools of heterogeneous datasets have allowed researchers to develop ever more demanding algorithms to analyse data and search for patterns, correlations and links that may be of significance. In this era of ‘big data’ there may often therefore not be a need to assemble groups of individuals to conduct experiments on, and gather information from. The need for the primary collection of data is no longer therefore a necessity for innovative research to occur. The availability of the free exchange of data and the development of new techniques for analysing it means that existing data can be scrutinized in novel ways to create different results. Such developments have been described as creating a situation where old data has been transformed from ‘refuse to riches’. Such re-utilisation could for example involve the use of data from one of three likely candidate sources.

The first involves the re-use of data from previous experiments. Such data may be available through forms of open access or may be made available on a contractual basis. Whatever the commercial or legal arrangement, the internet and related forms of communication have made the transfer of such data between researchers simple. Datasets can be transferred between researchers, institutions and organizations situated in different geographical locations and even different legal jurisdictions. The use of modern processing techniques means that innovative research can involve the application of new processing methods and analysis to primary data that may have long existed in order to generate new forms of secondary data. Realization of the potential for data recycling has been growing in the scientific community in recent times.

---

8 Ibid., p. 1351.
10 Supra note 1, p. 395. ‘Biomedical research increasingly uses methods from data mining, machine learning and text mining to investigate, for example, disease comorbidities, patient stratification, drug interactions and clinical outcome’. 
Preventing (where possible) ‘waste data’ has, as a consequence, been increasingly put forward as a way to reduce costs and organizational difficulties.\textsuperscript{11}

The second involves the use of biobanks and related repositories of health data. Biobanks represent depositaries of physical samples of biological material and often data relating to such material. The purpose of creating such sources is to provide the ability for research organizations to access material and data they would not be able to assemble themselves given the costs and the ethical and organizational constraints that it would entail. Biobanks are valuable because they have enormous sample sizes allowing researchers to utilise large and representative datasets.\textsuperscript{12} They are also important given the breadth of material (and associated data) that they may contain. This permits types of research that may not have been envisaged when the data in question was collected.\textsuperscript{13} Such repositories are usually based on the ‘open consent’ of individuals concerned to allow their data to be collected and used for unknown or even unforeseeable research purposes.\textsuperscript{14}

The third involves the use of electronic health records (EHRs). The gradual acceptance of the EHR by patients, practitioners and health systems across the world has provided for a rich depository of information that was not previously accessible. The value of EHR’s has been described as representing ‘integrated sets of fine-grained longitudinal-phenotypic profiles, facilitating cohort wide investigations and knowledge discovery on an unprecedented scale’.\textsuperscript{15} Such information (where it is accessible) represents a potential ‘gold mine’ for researchers who may be able to use data mining operations to discern relationships and correlations that would not be possible otherwise. This may for example allow important relationships to be detected and analysed that might allow new recommendations in terms of treatment options.\textsuperscript{16}

\begin{itemize}
  \item \textit{Supra} note 1, p. 395.
\end{itemize}
could include issues related to ‘disease comorbidities, patient stratification, drug interactions and clinical outcome’.17

The categories of data above are by no means exhaustive, nor are they exclusive. Indeed, as was discussed above, the desire to harness ever larger and more heterogeneous data sets means that researchers may desire to combine elements of these. Combining for instance, genetic data with EHRs, can allow information concerning phenotype and genotype to be linked.18 This can provide important research opportunities, potentially allowing researchers to look for the effects of various genetic polymorphisms on phenotype (e.g. the effects it might have on the manifestation of disease or the reaction to medication or environmental factors). Together with the socio-economic information that may exist in EHRs such a combination may allow valuable research to be conducted on the influences of genetic and environmental factors on the manifestation of diseases or the reaction to medication.19

