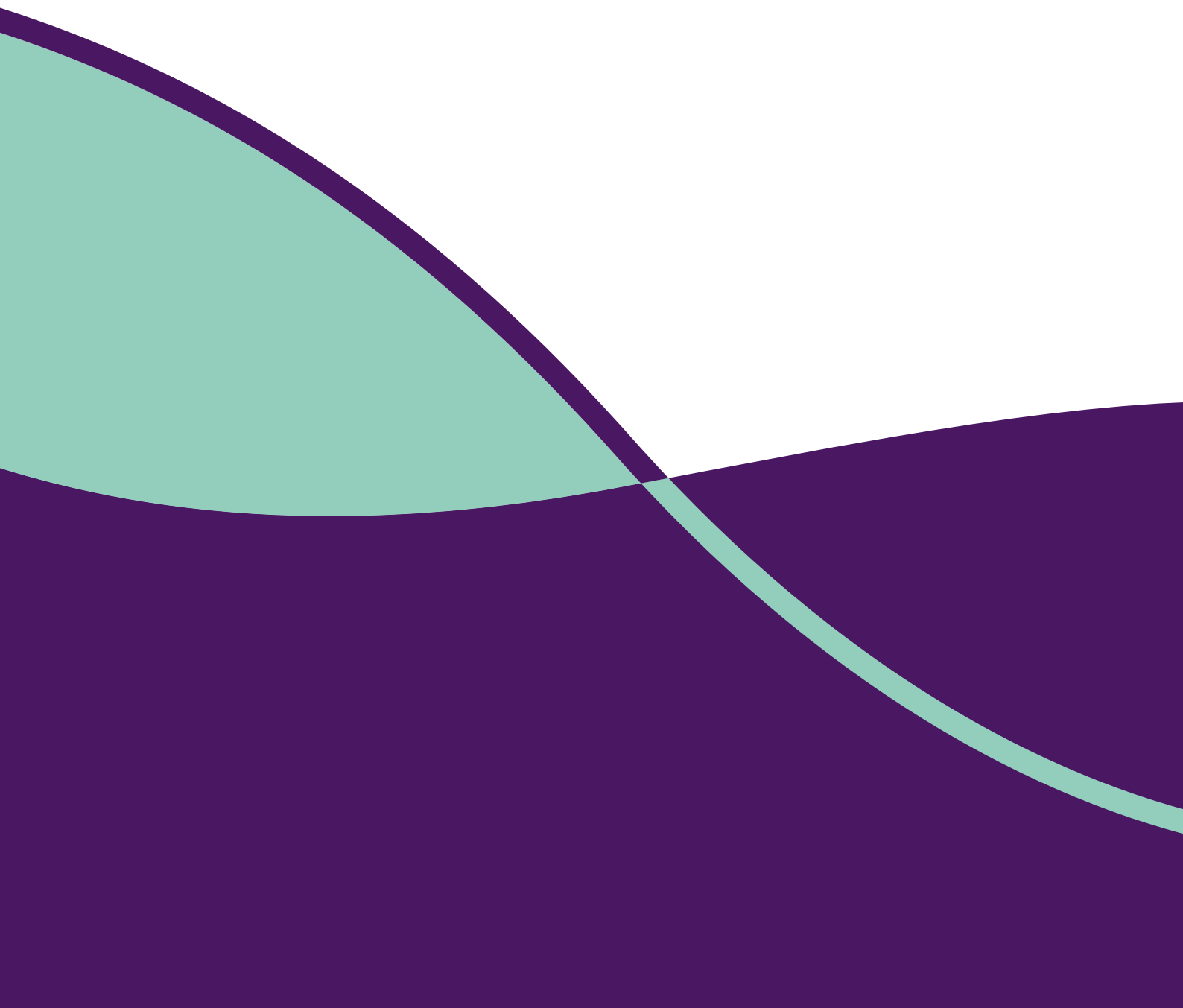
phgfoundation
making science work for health

Identification and genomic data

December 2017



Authors

Thomas Finnegan, Alison Hall

Acknowledgements

The PHG Foundation is grateful to the expert contributions and advice provided by: Dr Jeffrey M Skopek, University of Cambridge, Neil M Walker, University of Cambridge and Susan E Wallace, University of Leicester

NB: URLs in this report were correct as at October 2017

This report can be downloaded from our website:
www.phgfoundation.org

Published by PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
UK

+44 (0) 1223 761 900

Correspondence to:
allison.hall@phgfoundation.org

How to reference this report:

Identification and genomic data
Finnegan T, Hall A
PHG Foundation (2017)
ISBN: 978-1-907198-22-9

Contents

Executive summary 4

1. Introduction 9

2. What are genomic data? 11

3. Challenges and limitations of anonymisation 13

4. Techniques and terminology 18

5. How are data governed? 24

6. How can anonymisation techniques be undermined? 28

7. What do people think about genomic data? 33

8. Conclusions and recommendations 35

References 37

Executive summary

For many decades, various techniques have been used to obscure the source of personal data to protect against foreseeable harms such as stigmatisation and discrimination. The effectiveness of these strategies are strained when they are applied to genomic data where the potential identification of people using anonymised data is a growing challenge. There are several reasons for this: the very nature of genomic data, regulatory requirements, complicated new techniques for identifying people within anonymised datasets, the increased ability to link unrelated datasets and the need for the appropriate data to be used in legitimate ways to support clinical care and scientific research. It is a complex subject with a significant academic literature, but there have been only a few attempts to succinctly describe the issues to healthcare professionals, those who process data and policy-makers; this report aims to help redress this.

To what extent is the possibility of identification within supposedly anonymous datasets a real problem? There is an argument that because the techniques used to identify people within such datasets are complicated and difficult to apply, the likelihood of it happening is low enough that we need not be concerned. It is true that the techniques are difficult: only a relatively small number of people are capable of achieving reidentification, and the probability of success is low. But the topic is important: because the possibility of reidentification has potential to undermine public trust and confidence in the systems designed to keep people's data secure.

Genomic data have certain characteristics that make the problem more complicated than for other data. The genome is an organism's complete set of DNA including all of its genes; genomic data may include sequenced DNA that can be in the form of raw data derived from sequencing a person's genome in whole or in part (whole assembled genome sequences or whole exome sequences – the genes that encode proteins) or individual DNA variants. Therefore, genomic data can encompass data with quite different characteristics and which can be used in many diverse ways.

Genomic data are exceptionally useful, and promising, as a medical and scientific resource. But in this context, they are particularly challenging:

1. Genomic data are difficult to render anonymous *while also using them productively* – they are strongly identifying and the uses to which they are put are easily undermined if the data are manipulated in certain ways
2. When considering why genomic data are strongly identifying, significant flaws in the EU data protection regime arguably become apparent

In the UK and other countries, privacy and anonymity are considered worthy of protection but they are not of absolute importance; they are subject to other contending interests such as the societal or economic value of the use of data. To enable the use of data, the law tends to rely on specific legal exceptions, obtaining consent from the person to whom the data relate, or attempting to manipulate data so that it cannot be used to identify someone (and therefore protect their privacy). In many jurisdictions, including the UK/EU, data that cannot be used to identify a person have fewer or no protections.

Some mechanisms that might otherwise enable lawful data processing, such as obtaining detailed consent, or justifying the processing on the grounds of public interest may be restricted because of the nature of genomic data and how those data are used. For example, those running large research projects that use sequence data from a great many people may find it unfeasible to obtain sufficiently detailed consent from all those people. In addition, sequence data are not always just personal data – those data can relate to the family too, so it is not always clear from whom one should obtain consent.

Consequently, where specific legal exceptions cannot be used, anonymisation techniques are often implemented. But making data truly anonymous can be difficult depending on the nature of the data and the context within which such data might be used. Again, genomic data can pose challenges. For example, whole genome sequences are complex data that can form many different links with other datasets and so are highly identifying – removing some specific identifiers is not always enough; obscuring or adding values within datasets may be effective, but may be especially undesirable with genomic data because it reduces accuracy and therefore diminishes usefulness; moreover, each genome sequence is unique, so some forms of aggregation do not work.

Data can be encrypted, but genomic data are likely to remain relevant long after the encryption system becomes obsolete, because their relevance continues throughout an individual's life. The data will also be relevant to the family of that person, extending the lifespan of the data even further. In addition, data analysis tools applied to encrypted genome data tend to be less efficient and the computing requirements for working with such data can be problematic. This means that other techniques that do not rely on manipulating the data to avoid identification, but instead rely on restricting access to data, might be less effective.