2 The Secondary Use of Data Raise Particular Privacy Concerns

Whilst the possibilities in terms of innovative research are ever expanding, developments such as those described above are often perceived of as concerning from a privacy perspective.20 This is often because the gathering of such data and its further processing may often not be based on the explicit and informed consent of those who were involved, as would often have been the case with more ‘traditional’ types of research.21 In many respects this situation represents a ‘decoupling’ of the physical and informational aspects of research

17 Supra note 1, p. 395.
18 P. Chow-White, M. Macaulay, A. Charters and P. Chow, ‘From the bench to the bedside in the big data age: ethics and practices of consent and privacy for clinical genomics and personalised medicine’, Ethics and Information Technology 17 (2015) 189-200. Genetic analysis can now even be used on the individual level for acts of individualized medicine with attendant privacy risks.
20 For a discussion on privacy risks linked to the proliferation of individual genetic profiles available online see supra note 18.
21 Supra note 13.
that in the past would have been inextricably linked. This is because in most traditional forms of research, the data generated would have represented the observation of various characteristics observed in an experimental research setting. The creation of such a setting and the potential physical impositions on individual integrity demanded that informed consent from individuals was secured. Indeed the existence of such consent is viewed as *sine qua non* where the aim is to subject individuals’ bodies, minds, tissue, cells or genetic samples to certain experimental conditions. Given that the generation of research data was inextricably linked to such ‘physical impositions’, they would also receive the legitimacy that a requirement for informed consent provides. This traditional linkage served to provide legitimacy for the processing of the data generated given the arguably more onerous nature of physical impositions. The need to conduct physical experimentation in order to obtain research data therefore imposed, in a *de facto* manner, a culture of obtaining consent for the use of health information for research purposes.

The re-use of health data for research purposes has, in many contexts, weakened this link, raising concerns about the legitimacy of such practices. Such concerns usually focus on the legitimacy of relying on the original consent that was provided when the data in question was collected from the data subject. This is because it may be difficult to reconcile such consent with the proposed research use. This could, in the case of EHRs, for example be because no mention of use for research was made when the EHR was created or, even where data has been taken from previous research given that it may differ from the type of research that is subsequently being proposed. Given the speed at which data processing techniques are evolving, unforeseeable increases in scientific knowledge and ever increasing connectivity, there is therefore a risk

---


24 Supra note 4.


26 Such concerns relate to the idea of ‘purpose limitation’, an important concept in data protection enshrined *inter alia* in Directive 95/46/EC (described in Article 5).
that individuals who consent to their data being used for one purpose may eventually have their data used for entirely different purposes.\textsuperscript{27}

3 Data Protection Law Generally Recognises Two Grounds for Using Personal Health Data in Research

Data protection law in Europe recognizes that health data raises important concerns in terms of personal privacy.\textsuperscript{28} Article 8 of Directive 95/46/EC clearly states that in general, the processing of sensitive data (including health data) is in principle forbidden. There are however a number of exceptions to this. Two of these are potentially applicable to the potential use of health data in research. These are:

3.1 Consent

Article 8(1) of Directive 95/46/EC specifies that sensitive data may be processed with the explicit and informed consent of the data subject. The description of the type of consent required here is noticeably more onerous than that envisaged elsewhere as a legal basis for the processing of (non-sensitive) personal data.\textsuperscript{29} The emphasis on ‘explicit consent’ in particular noticeably differs from the notion of an ‘indication’ used to describe the type of consent required for non-sensitive data.\textsuperscript{30} This difference essentially requires that individuals clearly outline what they are consenting to and demarcate the limits of such consent. The purpose of this is to prevent potential data controllers from using open ended or general consent to justify forms of processing that were not

\textsuperscript{27} Supra note 13.

\textsuperscript{28} For more discussion of this see: Mantovani and Quinn, supra note 22; P. Quinn, A. Habbig, E. Mantovani and P. De Hert, ‘The Data Protection and Medical Device Frameworks — Obstacles to the Deployment of mHealth across Europe?’, European Journal of Health Law 20(2) (2013) 185-204.

\textsuperscript{29} For discussion on the different types of consent that Directive 95/46/EC may require see the Article 29 Working Party Opinion 15/2011 on the definition of consent (2011) 0197/11/EN WP187 p10 Consent, considered as an authorisation by the individual to allow the processing of data pertaining to him/her, may be expressed in different ways: Article 2(h) refers to any ‘indication’; it must be unambiguous (Article 7a) and explicit regarding sensitive data (ex Article 8).

\textsuperscript{30} For more on the need for explicit consent in the health care context see Article 29 Working Party Document on the processing of personal data relating to health in electronic health records (EHR) (2007) WP 131; see also supra notes 22 and 28.
foreseen in the consent provided by the data subject. This creates problems which techniques such as ‘dynamic consent’ (where individuals are in theory able to provide consent on a granular basis) attempt to address, though such techniques present their own difficulties. Although Directive 95/46/EC does not require, some member states have gone further in their imposition of the directive to demand that consent for the processing of health data be in writing and even that it be separate or apart from other consents that may be being presented at the same time (e.g. consent for medical treatment). Such requirements, where they exist, further complicate the mechanics of obtaining consent from data subjects, especially where large datasets are required.

Given such difficulties, relying on consent as the legal basis for the processing of health data in research will often entail the need to collect fresh and specific consent from all data subjects allowing their personal health data to be processed in the manner that is envisaged. Obtaining such consent may come with a significant price, both in administrative and financial terms, given that, depending on the research matter in question, consent may need to be gathered from a large amount of individuals who may not be readily accessible (or even alive).

3.2 **In the Public Interest**

In addition to consent, another relevant legal basis (for the processing of data for research purposes) is provided by Article 8(4) of Directive 95/46/EC. It states that where safeguards are provided, Member States may legislate for processing of sensitive data where consent is not obtained in instances of ‘public interest’. This provision has been used by Member States in their transposition of Directive 95/46/EC (and in other legislation) to permit processing of sensitive data for a range of purposes, including for scientific research (in certain instances where consent has not being obtained). This exception however must be read in a narrow form and comes with important conditionality. Measures that seek to profit from this exception must be

---

31 For more on problems related to open consent and medical research see *supra* note 16.
32 Written consent is required for example by Article 4 of the German Data Protection Act. Consent is also required for the processing of sensitive data in Belgium. *Supra* note 22, p. 8.
33 In Germany for example the use of health data without consent for scientific research is permitted by Article 4(1) of the German Data Protection act. Article 13 stipulates conditions see: Deliverable 9.3 ‘DELSI tools for standardisation and harmonisation to use data from different biobanks’ from the project BIOSHARE, European FP7 Research Project Grant No. 261433.
34 Recital 34 of Directive 95/46/EC.
clearly described in law and be both necessary and proportional in order to achieve the public interest related aim in question. Meeting such requirements may sometimes require conditionality that is demanding and difficult to fulfil. A common example of such conditionality relates to who exactly can engage in such processing. Member State legislation may often specify for example that processing is only to be carried out by individuals subject to obligations of confidentiality. In the Netherlands for example the Data Protection Commission can grant permission for the processing of personal health data without consent with the attachment of certain conditions. These may include that processing is only carried out by persons ‘subject to an obligation of confidentiality by virtue of office, profession or legal provision or under an agreement’. In addition, the research in question must be carried out in a way that guarantees that the private life of individuals is not disproportionately harmed.

In addition to such general provisions in national data protection legislation, Member States may have also enacted special legislation that permits the processing of personal health data specific circumstances. Such legislation usually describes the circumstances in question and may include requirements concerning the use of the data, requirements concerning who has access to various types of data, requirements concerning security (e.g. storage and encryption) and any number of idiosyncratic factors deemed necessary to ensure the proportionality of the measure in question. A prominent illustration of such specialist legislation is the Health and Social Care Act (2012) which paves the way for the controversial ‘care.data’ project in the UK. This project involves a centralization of elements of patient EHRs for inter alia research purposes.