Another complexity is the claim that the genome is in some way intrinsically identifying, which arguably it is not. Examining why it is not reveals a common, but false, understanding of what it means to be identifiable: the belief that there is some trait that makes data identifying, and the removal or modification of such a trait will render those data non-identifying. This is not the case. Rather, the state of being identified is a consequence of the interplay between the uniqueness of the data and the connections made between different data; it is the outcome of a network of associations. Genomic data, or more precisely genome sequences, are confused as being intrinsically identifying because they are both unique and capable of being connected to many and diverse data. This does not make them intrinsically identifying, because no data in isolation can be identifying. Instead, it means they can be used to make many connections; thus genome sequences are strongly identifying, not intrinsically identifying.

An important distinction that is frequently overlooked in both the technical and policy literature is the difference between 'identification' and 'individuation'. Philosophers use the latter term to describe differentiating the individual from the general and universal. Characterising certain sets of data as belonging to a unique individual is not synonymous with identifying that individual: the potential harms that arise from identification and individuation are therefore different.

EU data protection law categorises data into different types based on both the ability of those data to identify a person and also on their relative sensitivity. Under EU data protection law, data are offered varying levels of protection based on those categories. Only data capable of identifying people directly or indirectly are offered protections. But this approach perpetuates the misconception that the ability to identify a person is an inherent trait of data rather than data being identifying by reference to other data – deleting or disassociating a particular type of information cannot necessarily prevent identification: both the connections and the nature of the data matter.

Conclusions and recommendations

Genomic data are extremely useful medically and scientifically and can be used to great benefit. Although somewhat flawed, anonymisation techniques as applied to genomic data are useful tools when attempting to maintain privacy. However, genomic data do not sit comfortably within the current legal and regulatory framework as a consequence of their nature and an overall lack of regulatory coherence. This report proposes a number of responses to this challenge:

- **Take greater account of societal and technical change:** questions of privacy and anonymity have become pervasive and deserve urgent and extensive discussion as their implications extend to every level of society. This applies particularly to health data of which genomic data are part.
- **Do not rely solely on anonymisation:** greater emphasis should be placed on the use of systems that appropriately limit access to the data. Important elements include access control systems, audit, and legal sanctions. Although sophisticated techniques that manipulate data do have a very important role to play in protecting anonymity, they cannot be relied upon in all circumstances or over long periods of time.
- **Ensure transparency:** anonymisation techniques/mechanisms should not be hidden and their effectiveness should not be taken on trust. This is especially the case for systems that rely on public funds and use data derived from publicly funded systems.
- **Understand the difference between 'identification' and 'individualisation':** the distinction between these two activities needs to be critically evaluated, both in the scientific and the policy literature. This is necessary in order to ensure that the risks and harms associated with the two activities are correctly understood.
- **Move away from language that appears to be absolutist:** there are strong reasons to continue processing, linking and releasing genomic data despite failures of anonymisation, but public discussion, and comments made to patients and research participants, should make clear that in many cases anonymisation is not absolute and that it is not possible to fully determine the risks of future deidentification. Claims made about anonymisation ought to be moderated – patients, research participants, consumers, and doctors are unlikely to recognise the limitations of anonymisation without being told of them.
- **Pursue further research:** research into "*the potential harms associated with abuse of biological and health data, as well as its benefits*"[14] is needed to evaluate the nature, extent, and likelihood of such harms arising and determine how potential challenges should be optimally addressed.

Conclusions and recommendations continued

- **Consider regulatory change:** The EU data protection regime emphasises the nature of data as a way of moderating its use, rather than the use itself. Instead, there have been calls for greater monitoring and sanctions for illegitimate identification of individuals, and the PHG Foundation supports ways of providing additional protections. These includes stronger, and in some cases criminal, penalties for misuse of data and the possibility of bringing inappropriate identification within the scope of data protection law. We welcome the fact that the UK Data Protection Bill includes sanctions for unauthorised re-identification of de-identified data within its scope[89,90].
- **Establish a Council of Data Ethics:** the Government should empower a multidisciplinary transparent forum to address the technical, bioinformatic and regulatory challenges associated with anonymisation in order to optimise robust and proportionate data sharing practices in healthcare and medical research. This body could also help to promote national debate and resolve challenges associated with anonymised data in ways that build public trust and confidence.

1. Introduction

In this report we discuss the use of genomic data to identify people in situations where identification is unwanted and efforts have been made to prevent it. We aim to explain to non-technical readers the complex and much-discussed challenges posed by identification and anonymisation, why those challenges are important, and how they relate to genomic data.

Society places value in people sometimes remaining unidentified. Identification might result in an observable and measurable harm to a person or could pose a more intangible sense of harm – a feeling of violation, for example. The exact nature of harm resulting from identification, or loss of privacy or anonymity, is a matter of considerable controversy: is the harm the breach of privacy itself or must some concrete harm have occurred? Is the mere possibility of identification a harm?

To some extent, the law reflects these risks by relying on techniques intended to ensure data cannot be used to identify people. However, there is a misalignment between the objectives of those laws, what those laws require, the nature of the techniques for keeping people unidentified, and the expectations created by the language used. This is especially the case with regard to genomic data – partly because of the nature of genomic data and partly because of the way these data are used.

In the UK (and the European Union generally), the law prohibits the processing of genomic data unless certain conditions are met. These conditions include specific exceptions (e.g. clinical care), the agreement of the person to whom the relevant data relates (consent), or situations where data cannot be used to identify people. The different techniques for achieving the last of these are generally referred to as 'anonymisation', but the term has different meanings in different contexts. Consent is not the focus of this document, and it is relevant here only because in the absence of consent, anonymisation techniques have traditionally been used to permit the use of data.

Anonymised data are not protected or regulated in the same way as other data. In the Member States of the EU, for example, data protection laws influence the use of genomic data very strongly and anonymised data fall outside the scope of those laws.

Some reasons for the misalignment mentioned above include that:

- It is often impracticable to obtain consent in many of the situations in which genomic data are used. This results in reliance on other approaches, such as manipulating data to reduce the probability that identification will take place
- The techniques for manipulating data to make them non-identifying are useful but do not always work and may work less effectively when applied to genomic data
- The language used to describe what will happen is sometimes inaccurate or ambiguous

This can lead to the distribution and use of data that are not technically anonymous but are treated as such by the law. This increases the risk of a person being identified from those data and raising fears of a breach of privacy.

This widely recognised problem is yet to be resolved. In fact, the laws that embody this approach have been recently entrenched by the EU General Data Protection Regulation (GDPR), which was passed by the EU in 2016 and has 'direct effect' upon all EU Member States without the need for them to pass implementing legislation. The provisions of the GDPR will become applicable in May 2018. A forthcoming UK Data Protection Bill will clarify how the exemptions to the GDPR will operate in practice. This will have implications for the way in which data can be processed for health and medical research within the UK[1].

There has not been enough transparent, high-level, public debate about the challenges posed by and to anonymisation, despite the fact that the issue is extensively discussed in the literature and genomic data are central to a number of national health-science programmes. We hope this document will encourage debate and raise the visibility of these issues by clearly and simply describing the challenges and what might be done to manage them. We do not claim that anonymisation techniques have failed completely and should not be used, or that the risk of reidentification is extremely high; the challenges to anonymisation in the context of genomic data require action, not panic.

2. What are genomic data?

How one determines the nature of genomic data is a question of technology, law and policy. Policy makers tend to treat genomic data as exceptional on the basis of its characteristics which raise questions of privacy and anonymity: we argue that this approach is erroneous because only a sub-set of genomic data is identifiable and predictive.

From a very basic technical perspective the genome is an organism's complete set of DNA including all of its genes; 'genomic data' is a broad term referring to sequenced DNA that can be in the form of raw data derived from sequencing, a person's genome in whole or in part (whole assembled genome sequences or whole exome sequences – the genes that encode proteins), or individual DNA variations.

Each human genome consists of around 3 billion points or bases: of these, about 3-4 million differ from other human genome sequences. While the great majority (around 99.9%) of an individual's genome is identical to the genomes from all other people, these differences, known as genomic variants or variations, occur across the genome, and can be associated with populations with different genomic ancestries or genetic diseases. These variations can be used to distinguish between people, but the similarities between genomes mean that genetic variations have a limited role[2]. These variations can be in the form of single point changes – single nucleotide polymorphisms (SNPs) or larger sequences. The academic literature often talks of genomic data or genomic privacy but it is not always clear what data are at issue. Access to different data can have varying implications as some types of data are richer than others: the information that can be derived from an entire genome sequence may be very different from that derived from a few dozen SNPs. But the ability to identify someone is dependent on the relationship between the characteristics of the data and the techniques used to interrogate those data (see chapter 3). The richness of data is relevant to anonymisation only to the extent that it facilitates identification; sometimes relatively sparse data can be used to identify someone if it is combined with other data.

From a policy perspective the most important issues to consider are those characteristics that cause people to assign special meaning to genomic data. Genomic data are sometimes treated as 'exceptional': intrinsically unique as compared to other types of data. Our position is that genomic data are not exceptional but sometimes do possess characteristics that challenge the law and policy relating to anonymity[3]. Perhaps the four most important characteristics are that genomic data can be 1) strongly identifying, 2) familial, 3) static, and 4) commonly understood to be special.

1. Genome sequences are often claimed to be inherently identifying, 'the ultimate identifier' or otherwise very strongly identifying[4]. In this view, genome sequences are a type of data so bound to the identity of the individual that identification is unavoidable given sufficient analysis. This is an exaggeration, as there is no single type of data viewed in isolation from other data that completely identifies a person. (We discuss the nature of identification in greater detail, in chapter 3.) However, genomic data (in general) can be highly identifying when combined with other data. Identifiability is dependent on many factors, mainly the kind of genomic data at issue and the type of data with which they can be combined. A whole genome sequence is 'more identifying' than several dozen SNPs, but both are technically genomic data and both can identify a person given the right context.

2. The genomes of one person can reveal information about someone closely related to them. They are therefore familial data. Depending on the context, those data could reveal a person's biological paternity or their susceptibility to certain diseases[5]. Genomic data may therefore reveal things about people other than the person from whom they were derived. This creates more challenges to the release of genomic data than for some other types of data. For example, consent, which is often regarded as one of the cornerstones of lawful data use, becomes a weaker justification to the release of genomic data because family members may not have a say in decisions made about those data. Thus both consent and making data non-identifying are weakened by the nature of genomic data.

3. Genomic data are relatively static because they remain relevant to the individual to whom they relate over long periods of time, even between generations[6]. This means the value of genomic data – medically, scientifically, and economically – is likely to increase over time because the information we are able to derive from studying those data will improve. For example, disease susceptibility estimations based on the same data are likely to get better as our knowledge increases[6]. The consequences of the release of genomic information are therefore not limited in time[7] and apply equally to positive developments as well as to new threats. There are limits to this because the methods used to obtain sequence data, for example, will themselves improve[8]. A genome sequenced in ten years' time may contain different data to one sequenced today (it could be more complete, for example have greater depth and coverage), and genomic data relating to a person can change to a certain degree in some contexts (tumour DNA for example, have greater depth and coverage). But data released now will continue to have relevance almost indefinitely: the presence of newer and richer data does not make the older and poorer data worthless.

4. That genomic data are believed to be special is, perhaps tautologically, one of the characteristics that cause genomic data to be treated as unique. This belief has been described as a "*mystique*"[8]. This mystique is important even if it is non-empirical; beliefs play a significant role in informing policy-making.

Genomic data have other important characteristics. Genomic data:

- Can reveal information about susceptibility to diseases and other physical conditions[9]
- Contain information about ethnic heritage[10]
- Have a physical and digital existence,[6] and in its physical form can be obtained relatively easily from a variety of sources (blood, hair, skin etc.)[7]
- Are high resolution[11] and have many dimensions[6]

3. Challenges and limitations of anonymisation

In Western countries, privacy and anonymity are usually considered worthy of protection, but are not absolute values and are subject to other contending factors such as public interest, societal or economic value in the use of data. Where interests collide, the response is often to obtain consent from the person to whom the data relate or to make those data incapable of intruding on privacy or anonymity. But making data truly anonymous is difficult and this is especially the case with respect to genomic data.

The challenges of identification

A number of techniques can be used to make the identification of people within data more difficult: however these are sometimes challenging to implement successfully, especially when applied to genomic data. There has been a tendency in law and policy to treat such techniques as more robust than they actually are but, as one notable commentator put it, "*as the utility of data increases, the privacy decreases*"[12]. The law takes insufficient account of the weakness of some of these techniques and public policy and discourse are insufficiently transparent about those weaknesses. This problem is well known and there is a large literature on the issue, often as part of a wider debate about the use of data in general (and health data in particular)[13]. Some of the challenges have drawn the attention of senior and non-specialist policy makers and there are claims that in policy circles there is an increasing acceptance that "*irreversible anonymisation of meaningful data is practically unattainable*"[14]. This acceptance is not properly reflected in law or policy, in part, due to the changing context: statistical methods have developed quickly, more data are available (including genomic data), and public attitudes have changed. Twenty years ago, being identified and losing privacy were not common fears as they are now[15].

Paul Ohm, a legal academic, was one of the first to discuss outside technical fields of study the threat posed by the ability to identify individuals using supposedly-anonymised data. He neatly summarised the problem: *"Reidentification science disrupts the privacy policy landscape by undermining the faith we have placed in anonymization"*[12]. This is a useful summary of the general problem, but a more specific set of problems relate to genomic data.

In Western countries, laws relating to the use of data tend to encourage masking identities within datasets unless people have agreed that those data may be used in a way that could identify them. This reflects the commonly held view that it can be ethically problematic to use data to identify a person without their permission, especially if their inclusion within a particular dataset reveals something important, or sensitive, about that person. For example, data about hospital admissions, health records or internet searches might reveal varying amounts of potentially sensitive information. People may not wish to be identified based on those data and their reasons for this may vary. There might be an observable and measurable 'harm' that befalls the person as a consequence, such as

"erroneous or malicious identity disclosure and consequent embarrassment; legal or financial ramifications; stigmatization; or discrimination for insurance, employment, promotion, or loans"[16]. Alternatively, the harm might be something more intangible, like violation of dignity or personal or familial identity[17]. Some reasons may be stronger or more reasoned than others, but ultimately it ought not to matter – the law reflects and manifests (to some degree) the view that identification should sometimes to be avoided.

Traditionally, the response to this problem has been to either obtain permission to use those data from the person or to manipulate the data to prevent them being used to identify people. This approach is known as 'consent or anonymise',[14] and there are many possible ways to anonymise data. Historically, anonymisation has been seen as something of a panacea but it has now become more difficult to secure, as sophisticated data analysis has made it harder to prevent people from being identified.

The law must therefore take into account the perceived utility of the data, risk of identification, the security of the data and difficulty of achieving anonymisation. The law seeks to protect against the perceived harm of identification while allowing data to remain available for use, because of the possible benefits: public health benefit, economic growth, or national security, for example.

In identifying and describing these challenges, we are not claiming that the system has completely broken down – broadly, it continues to function: data moves, people conduct research and provide care, and the likelihood of identification remains low[18]. But the system does not necessarily function rationally, and harms – possibly unidentified – may still occur as a consequence. For our purposes, there are two significant dimensions to these challenges: how to manage the complex relationship between the many interests at stake (e.g. protecting privacy, furthering medical research, and improving healthcare); and whether there is something about genomic data that makes a loss of anonymity or privacy more likely when compared to other data.

Discussion of how competing interests are managed is outside the scope of this document. However those interests underpin how people approach this topic and so it is worth noting that these interests tend to be understood in terms of the balance between private and public interests (with 'private interests' referring broadly to the interests of individual persons rather than commercial entities). For example, the private interest in maintaining the privacy and anonymity of the individual against the public interest in the furtherance of medical research. This view tends to see public and private interests as inherently opposed. A more nuanced position recognises the possibility of private interests in the public good and a public interest in private goods - e.g. an individual has a private interest in fostering the public good because they are a part of that public. Likewise, there is a public interest in fostering private interests because doing so helps maintain the trust necessary for a functioning society. The Nuffield Council on Bioethics notes *"we all have interests on both sides, private and public, as individuals, members of families, groups, communities and nations"*[14].