Whilst the public interest option in Directive 95/46/EC allows states to legislate for the possibility of using personal health data for scientific research without consent, the conditionality that is required means that such options cannot be considered as ‘constraint free’. Imagine for instance conditionality that requires an extremely high level of pseudonymisation. Another

---

35 Supra note 30.
36 Supra note 33 section 2.2(c) Dutch Data Protection Act Article 23(2).
38 Supra note 2, Despite the legal authorisation of the project and the existence of safeguards that were argued capable of protecting privacy, this project has nonetheless run into trouble given concerns regarding perceived privacy issues.
39 Section 251 of the Social Care Act (2012) in the UK for instance requires that data be pseudonymised for a range of uses including related to research.
requirement may (depending upon the jurisdiction in question) require that approval is sought and obtained from a national, regional or organisational ethics body. As section 4 discusses below, such bodies may impose further requirements that go beyond the strict letter of the law. In short, one can therefore say that whilst Article 8(4) makes non-consent-based use of personal health data possible, it only does so under certain strict conditions.

4 Regulatory and Ethics Bodies Fear of the Public Interest Option

Despite the clear existence of a legal option for research using health data that has not been obtained with consent, regulatory authorities and ethics bodies have, in many cases, seemingly been reticent to use this option, preferring to insist that researchers obtain consent or use anonymised data. This has been most noticeable amongst ethics bodies which, often to the consternation of research scientists, have become ever more demanding in the requirements that they formulate when personal data is to be used. In many instances such requirements go beyond what is required by both European and national data protection law (i.e. by Directive 95/46/EC and various legislation implementing it in Member States). The position that ethics bodies or other similar entities take can be critical because, depending on national, local, or sectorial regulations, they may have the ultimate say in deciding whether a research proposal concerning personal health data is approved or not. In particular, many ethics bodies seem to have developed a strong aversion for research that aims to utilise personal data that is not accompanied by the consent of the data subjects involved. Such reticence exists even though the possibility for the use of personal health data without consent is envisaged by the law as a ‘public interest option’ in certain cases (see above). The consequence has often been a de facto choice for researchers between opting for consent or anonymising the data in question. This has, according to some research scientists, led to situations where research that is legally permissible is not even contemplated because of an overzealous approach by ethics committees.

Such fears have been intensified in recent years by the process surrounding the formulation and approval of the proposed European Data Protection Regulation. Whilst negotiations between the different institutional players

40 Supra note 2.
41 Supra note 25, p. 5.
42 Supra note 4.
43 Supra note 2.
in the EU legislative process are ongoing (at the time of writing), some of the positions put forward have been the subject of concern for the research community.\textsuperscript{44} In particular, the proposed amendments to the Commission’s original proposal by the European Parliament has raised a number of eyebrows in the research community.\textsuperscript{45} It has \textit{inter alia} demanded a more prominent role for consent in the processing of health data for research purposes, whilst at the same time proposing that the potential scope of application of the ‘public research exception’ be narrowed significantly. In the proposed revisions, produced by the Committee on Civil Liberties, Justice and Home Affairs (the LIBE committee) and supported by the Parliament, the exception permitted for scientific research in Directive 95/46/EC and in the Commission’s proposal\textsuperscript{46} for a new regulation was called into question. In an explanatory note with its proposed amendments the LIBE committee stated:

processing of sensitive data for historical, statistical and scientific research purposes is not as urgent or compelling as public health or social protection. Consequently, there is no need to introduce an exception which would put them on the same level as the other listed justifications.\textsuperscript{47}

The committee recommended that consent should always form the correct basis for the processing of personal health data in a research context unless such research serves a purpose of ‘exceptionally high public interest’. It also recommended that, where possible, health data was to be anonymised or at least pseudonymised to the highest possible technical standards. The standard of ‘exceptionally high public interest’ represents a much higher bar than that described under Article 8(4) of Directive 95/46/EC. It has been suggested that such a high bar will rule out the use of the public interest exception in all instances where the aim of the research in question is not to ease an

\begin{thebibliography}{9}
\bibitem{45} \textit{Supra} note 44.
\bibitem{47} \textit{Ibid.}
\end{thebibliography}
immediate pressing need of public health. This may for instance rule out forms of ‘blue skies’ research where the potential practical application of the results of the research is not yet clear.\footnote{48} Whilst it cannot be argued that such research serves a pressing need, history has shown again and again that such research is important in allowing breakthroughs in the development of scientific knowledge and the treatment of disease.\footnote{49} In addition, the European Parliament’s requirement that the public interest option be available only where individual consent ‘cannot possibly’ be garnered is particularly onerous. Such a formulation appears to bar types of research where, whilst consent could in theory be obtained, it would be disproportionate to do so, making the research in question practically unfeasible or unaffordable.

It is important to note however that the Council appears to have sided with the Commission in this particular institutional ‘tug of war’, meaning that the EU parliament is unlikely to have its way entirely.\footnote{50} The parliament is nonetheless a force to be reckoned with, not only because of its increased powers to block legislation but also because it is one of the few EU institutions that has an obvious source of democratic legitimacy.\footnote{51} What is particularly striking is the fact that the EU parliament seems to want to reinforce the ‘dichotomy’\footnote{52} of consent or anonymisation as the main legitimate grounds for the processing of health data in the vast majority research related contexts (in contrast to Directive 95/46/EC which gives an equally prominent place to the public interest exception). If the EU parliament was to have its way this would arguably

\footnote{48 Supra note 44, p. 1.}


\footnote{50 The Council set out its general position in Inter institutional File: 2012/0011 (COD) Subject: ‘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 (General Data Protection Regulation) — Preparation of a general approach’ Brussels, 11 June 2015 (OR. en) In its position the council supported a public interest option for the processing of sensitive data for historical and scientific research as proposed by the Commission. Such a position is also similar to that described in Directive 95/46/EC. See for example, para. 125(a)(a) on p. 69.}

\footnote{51 The ordinary legislative procedure (which foresees a blocking/amending role for parliament) is described in Article 289 of the Treaty on the Functioning of the European Union. For more on the perceived democratic legitimacy of the European parliament see: V. Schmidt, ‘Democracy and Legitimacy in the European Union Revisited: Input, Output and “Throughput”’, Political Studies 61 (2013) 2-22.}

\footnote{52 For more on this idea of a dichotomy, supra note 2, p. 2.}
represent the ‘setting in stone’ in law of the conservative approach that many ethics and other related bodies have adopted.

5 Anonymisation as a Pyric Victory for Privacy?

The reluctance of regulators and ethics bodies to grant permission for the use of personal health data without consent often means that researchers are faced with a de facto choice between gaining consent or anonymisation (the EU Parliament position, described above even goes as far as calling for anonymisation where possible, even where consent has been provided). The latter option is not explicitly described in Article 8 of Directive 95/46/EC as a legitimate ground for the processing of health (or any other type of sensitive data), nor is it mentioned as a legitimate basis elsewhere in Directive 95/46/EC. Rather than providing a legitimate basis for the processing of health data (or other personal data for that matter) under Directive 95/46/EC, what anonymisation can provide is a way to escape the application of the data protection framework altogether. This is because the anonymisation of data removes the possibility that it can be linked to identifiable individuals. It is this criteria that is identified in recital 26 of Directive 95/46/EC as being requisite in order for data to be described as personal data.

53 European Commission. 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 (General Data Protection Regulation). COM(2012) 11 final. 2012, http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf. Amendment 238 states ‘...The data in question shall be anonymised, or if that is not possible for the research purposes, pseudonymised under the highest technical standards, and all necessary measures shall be taken to prevent re-identification of the data subjects'.