The challenge

- Identifying associations between data can reveal the identity of a person to whom those data relate (breaching anonymity) or reveal facts about a person whose identity is already known where they would rather that did not happen (breaching privacy)
- Once a person has been identified, the opportunity exists for further information to be revealed about them[19]
- In many circumstances revealing identity is considered wrong (hence the fact that the protection of anonymity in medical research and practice is often enshrined in governance)[20]
- Genomic data can be a powerful way of revealing those associations and it is difficult to prevent them from being used to do so
- The characteristics of genomic data that give rise to this are the very characteristics that make them medically and scientifically useful
- Much of the study and use of genomic data relies on giving multiple people and groups access to those data and there is a push to increase this access in order to improve health and scientific research
- There has been a large increase in the amount of genomic data generated and used because the associated costs have fallen steeply
- There has also been a marked increase in the availability of other data from many different sources[21]
- Increased availability increases the risk of those data being accessed, disseminated, used, or linked. **This increases the risk of breaching anonymity or privacy**
- To allow for the extraction of value from data (medical, scientific, commercial) while protecting the interests of those to whom the data relate, the law often depends on consent or the manipulation of data to mask the identity of those to whom the data relate[22]. **The latter is often known as anonymising data**
- It is possible to fulfil the legal requirements of anonymisation without making the data technically anonymous; data do not have to be 100% anonymous to be freed from the legal restraints applied to identifying data
- The nature and increasing availability of genomic data threatens this approach because it is difficult to make genomic data non-identifying. This challenges the accepted degree of anonymity
- Attitudes are changing towards privacy, anonymity, and genomic data specifically; the societally accepted balance between risk and benefit is in flux[21]
- Therefore, there are significant differences between what can be done to protect genomic data from potential breaches of privacy or anonymity, what the law requires and accepts with respect to such protection, the needs of science and medicine, and what is more widely acceptable
- **Consequently, a concerted effort must be made to reach a general consensus on appropriate regulation, in ways that respect cultural change and that manage expectations**

All of these issues are pertinent to genomic data. Genomic data are extremely useful, medically and scientifically, because of the characteristics listed above (especially that they are predictive, high resolution and of high dimensionality). There are considerable benefits to be derived from the free-flow of genomic data and extensive linking: transfer of access to genomic data is needed to achieve those benefits[16]. For example, diagnosis of rare genetic diseases often necessitates a comparison between the patient's own genomic data and pre-existing data from unrelated patients with the same or similar disorders[23].

There is also a strong political and economic drive to release and link genomic data. It is common for genomic data to be deposited in shared databases in order to improve their utility and exploitation – such deposition is sometimes a funding requirement[24]. But genomic data are sensitive and unusually difficult to keep secure using common techniques because they are strongly identifying, familial, and static.

Anonymisation methods are not absolutely secure

There are many different methods used to try and prevent people being identified from data, but there are two broad technical approaches: manipulating data and restricting who can gain access to the data. The former can be statistical (modifying the data themselves to prevent them being used in a particular way) or cryptographic (encoding them to prevent unauthorised use)[25]. Both can provide mathematical proofs defining what can and cannot be inferred by the person in possession of the data[25]. These techniques might be undermined in different ways, and we discuss some examples in chapter 6. Controlling access to data relies on developing and applying systems and processes that mediate who can access data and how that access takes place. For example, requiring that the user is a bona fide researcher or (relevant) clinician, or requiring a potential user to access the data only from a particular physical location.

These approaches are not mutually exclusive and no system is ever completely secure. Genomic data poses a greater challenge to those techniques that rely on simply modifying data and then releasing widely ('release and forget'[12]) than techniques that actively moderate access. Statistical techniques suffer because the genome is strongly identifying and because much information that is available elsewhere can be combined with genomic data in order to draw identifying associations even if those genomic data are stripped of identifying information: the more available data, the greater the likelihood of accessing the right combination of data to identify a person.

In the UK, and other EU Member States, this is especially relevant because data protection law currently applies only to data that relates to an identified or identifiable individual – and this will remain the case under the EU GDPR to be implemented in May 2018. Laws around the world make allowances for 'non-identifying' data. All of this leads to the fundamental problem identified above: the imperative to disseminate or link genomic data in order to obtain their full scientific, medical, and economic value is threatened by a combination of the regulatory framework and the characteristics that make those data so valuable in the first place. Thus there is an inverse relationship between the overall usefulness of genomic data and how difficult it is to use those data to identify a person[26]; the less identifying, the less useful.

Whilst there have been many instances where hackers have successfully gained access to other types of sensitive data[27], there have been no reported, significant, and successful real-life attempts to breach privacy or anonymity using genomic data[28]. But academic studies have shown that it is possible to do so (we discuss this in more detail, below). It is likely however that the risk of such an attack will increase as genomic data become more prevalent[29], prevalence increased by the falling cost and difficulty of generating those data and the inroads genomic medicine is making into mainstream medicine. Not only that, when genomic data are placed in the public sphere they are almost impossible to retrieve[11], whether they were put there deliberately, accidentally, or maliciously. Because of this, there have been a number of investigations into techniques used to protect genomic data, some of which have resulted in changes to policy. (We discuss these issues in more detail in chapter 4.)

One of the difficulties of regulating this area is predicting what data are likely to be available to a person attempting to identify someone. Yet ultimately this is necessary to determine identifiability under UK law. When developing attack and countermeasure scenarios, those who specialise in the technical aspects of data security often assume the availability of those data relevant or necessary for the attack[12]. This is partly a function of how security measures are assessed, but as a position it is at least to some extent supported by the relatively static nature of genomic data: any data previously released continues to be useful. But the law is not single-mindedly focused on the security and privacy of data – other issues are in play, such as protecting national security, and encouraging scientific research, economic development, or improved healthcare.

4. Techniques and terminology

There are many ways of protecting the identity of people within datasets, for example, changing the data or limiting who can use the data. Sometimes these techniques can be difficult to apply to genomic data. Therefore, simply restricting how people access genomic data is more effective in protecting privacy while retaining the usefulness of the data. No single piece or type of data are capable of identifying someone, and laws based on this idea are therefore fundamentally flawed. The potential harms associated with identification and individuation should be carefully distinguished.

How are data 'anonymised'?

There are a number of methods used to prevent identification of an individual within or by a dataset: some change the data in some way and others change the way it is accessed. The focus of this chapter is on techniques that change the data or facilitate the wider release of those data, and include anonymisation, pseudo-anonymisation, obfuscation, and aggregation. Other methods attempt to prevent people accessing the data, and include cryptography and data access controls[30]. Cryptographic techniques seek to make the data meaningless to those without the right way of accessing them and data access controls restrict who, when and how these data can be accessed.

In an information-rich society like our own it is difficult to guarantee that people will not be identified using their genomic data if the data are accessible. As the use of genomic data expands beyond research and health, the problem will be exacerbated[31], making genomic data more readily available to groups with motives and standards different to those traditionally found in the research and health sectors. Those sectors employing 'release and forget' protections seem likely to pose the most significant threat to anonymity[12].

Anonymisation and pseudo-anonymisation

'Anonymisation', 'anonymity', 'anonymised', and 'anonymous' have different meanings depending on context. For example, 'anonymisation' refers to a particular technical process necessary to achieve anonymised data. A person referred to by those data has been and is anonymous to the *extent that the process allows*. This technical process is usually associated with the removal of specific identifiers such as name, date of birth, address etc[14]. Alternatively, 'anonymity' might refer to the state of actually *being* anonymous, regardless of technique used and so 'anonymisation' would refer to the process of achieving that true anonymity. Or the process and state might be simply legal in nature, rather than technical or actual – anonymous to the extent the law requires.

Confusion is inevitable when using a term with so many different interpretations: a lay-person might think the term describes the state of true anonymity, a lawyer merely that the requirements of the law have been fulfilled, and a statistician something far more mathematically nuanced. Interpretation of 'anonymous' is also complicated by confusion about whether it means 'without name' or 'unidentifiable'. What seems clear is that researchers and policy makers often conflate the notions of de-identification and anonymity: whilst the latter is 'an absolute state', the former is 'a process that seeks to mitigate disclosure risk through careful application of rules and statistical analysis' and the consequence of this conflation and its downstream effects is profound[32]. (We discuss these issues in more detail, below on pages 20-23).

Pseudo-anonymisation is a way of distinguishing individuals within datasets without identifying them[33]. It replaces with codes the identifiers that anonymisation removes. These codes are associated with an individual using a 'key' that is kept separately: this allows particular parts of data to be linked with a person without them being explicitly identified and is easily reversible by those who hold the key[14,34]. Pseudo-anonymisation is sometimes known as 'coding' and pseudonymised data as 'coded data'[35].

Anonymisation as a process was one of the earlier methods suggested as a way of protecting people from being identified by and within genomic data, but has been proven unsuitable in some circumstances[36]. In general, in this report, when we refer to anonymous data we mean 'data that cannot be used to identify a person', unless otherwise noted.

Obfuscation

Obfuscation changes some part of a dataset with the aim of preventing people from interpreting it in specific ways. In the context of genomic data, this might include adding incorrect data about SNPs, or hiding data about them[36]. However, some methods of obfuscation, such as adding fake values, tend to reduce the accuracy of genomic data, making it less useful[37]. This is particularly problematic given the health-related nature of genomic data and the attendant need for accuracy[8].

Aggregation

Aggregated data are data about groups of individuals presented either as averages or frequency records (listing the number of instances of an event or result type). They show trends and general values[38]. This is done in an attempt to avoid revealing any particular individual within those data. There are a number of ways of doing this, such as supressing values that include less than a certain number of people, or perturbing data by randomly altering values by a certain amount. Frequency records are inappropriate for genome sequence data because every record is unique.