55 Supra note 46. Article 4 of the proposed regulation sets out a similar position. The Working party has also confirmed that where data does not possess such a property it falls outside the scope of application of Directive 95/46/EC stating 'Anonymisation may be a good strategy to keep the benefits and to mitigate the risks. Once a dataset is truly anonymised and individuals are no longer identifiable, European data protection law no longer applies'. See Article 29 Working Party Opinion on Anonymisation Techniques (April 2014) 0829/14/EN WP216, p. 5.
Given this, the potential to use anonymised data for research purposes would appear to be an extremely attractive prospect from a legal and ethical perspective. This is because the use of anonymised data would immediately avoid many of the legal and ethical obstacles to using health data for research. In terms of the law, there would be no need to give thought and consideration to the demands of meeting one of the legal bases for the processing of sensitive data (i.e. under Article 8 of directive 95/46/EC as described above). This includes the need to consider an effective formulation for consent (i.e. that it be both explicit and informed) or the need to meet the conditionality required by a public interest exception for research (where such an option exists). In terms of ethical concerns, the use of anonymised data would alleviate many (but not all) of the concerns that ethics bodies have (and which often go beyond the letter of the law) concerning the use of health data. The option of anonymising data thus represents an attractive option, legally speaking, for those considering research that will require health data. The use of anonymised data would, to a large extent, free researchers from hindrances associated with data protection law and allow them to process datasets in far more constraint free manner.

In reality however this option is far from simple and may be difficult or even unachievable in many research contexts. This is because data that is truly anonymous may often offer little or no potential in terms of research value. Very often data is only of use where it contains personal (or quasi personal identifiers) that allow the data in question to be analysed within specific

---

56 Supra note 25.
58 Supra note 2, p. 1. The use of anonymised data on some occasions may still engage some legal spheres; ‘measures such as anonymisation (even when possible) do not solve all ethical, legal and technical problems; people may, for example, have religious or moral objections to particular studies or concerns about stigma and breaches of privacy’. The use of anonymised data can still give rise to concerns related to surveillance issues.
59 It is possible that on certain occasions other legal rights may be attached to non-personal data — these could for instance be in relation to any intellectual property that the data might represent. Other legal rules may apply, these include (as identified by the Article 29 Working Party) Article 5(3) of the e-Privacy Directive and the principles and case law identified by Article 8 of the European Convention on Human Rights and Article 7 of the EU Charter of Fundamental Rights, *supra* note 3, p. 11.
contexts. These could include a range of useful metadata that link conditions to a particular medical dossier, or attributes that allow a number of dossiers to be linked together (e.g. showing familial relationships). Other attributes might relate to factors that could be linked to important socioeconomic factors, e.g. the date of birth, gender, and area of residence). Such factors might be important in data analysis, especially related to health matters where they can be used to gage environmental effects upon the manifestation of a particular phenotype. Whilst they may therefore be beneficial in terms of research value, their presence, especially in concert with other potential identifiers may mean that it is possible to identify the individual or individuals which the data relates to (under certain conditions). Moreover, it may be possible to use such information in concert with data kept elsewhere to link information that may appear ‘non-personal’ to individuals and thus make it ‘personal’ data. Whilst it may be possible to anonymise such data (e.g., by thoroughly removing all linking elements), doing so may render it devoid of usefulness in terms of potential research value.

6 A High Bar Set for Anonymisation by the Article 29 Working Party

The working party has confirmed (or even highlighted) these problems in its recent opinion on the matter of anonymisation. In that opinion it set what can be considered an extremely ‘high bar’ to be met for actual anonymisation to occur. In particular, the working party was concerned that anonymisation not be confused with processes of ‘pseudonymisation’. This has long been an issue in efforts to improve compliance with data protection demands. Pseudonymisation implies that efforts have been made to reduce the possibility that a particular data set can be identified as belonging to a particular individual. This may involve removing unique identifiers such as names, social security numbers and dates of birth. Whilst intuitively such measures