Cryptography

Cryptographic techniques, which include encryption, transform data into a form unusable by anyone who does not have the right key[30]. Encryption techniques are the most recent to be suggested as a way of protecting genomic data[36]. Where data are encrypted, they are usually considered safe if decrypting them would cost more than the value of the data, or if there is no longer value in the data by the time decryption has been achieved[6].

There are limitations to cryptographic techniques with respect to genomic data:

- Encryption systems are – in time – broken, but genomic data are very unlikely to lose their value over the same period of time[39]. Given the familial nature of genomic data, even protections that last the length of the life of the person to whom the data relates may not be enough and the best techniques are likely to be broken after approximately 30 years[40]
- Cryptographic techniques usually make it harder to work with genomic data: data analysis tools applied to encrypted data tend to be less efficient, computing and data storage needs are problematic and the data cannot be directly viewed by the user[41]

Data access controls

Data access controls are ways of restricting data access only to authorised people[30]. Evidence may be needed that the person accessing the data is a bona fide researcher or (relevant) clinician; sometimes the person using the data cannot keep or copy it, there may be no access granted to 'individual-level' data, and various protections can be applied to prevent or discourage unauthorised access, such as tracking user behaviour[42]. Where very high levels of security are needed, data access systems may prevent data being accessed via the internet[42].

How do we understand anonymity, anonymisation and identification?

Different legal and academic interpretations have resulted in a complex and fragmented terminology. The various meanings ascribed to different words confuse discussions and cause difficulties in setting up frameworks for the use of data. As alluded to above, the meaning of 'anonymous data' under EU data protection law may vary significantly from the understanding of a non-expert member of the public to that of a technical expert.

Developing a standard lexicon and a list of concordant terms would be a useful starting point in resolving this problem, and some organisations have already attempted to develop such a list[43]. In the short term, the most useful thing such a list could achieve would be to encourage transparency about the difficulty of reaching genuine consensus. It is unlikely that there will ever be true harmonisation of terms between different jurisdictions, and local regulations will always need to be followed, but recognising the limits of the discussion is useful in itself. We do not attempt to recreate other glossaries, but instead focus on the meaning of a few key terms that in their ambiguity cause confusion[44].

The meaning of anonymity is highly contested, especially as it relates to privacy. They are not the same, but they are related. Two useful ways of understanding anonymity and privacy are:

1. As part of different 'sets'
2. As inversely related ways of hiding a personal fact (subject/predicate)

These approaches reach broadly the same conclusion but are worth considering separately because examining their reasoning can offer useful insights.

Anonymity sets and privacy sets

The Nuffield Council on Bioethics defines 'technical anonymity' and the relationship it has with privacy by describing the interaction between different groups (or sets). Under this rubric, an anonymity set is a group comprising all the people who might be confused with the person to whom the dataset does truly relate. For example, if the data provide the name 'Alice', the anonymity set becomes all people called Alice. In this scenario, a privacy set comprises those people which the person Alice, requires to remain in ignorance of a specific piece of data. Privacy sets are contextual, following the circumstances of the person concerned; they are subject to change.

In this understanding of anonymity and privacy, anonymity is broken *"if the anonymity set is reduced to one from the viewpoint of anyone in the privacy set"*[14]. This will happen if the dataset is available to someone in the privacy set and they have the desire and ability to access the data in a form that would identify the person (i.e. narrow it down to one person).

Subject/predicate

Jeffrey Skopek, an influential legal scholar, explains both 'anonymity' and 'privacy' as states of affairs in which a personal fact is unknown to others, with each term referring to a different way of preventing people from learning that personal fact. Skopek notes that a personal fact comprises two elements: a subject (which identifies a person) and a predicate (which informs us of a fact about that person). To prevent a person from learning a personal fact, one can either hide what makes it personal (the subject) or hide the fact (the predicate). Under this view, **privacy is a condition in which we know a subject but not a relevant predicate and anonymity is when we know a predicate but not the relevant subject.**

Skopek illustrates this using an example of a blood test result on a piece of paper in a medical file. If the piece of paper is removed and access is subsequently gained to the patient's medical file, that patient has privacy in relation to the blood test result, but not anonymity. If access is gained to the piece of paper alone and unconnected to the medical file then there is a state of anonymity but not privacy. This is why anonymity is sometimes confused with privacy: it protects privacy because the one entails the existence of the other. Because of this, if access is gained to the test results but they cannot be associated with a person then the test results will be anonymous and as a result the privacy of the patient will be protected with respect to the results. Conversely, if access is gained to the medical file but the test results are not associated with the file, the patient's privacy will be protected, and so the results are anonymous[45].

This example illustrates the fundamental point: whether something is a subject or a predicate is not based on the type of information but on the relationship it has with a given piece of knowledge. Anonymity and privacy are each therefore reflections of incomplete knowledge of a given piece of personal information: knowing a thing about a person requires knowing both the subject and the predicate.

Identifiability, anonymity and associated concepts

The most obvious conclusion to draw from both of these approaches is that anonymity is not the same as 'without name', despite its etymology. Being without a name is not enough to be anonymous and being named is not always enough to stop an individual being anonymous (more than one person can have the same name). Instead, a better way of thinking about it is to understand **anonymity as a state of being unidentified**. It is not necessary to know a person's name to be able to determine who they are. This is unsurprising – even obvious – but it reminds us that the concept of anonymity is caught in the extremely complex web of factors that define identifiability.

From the above, we can see that a state of being unidentified is not based on the presence of a particular type of information because there is no one type information that is inherently identifying. There is no single type of trait that can be removed to ensure a person is unidentified; there is no such thing as a perfect identifier. Instead, what identifies someone is a combination of the uniqueness of the data and the nature of the connection between different data. Identifiability is the outcome of a network of associations: datasets that allow more connections to be drawn more easily are more likely to result in identification.

In the context of genomic data an apt comparison might be a very common genetic variant which is found frequently within the population. Here, knowing the precise configuration of the variant will be neither identifying or individuating as it is shared with a large percentage of the population. If however, a very rare variant is identified, if the genomic information alone is disclosed, this might be done anonymously, but linkage with specific clinical signs and symptoms (phenotype) particularly on a publicly accessible database might make identification more likely. This is one of the reasons why genomic data are important – the depth and complexity of genomic data make them very rich in terms of the way they can be linked to genomic and to other data (depending on how genomic data are conceived).

Where information is not unique it is sufficient to ensure that the person to whom it actually relates remains unidentified. This is because it cannot be used to single out an individual. A piece of information must be unique in order to identify someone, but that uniqueness is not enough to ensure that person is identified. Uniqueness is therefore necessary but not sufficient for identification to take place: *"identification requires more than individuation"*[45]. It is connectedness that makes a fact associated with a specific person.

Because of this, deleting or disassociating a particular type of information cannot necessarily prevent identification. Data that are capable of extensive connection with other data are more powerfully identifying than data that have minimal connections, but those connections are not completely reliant on the presence of specific properties or data type. Some data types may be more likely to provide extensive connections, and therefore be more identifying, but their removal does not make the data inherently incapable of identifying people. Since the ability to identify someone exists on a spectrum, being unidentified – being 'anonymous' – is not a state of being unidentifiable but instead a state of being unidentified in a particular context[45]. This is important to recognise and is the correct way of understanding both terms. But the fact remains that the term anonymous is in widespread use, often with little in the way of clarification, and much data-relevant law fails to properly recognise that someone cannot be identified using a single piece or type of data.

5. How are data governed?

Data are categorised into different types according to the 'sensitivity' of the data, with protections dependent on the category. Data classified as anonymous have very few protections and are not subject to data protection law. There is no single standard defining legally anonymous data in the EU, but data do not actually have to be anonymous to be legally classified as anonymous. Guidelines and case law do not perfectly align in what they require for anonymity to be formally achieved, which can lead to confusion over how to proceed.

In Western countries data-relevant laws usually establish systems that simultaneously limit the risk to people's privacy and encourage the benefits of data use[11]. This is particularly the case with research and the provision of medical care to individuals and populations, partly because the data at issue are regarded as sensitive, and therefore deserving of greater protection than other types of data. Some see these protections as inappropriately hindering research and the delivery of healthcare[34].

In the context of genomic data, the two most important aspects are how the law approaches types of data and how anonymisation is achieved.

How does the law treat data?