---

61 Supra note 1, pp. 395, 397, 399.
62 Supra note 54, pp. 403-405.
63 Article 29 Working Party Opinion on Anonymisation Techniques (April 2014) 0829/14/EN WP216, p. 3.
65 Whilst pseudonymisation may not constitute anonymisation it can nonetheless be an important process in protecting the privacy of individuals. See: S. Lusignan, ‘Effective pseudonymisation and explicit statements of public interest to ensure the benefits of
may seem to anonymise the data in question, the reality is that they just make identification of the data subject more difficult.66 Those in control of such (pseudonymised) datasets may be able to take certain measures to allow ‘reidentification’ of the data subjects in question. This could for example include referencing pseudonymised datasets to master datasets where cross referencing will allow data subjects to be identified. Even where the controller of the pseudonymised dataset does not have access to such a master dataset (as is the case in more stringent forms of pseudonymisation) it may be possible, without an inordinate amount of difficulty, to identify data subjects by reference to other datasets that the data controller may be able to gain access to or which are even publicly available.67 One common example may include records that are made public such as records of births, deaths or electoral registers. Given that the efforts required to perform such re-identification are not very onerous, such data cannot be considered as being anonymised, but merely pseudonymised.68 Pseudonymised data may also be more vulnerable to deanonymisation by malevolent third parties.69

Indeed, it is the effort required to re-identify or (or ‘deanonymise’ as the working party terms it) data subjects that the working party uses as one of its primary criteria, stating that it is necessary to look at the ‘means…that are reasonably [sic] to be used’ to deanonymise data in order to determine whether the efforts made to anonymise the data in question are sufficient.70 This test essentially requires a determination of whether identification of


68 Supra note 55 in Annex. In its opinion on anonymisation the working party suggested several examples of methods that could be considered as anonymisation and not just pseudonymisation.


70 Supra note 63, p. 3.
individuals using the anonymised data is ‘reasonably impossible’.71 Whilst the use of the word ‘reasonably’ may seem to connote a low standard for anonymisation, the working party has made it clear, in particular with its juxtaposition to the word ‘impossible’, that the standard is actually very high. Several factors identified by the working party are testament to this (especially when taken in combination).

First, it requires data controllers to focus on the means that would be necessary to bring about deanonymisation.72 This requires a consideration of evolving technical possibilities in terms of computing power and the availability of algorithms that are able to deanonymise data that was thought to be anonymous. In doing so it is necessary to balance anonymisation effort and costs (in terms of both time and resources required) against the increasing availability of technical means to identify individuals in datasets and the increasing public availability of other datasets (such as those made available in connection with open data policies) that may be of use in such deanonymisation.73

Second, the working party has stated that, in making such a determination, it is necessary to take into account the fact that many types of publicly available datasets that are claimed to be anonymised may not meet the requisite standards of anonymisation.74 Such a standard requires that the party anonymising data, not only consider their own ability to deanonymise the data in question, but also the ability of other known and unknown parties given the state of technological development and other potential sources of data that may be publicly available. Given the rapidity with which computing power is increasing and the increased availability of research data online, the threshold for true anonymisation to occur may be extremely high. Imagine for instance the use of genetic data or its publication online following research. Given the nature of the data involved (where even small DNA sequences may provide a link to specific individuals) and the potential for related information concerning the individual or a family member to exist in an accessible version elsewhere, it may be difficult to speak of genetic data as ever being truly anonymous.75

---

71 Ibid., p. 8.
72 Supra note 69, p. 75. Such a focus is drawn from recital 26 of Directive 95/46/EC; supra note 74, p. 75.
73 Supra note 63, p. 9.
74 Many such datasets may more realistically be described as ‘pseudonymised’. For more see also supra note 54.
75 H. Schmidt and S. Callier, ‘How anonymous is ‘anonymous’? Some suggestions towards a coherent universal coding system for genetic samples’, Journal of Medical Ethics 38(5)
A third important factor is that one cannot depend upon the ‘good motives of the data controller’.76 This means that the data controller, when assessing whether a dataset he or she possesses is truly anonymous, must take into account what other data they have access to. If the controller of the supposedly anonymised data set has access to other data that will allow the identity of individuals to be decided through cross referencing the two, then it is not correct to speak of anonymised data. Given this, the data controller must make sure that those with access to anonymised data (even within their own organization) do not have access to other datasets that might facilitate deanonymisation or, must employ such a level of anonymisation so that, even with reference to other datasets, deanonymisation will not be possible. Once again such measures, where employed are likely render the dataset in question less valuable in terms of its research potential (or even remove such value altogether).

Fourth, the working party confirmed that in its opinion, the act of anonymisation itself constitutes an act of processing of personal data.77 This is logical as in order to anonymise data the data controller must have been in possession of data that was not anonymised i.e. personal data.78 Given this it is also logical to expect that the original dataset in question was obtained in accordance with one of the legal bases described above. This may create a ‘catch 22’ because it means that in order to collect personal data, even if the intention was to immediately anonymise it, it would be necessary to have the consent of the data subjects involved. Where the purpose of anonymisation is to avoid the need to obtain consent, this will present immediate problems because, where such consent has not been obtained, it may not be possible to gather the data in the first place.

---

76 Supra note 63, p. 10.
77 Ibid., p. 2. The working party states ‘Anonymisation constitutes a further processing of personal data; as such, it must satisfy the requirement of compatibility by having regard to the legal grounds and circumstances of the further processing’.
78 Supra note 69, p. 79. As Khaled states: ‘In order to anonymise data, it is necessary for an anonymisation engine to ingest personal data, apply anonymisation techniques to it, and then output anonymised data. The input is personal data’.
8 Conclusion

Whilst anonymisation may appear attractive from a legal perspective, it is likely to come with many drawbacks, mostly related to a reduction in the research value of a particular dataset that may occur. This has been confirmed by a recent opinion issued by the Article 29 Working party which was keen to emphasize the distinction between actual ‘anonymisation’ and processes of ‘pseudonymisation’ that may often be assumed incorrectly to constitute the same thing. For anonymisation to occur it must be established that it is ‘reasonably impossible’ to re-identify particular individuals from the dataset in question. Given advances in modern computing power, the development of sophisticated deanonymisation algorithms and the increased availability of potentially complementary data online, the level of anonymisation needed to meet the standard of ‘reasonably impossible’ as described by the working party may be very high. As such possibilities for deanonymisation multiply with the accumulation of further technological and analytical advances, the standard required to achieve anonymisation is likely to become ever higher.

The effect of the Article 29 Working Party’s high bar for anonymisation is that truly anonymised data may not be suitable for many forms of research. Whilst it may therefore be suitable in certain circumstances as an alternative to obtaining consent legally speaking, in many cases it will not represent a viable option given that it will in reality render data largely useless. Pushing a vision of (as many ethics bodies have done) a choice between obtaining consent (which is also often not feasible) and anonymisation is therefore not realistic and not conducive to fostering an innovative research environment that will help solve problems that we collectively face.

In place of pushing such a false dichotomy this author would submit it is necessary to bring about a realignment in the attitudes that exist concerning the potential harm to individual privacy through research that is conducted under correct conditions. As Directive 95/46/EC recognised, processing of personal health data may be required in certain instances in the public interest, including in matters of scientific research. The availability of such an option as an alternative to consent needs further appreciation and demystification. In particular, it is necessary to realize that there are valid reasons for the existence of the public interest option and that it has been formulated in order to balance any risks that may be created to individual privacy with the potential gains that may be realised from the research in question. Whilst frivolous and unnecessary use of such an option should be prevented, it is necessary to recognised that ‘explicit consent’ is not and cannot be the only acceptable basis for the processing of personal health data.
The proposals made by the European Parliament have however arguably made such demystification more difficult and raised confusion about the future direction of the law. Even if the final period of institutional negotiation may go some way to softening the proposals put forward by the Parliament, the harm done in terms of pushing the dubious dichotomy of obtaining consent or employing (full) anonymisation may be long lasting.