Legislation often categorises data into various types and affords different protections to these categories and this is how EU data protection law works: 'personal data' are protected, and some types of personal data are given additional protections. To qualify as personal data, data must relate to an identified or identifiable person[46]. This means that if data are classified as anonymous (non-identifying) those data are not technically personal data and fall outside data protection law. It also means that personal data must only be potentially identifying – a person does not have to actually have been identified, merely capable of being identified by those data[47].

Health data, a category that almost certainly includes genomic data, are usually given stronger protection than less sensitive data. However, under EU data protection law, there are a number of situations that permit data to be used. Two of the most important of these are:

1. That the person to whom they relate has given their consent
2. That data are classified as anonymous (determining this is not simple – see pages 26-27)

There are also significant exceptions for the use of data for research and healthcare purposes that allow data to be used in situations where these two criteria have not been met. Data protection law does not exist to obstruct legitimate research and the UK research community is mostly compliant with the research exemptions that apply in data protection law[20].

Genomic data are likely to be classified as health data in any jurisdiction that has such a category and are sometimes afforded even greater protection than normal health data. But such categorisations are intrinsically ambiguous. There is no definitive, and satisfactory, definition of health data in UK-relevant legislation. The Article 29 Data Protection Working Party, an authoritative EU organisation that offers non-binding advice and guidance, describes three forms of data that can be classified as health data:

1. Inherently/clearly medical data, such as medical histories or data about the physical or mental health status of a data subject that are generated in a professional medical context[48]
2. Raw sensor data "*that can be used in itself or in combination with other data to draw a conclusion about the actual health status or health risk of a person*"[48]
3. Conclusions about a person's health status or health risk (irrespective of whether these conclusions are accurate or inaccurate, legitimate or illegitimate or otherwise adequate or inadequate)[48]

This approach has been broadly reaffirmed by the recent EU GDPR: data are still separated into these kind of categories, but genetic data now have a separate definition outside of health data (although it is related and likely includes genomic data)[49]. Although anonymous data remain outside of the meaning of personal data, personal data which have undergone pseudonymisation, which '*could be attributed to a natural person by the use of additional information, should be considered to be information on an identifiable natural person*'[49]. There is considerable uncertainty about the scope of the pseudonymised data which will fall within the GDPR.

Finalising the GDPR took a long time and was extremely complicated and costly. We are unlikely to see significant change to data protection law at the EU legislative level in the near future – the GDPR is likely to form the basis of EU law for a long time to come.

Following the Brexit vote, the UK plans to implement the GDPR, and this has been reaffirmed in policy documents and in draft legislation[50, 1]. In the longer term, the scope and effect of data protection law is uncertain although there are strong justifications for retaining a harmonised regulatory framework in order to support harmonised health care provision and medical research.

The sense behind this kind of data categorisation breaks down when faced with predictive analytics that can produce category-relevant knowledge from data that would not, in a common-sense view, be categorised in the same way as that knowledge. Techniques that allow the drawing of unexpected inferences challenge the boundaries between the theoretical constructs developed to manage what can and cannot be done with data. The now-famous and controversial example of the US company Target correctly predicting a girl's pregnancy based on her not-obviously-pregnancy-related shopping habits shows the difficulty of categorising data in this way[51].

If data on shopping habits can predict pregnancy with a reasonable degree of accuracy, should those data be classified as health data and thus be given the same protection as whole genome sequences? That seems counter-intuitive. Yet common concerns about the use of data, especially health-related data, are unlikely to be based on the incoherence of the legal categorisation of data. Instead, they are more likely to be based on fears about what can and cannot be done with data and whether the law gives people enough protection from harm.

Categorisation matters in terms of anonymisation and identification because of the ability to draw category-related inferences from data legally outside that category, like the pregnant girl in the example. Data not subject to the stricter provisions of data protection laws can still be used to reveal things that relate to the more strictly protected categories, and this includes being used to identify individuals within datasets. Add to that the fact that genomic data can often not only identify a person but reveal much about health-related issues, and it is clear why the intersection between genomic data and data protection law causes friction.

What counts as 'anonymous data'?

As noted, under current and future data protection law, non-identifying data are not subject to the principles of data protection. On the face of it, the term anonymous or non-identifying might feasibly refer simply to obtain a state of technical anonymity, but this is not the case. Instead, it has a specific legal meaning.

EU and UK legislation defines anonymous data, although the UK approach currently differs from the EU law on which it is based. Under current EU law, account should be taken of all means 'likely reasonably' to be used to identify the person to whom data relate[46]. The General Data Protection Regulation refers to *"means reasonably likely to be used… to identify the natural person directly or indirectly"* and this will be the law across the EU by May 2018[49]. UK law refers to how 'likely' identification would be[52] (although UK courts have sometimes used the 'likely reasonably' test), but the UK Information Commissioner's Office (ICO) notes that the practical problems that arise are much the same regardless of which test is used[53].

Although EU-level and UK law defines what constitutes anonymous data, there are no legislative definitions that describe exactly what must be done to make data formally considered anonymous. The system in the USA is somewhat different, as there is a more formal system of guidance relating to the requirements of the Health Insurance Portability and Accountability Act[54].

Many factors may be taken into account when determining what is necessary to achieve a state of anonymity: security and non-disclosure safeguards, removing obvious identifiers, coding, firewalls, and aggregating data have all been given as examples[17]. In the UK, there are a few authorities providing guidance on anonymising data (see for example guidance from NHS Digital and the ICO) including recent advice on anonymisation and big data[55].

The most important thing to understand is that to qualify as anonymous, data do not have to actually be anonymous. The ICO notes that it is not necessary to achieve *"100% anonymisation"* – the UK regulatory framework is concerned with the probability of identification, not the possibility that it may occur[53]. Therefore, fulfilling the legal requirements for anonymisation does not necessarily ensure either privacy or anonymity, although it may reduce the probability of identification to extremely low levels depending on the context. There are multiple techniques available for achieving legal anonymisation and they apply differently depending on context (for example, whether the data are published)[33]. (See chapter 4).

Although there is no legislative description of what processes must be undertaken to achieve legal anonymity, there is a legal test to determine whether data are anonymous. The test seems relatively simple: in the UK data are considered anonymous where anyone receiving those data would be unlikely to be able to identify any of the individuals to whom the data relate, taking into account all the means likely (or likely reasonably) to be used[56]. In some circumstances 'anyone' probably means people identifying themselves within those data, even where no one else is capable of doing so (such as a person identifying a unique genetic variant within a database)[23]. Essentially, the risk of identification must be greater than remote in order for data to be considered identifying[53]. Under UK law, the test applies only to people or organisations other than those prospectively releasing the relevant data. This means that the data are considered anonymous even if the person or organisation releasing them can identify people within those data. The interpretation of 'likely' remains ambiguous and is defined in one UK case as *"extremely remote"* [57] and under another as simply *"on the balance of probabilities"* [58]. These are very different standards, and do not fully accord with the guidance of the ICO and the Information Governance Review, which take the view that non-identifiable data are those data in a *"form that does not identify individuals and where identification through its combination [sic] with other data is not likely to take place"*[53].

How does the nature of genomic data affect its treatment by the law?

Genomic data do not lose their value or identifying strength over time and as time passes more data tends to become available and better linkage techniques developed. This poses a practical, if not necessarily legal, threat to the robustness of the test. It seems common sense that the criteria used to determine 'likely' should be anchored in the time period in which the data are released – in this context, attempting to predict the future would be futile. But equally it seems perverse to ignore the fact that it will almost certainly become more likely over time that someone will be identified within a given dataset.

Although not intrinsic to the nature of genomic data, the uses to which they are put – in industry and academia – mean that those data frequently cross jurisdictional boundaries. This raises a host of legal issues, as jurisdictions have different rules that are sometimes incompatible[59]. It can be confusing to know when and how which rules apply.

6. How can anonymisation techniques be undermined?

A number of techniques are used to identify people using genomic data that has been manipulated to disguise identities. These techniques tend to rely on accessing other data in addition to the datasets being attacked, sophisticated knowledge of associations between single nucleotide polymorphisms (SNPs) and physical traits, and the small number of SNPs necessary to uniquely identify a person within a dataset. This chapter summarises the methods used in the most well-known attacks but useful and highly technical reviews are available elsewhere[60].

Defeating anonymisation using pre-existing genomic data

In 2004, it was shown that it is possible to identify a person within some genomic datasets using only very small amounts of genomic data from the person to be identified. By comparing an individual's genomic data to publicly available SNP data, it was found that in some circumstances a person could be uniquely identified within those data using only 30-80 SNPs[61]; subsets of approximately 300 common SNPs could uniquely identify any person[60]. Commentators noted *"In such a case, the rest of the genotypic, phenotypic, and other information linked to that individual in public records would also become available"* [62]. Later work determined that approximately 45 SNPs, carefully chosen, could identify people within most populations[63]. Once a person is identified, their relatives can also be targeted[60]. These were important developments: they showed how little data was necessary to identify a person, and also that randomly altering a percentage of SNP data to prevent this kind of technique from working would be inappropriate because the required percentage would be high enough that the randomisation would make the relevant data far less accurate and thus less useful[62]. As a consequence, it has been claimed that most genomic researchers agree that raw DNA data ought not to be published in online repositories without proper agreements in place, because such data are thought to be too liable to this kind of attack[8]. Following the publication of these methods, 'pooled' summary statistics were seen as a possible solution, but that method has its own weaknesses[8].

The obvious limit to this approach is that it requires access to the DNA data of the person to be identified.

Defeating aggregated data with a pre-existing sample

In 2008 a team in the US showed that it was possible to identify whether a person was included in a particular study even when the data at issue comprised aggregate SNP variants of the people in the study[64]. The work was originally intended as a forensic method and applied to physical samples[65]. Depending on the nature of the data, identifying a person within an aggregated dataset may be an attractive proposition because their presence in that set may be indicative of them developing a disease[66]. This technique compares data about the SNPs of the person to be identified against data about SNP variation frequencies in the aggregated study data and a reference population available from public sources that are known to exclude the person to be identified[67]. A sufficient number of SNPs makes it possible to identify whether the person is in the study population or the reference population and to identify the person's data within the aggregate data with a high degree of confidence[66]. The authors needed a profile of 10,000-50,000 SNPs[68]. Those seeking to identify the person must therefore already have a SNP profile derived from a DNA sample.

There are limits to the applicability of this technique:

- The reliance on a SNP profile of the person to be identified is an obvious limitation, as it requires a sophisticated analysis of the DNA of that person. Although the cost of obtaining those analyses are dropping, making it at least feasible, it still needs a significant amount of information that can be difficult to obtain[66]
- Although strong theoretically, the actual risk to study participants is limited in practice because many of the assumptions made are not borne out in real-life examples[68]

The publication of this technique prompted a reduction in open access to the relevant kind of aggregate data, with both the Wellcome Trust and the US National Institutes of Health (NIH) removing those data from public websites[65, 69]. Researchers must now sign an access control agreement not to attempt identification. These agreements can take several months to complete[8].

Refinement using 'linkage disequilibrium'

Later work by other teams refined the original technique to make it more powerful by using different types of data or reducing the number of SNPs needed. For example, in 2009 another team of researchers in the US showed that it was possible to use a somewhat similar attack that needs far less information. The technique relies on 'linkage disequilibrium' (a measurement of the non-random correlation between SNPs). This allows the needed data – the allele frequencies – to be inferred from those data that are released. This eliminates one of the key weaknesses of the above technique and allows a person to be identified within a genomic dataset using far less data[8]. It is also possible to use this method to identify people even without having their genomic information as a comparator: if the data associated with a particular person has a sufficiently robust 'identification confidence', and that information contains information about observable phenotypes, phenotypic trait correlation attacks can be used as an identification method[66]. (See pages 31-32.)

Applying knowledge of linkage disequilibrium is not only used to identify people within large datasets: it was also used famously to predict the presence of the *APOE4* gene in James Watson's genome. *APOE4* is strongly associated with the development of Alzheimer's disease, and Watson had specifically ensured that the data relevant to that gene was not included in his published genome[26]. This demonstrates well that one cannot know what others will know in the future and so redacting a small portion of the data may not be sufficient.

Defeating anonymised data using publicly available data, without a pre-existing sample

In 2013, it was shown that it is possible to identify people from personal genomes published online in anonymous form. The genome data were stored in the open access database of the 1000 Genomes Project and the European Nucleotide Archive and the team responsible did not have access to a DNA sample of the person eventually identified unlike the methods described above[70].

Identification was achieved by combining the published genomic information with data found in other public databases: 'Y-chromosome short tandem repeats' were analysed and triangulated with the other data, such as recreational genetic genealogy databases, age, and location (US states, in this case). 'Y-chromosome short tandem repeats' are small sections of repeating DNA found on the male chromosome that are inherited by the son from the father. The variations are very similar between generations. This, combined with the fact that surnames are inherited via the father in many societies, means that there is a correlation between surname and Y-chromosome short tandem repeats. These repeats used in genealogical testing and included in online genetic genealogical databases and surnames are a powerful way of identifying people[60]. They can also be easily searched[60]. By analysing these sections of DNA and triangulating with the other data mentioned, it is possible in some cases to uniquely identify the people to whom the genome data relate.

The genome data complied with the relevant legal anonymisation requirements, however, in this case, those requirements were not enough to prevent identification[71]. The team that revealed the threat predicted that the risk of surname inference will grow as the technology to read the relevant genomic data improves. The technique used by this group has a number of limitations:

- The technique is difficult: extremely specialised tools and knowledge are used to analyse the data[60]
- The method is not perfect: there are a number of ways to break the link between surnames and Y-chromosome short tandem repeats: for example, spelling variants for surnames, 'non-paternity events', and genetic mutations[72]
- The method does not apply in all contexts: not all societies have strongly identifying or paternally inherited surnames[60] and at the moment most genome sequencing studies do not report on Y-chromosome short tandem repeats[60]

As a result of this work the NIH withdrew information about the age of the participants from public access[73]. The scientists responsible for developing the technique argued that existing policy tools could be used to mediate exposure to surname inference attacks (such as controlled-access databases with data use agreements)[74].

Defeating anonymisation using known phenotypic trait correlation

An individual's genotype is linked to physical traits, such as eye or hair colour. In 2015, a group of researchers tested the extent to which known physical traits can be used to identify people within genomic databases, even where the relevant data has been subject to anonymisation. Phenotypic information can be acquired from online social networks, medical databases, or simply knowing the person concerned. Genomic data can be acquired from online databases such as OpenSNP or the Personal Genome Project. If a hypothetical attacker has access to information about a person's phenotypic traits, those traits can be used to identify the genotype of this individual in a database. This allows other information to be inferred from the now de-anonymised data, such as predisposition to severe diseases. For example, someone might identify a target's genome using known phenotypic traits and subsequently infer a susceptibility to Alzheimer's disease.

The research group responsible for investigating the limits of this approach listed several possible scenarios in which this technique might be used. These included situations in which an entity:

- Wants to identify the genotype of a targeted individual in a database, knowing some of their phenotypic traits
- Has online access to a dataset of anonymised genotypes indicating some risks to certain severe diseases and who wants to know to whom they belong (e.g. for insurance purposes)
- Has access to a database with patients' diseases and another database with patients' anonymised genomes, and who wants to deanonymise these genomes (e.g. hospital IT staff)
- Has access to an identified genotype on an online database and wants to use it to identify the corresponding user on a collaborative patient information network, such as PatientsLikeMe

The team responsible for this work examined two deanonymisation attacks: (a) an 'identification attack' where the intent is to identify the genotype (among multiple genotypes) that correspond to a given phenotype, and (b) a 'perfect matching attack' where the intent is to match multiple phenotypes to their corresponding genotypes. Two forms of background knowledge were assumed. The first was an unsupervised approach that made use of existing knowledge of associations between SNPs and phenotypic traits. The second was known as a supervised approach and was based on learning the genomic-phenotypic associations from data containing both data types (i.e. genomic and phenotypic).

The experimental results were based on a database of 80 participants. Using the identification attack a proportion of 13% correct matches was achieved in the supervised case and 5% in the unsupervised case. In the perfect matching attack, the proportion of correct matches was 16% in the supervised case and 8% in the unsupervised case. When the database size was 10, these results were 53%/44% and 65% /58%, respectively.

Unsurprisingly, the distinguishability between two individuals had a significant effect on the success of the perfect matching attack: the more distinguishable two individuals are, the more likely their genomic (or phenotypic) data are to be deanonymised. The team concluded that the threat to genomic privacy posed by their deanonymisation attacks will become more serious as more SNP-trait association information becomes available.

The team also suggested a number of countermeasures. For example: encryption (usually acceptable for current healthcare purposes, but more difficult to ensure the accuracy of research data), removal from databases those SNPs associated with visible traits, or the addition of limited 'noise' to genome data to limit distinguishability between two genomes[74].

More recently, researchers have claimed to predict biometric traits using whole genome sequencing, detailed phenotyping and statistical modeling in a cohort of participants of diverse ancestry[75]. Although the significance of these findings have been challenged[76,77], this work raises the possibility that associating deidentified genomic data with phenotypic measurements such as height, skin and eye colour, facial structure and voice might have implications for personal privacy and consent practices.

In combination, these mechanisms pose a significant threat to the safe and secure use of personal data, especially as increasing amounts of data are combined in novel ways across sectors (such as is being envisaged from the use of big data in the life sciences)[78].

7. What do people think about genomic data?

Preferences as to what ought to be done with genomic data vary considerably depending on factors such as the apparent anonymity of the data or the age of the respondent. But knowing the number of people happy to do one thing does not necessarily reveal if and when society ought to override minority preferences.

It has been said that *"the willingness of individuals and communities to assume some risk to participate in biomedical research depends on the scientific community's ability to maintain the public's trust"*[79]. This is no doubt true, and attitudes to genomic privacy are complex and changing[79]. Research published in 2016 by the Wellcome Trust found that in the UK genetic data are little understood but are seen as inherently personal and more private than other types of data[80]. There is also evidence to suggest disquiet about a lack of control over genomic data – people are unwilling to allow total access to their genomic data, but they do understand and accept that there are limits to the extent to which personal identities may be secured[81].

Yet this disquiet is by no means universal: preferences vary considerably between groups and individuals, and people's views are changing rapidly[21]. For example, attitudes towards public revelations of private information by younger generations suggest (but by no means demonstrate) that they may be less concerned about the privacy of their genomic data than older generations[26], although this might be related to the fact that the young tend to have fewer life events they may wish to keep private[82]. Research published by the Wellcome Trust in 2016, suggests that a majority support data sharing for medical research: 75% of survey respondents were willing to share anonymised genetic data for the purposes of medical research, 22% unwilling, and 4% not knowing (anonymised was defined as not including *"name, date of birth, address or any contact details"*)[83].

Because attitudes seem to change relatively quickly, and vary between groups and individuals, some argue that blanket policies about what to do with data are inappropriate, whether designed to strongly restrict data sharing or to mandate their public release[84]. This is especially the case in terms of health data and health-related uses of data. Data collection, categorisation, and use are complex areas and people's attitudes towards use of 'their' data are similarly complex. Attitudes are not solely related to the idea of concrete harm: personal and familial identity as well as beliefs about equity and democratic ideals all play a role[17].

These worries are embedded in a wider societal concern about data use, and highly technical arguments regarding anonymisation, while extremely important, to some extent obscure difficult debates about the acceptability of overriding minority values and preferences – 22% is a fairly large minority.

As we have already noted, these concerns have been discussed in a deep and complex literature, but changes are modest. The National Data Guardian for Health and Care has proposed that a new opt-out model be introduced, to facilitate more systematic and streamlined secondary uses of personal data. Once implemented, the numbers of opt-outs might provide a more robust indication of the numbers of people who have reservations about data sharing. This should also involve a comprehensive assessment of the future uses to which health data and samples might be put including potential unforeseen uses in order to address whether existing protections, including anonymisation, are sufficient, and whether new approaches are needed[85]. In order to achieve effective policy change there is a pressing need for inclusive debate that draws on this literature, and that debate should seek to generate meaningful and coherent conclusions[86].

8. Conclusions and recommendations

Genomic data are becoming ever more important in medicine and science, and their optimal use could deliver great benefits[87]. However, concerns about the potential risks associated with their use, such as adverse impacts on individual privacy need to be properly addressed. To this end, anonymisation techniques are extremely useful tools when attempting to maintain the privacy of those people to whom genomic data relate but they do not resolve the fact that genomic data do not sit comfortably within the current legal and regulatory framework. This is true both as a consequence of their nature and because the framework itself lacks overall coherence in the face of a number of technical and social developments.

The challenges to the use of health data in general have been comprehensively analysed across the literature by hundreds of specialists and groups, and occasionally these challenges intrude into public life (such as with the *Care.data* controversy). These issues also tend to come to public prominence when a data breach occurs or data security is compromised for other reasons, such as through hacking or criminal activity[88]. Although these issues are well rehearsed, there remains confusion and disagreement as to the extent and true nature of the challenges and what might be done to overcome them.

In 2016 the House of Commons Science and Technology Committee recommended that the UK Government should set up a Council of Data Ethics to address the many data-related problems facing the UK[89]. In their response to the Committee, the UK Government agreed that independent oversight was necessary and that it would investigate the possibility of establishing a Council on Data Ethics[89]. This would go some way to providing a framework for managing these important challenges. However, the most important issue is the need for open discussion at the societal level – questions of privacy and anonymity are becoming more prevalent and increasingly germane, yet clear agreement about how to respond to those questions is lacking. This applies even more so to those data people think of as health data, of which genomic data are a part.

There are no simple solutions to the challenges identified in this report, and in some cases these challenges are firmly entrenched. The recent EU General Data Protection Regulation, for example, changed a great deal about data protection law in the EU but broadly and firmly reinforced the legal approach to data described above; the flaws remain because changes have been incremental.

However it is also important to understand that there is a political element to this debate: protagonists in favour of enhanced data sharing may regard existing protections (such as anonymisation coupled with managed access) as being more robust than those with the primary aim of protecting individual privacy rights. And the stakes are high: the implications for healthcare and society are profound, depending on which vision is ultimately adopted[87].

This report proposes a number of responses to this challenge:

- **Take greater account of societal and technical change:** questions of privacy and anonymity have become pervasive and deserve urgent and extensive discussion as their implications extend to every level of society. This applies particularly to health data of which genomic data are part.
- **Do not rely solely on anonymisation:** greater emphasis should be placed on the use of systems that appropriately limit access to the data. Important elements include access control systems, audit, and legal sanctions. Although sophisticated techniques that manipulate data do have a very important role to play in protecting anonymity, they cannot be relied upon in all circumstances or over long periods of time.
- **Ensure transparency:** anonymisation techniques/mechanisms should not be hidden and their effectiveness should not be taken on trust. This is especially the case for systems that rely on public funds and use data derived from publicly funded systems.
- **Understand the difference between 'identification' and 'individualisation':** the distinction between these two activities needs to be critically evaluated, both in the scientific and the policy literature. This is necessary in order to ensure that the risks and harms associated with the two activities are correctly understood.
- **Move away from language that appears to be absolutist:** there are strong reasons to continue processing, linking and releasing genomic data despite failures of anonymisation, but public discussion, and comments made to patients and research participants, should make clear that in many cases anonymisation is not absolute and that it is not possible to fully determine the risks of future deidentification. Claims made about anonymisation ought to be moderated – patients, research participants, consumers, and doctors are unlikely to recognise the limitations of anonymisation without being told of them.
- **Pursue further research:** research into *"the potential harms associated with abuse of biological and health data, as well as its benefits"*[14] is needed to evaluate the nature, extent, and likelihood of such harms arising and determine how potential challenges should be optimally addressed.
- **Consider regulatory change:** The EU data protection regime emphasises the nature of data as a way of moderating its use, rather than the use itself. Instead, there have been calls for greater monitoring and sanctions for illegitimate identification of individuals, and the PHG Foundation supports ways of providing additional protections. These includes stronger, and in some cases criminal, penalties for misuse of data and the possibility of bringing inappropriate identification within the scope of data protection law. We welcome the fact that the UK Data Protection Bill includes sanctions for unauthorised re-identification of de-identified data within its scope[89, 90].
- **Establish a Council of Data Ethics:** the Government should empower a multidisciplinary transparent forum to address the technical, bioinformatic and regulatory challenges associated with anonymisation in order to optimise robust and proportionate data sharing practices in healthcare and medical research. This body could also help to promote national debate and resolve challenges associated with anonymised data in ways that build public trust and confidence.

References

1. UK Government (2017). Data Protection Bill (HL Bill 66).

2. Wright C, Middleton A, Burton H *et al*. Policy challenges of clinical genome sequencing. BMJ. Nov 2013).

3. Ayday E, Rasisaro JL, McLaren PJ *et al*. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data Proceedings of the 2013 USENIX conference on Safety, Security, Privacy and Interoperability of Health Information Technologies 1-1. 2015.

4. Ayday E, Cristofaro E, Jean-Pierre H *et al*. Whole genome sequencing: revolutionary medicine or privacy nightmare? IEEE Computer. 2015 Feb 16; 48(2) 58-66; Cristofaro E. Genomic privacy and the rise of a new research community. IEEE Security and Privacy. 2014 April/May; 12(2) 80-3; Raisaro JL, Ayday E, Hubaux J-P. Patient privacy in the genomic era. PRAXIS 2014 May 7; 103(10): 579-86; Danezis G and Cristofaro E. Fast and private genomic testing for disease susceptibility. Proceedings of the 13th Workshop on Privacy in the Electronic Society. 2014; 31-4; Ayday E, Cristofaro E, Hubaux J-P *et al*. The chills and thrills of whole genome sequencing. IEEE Computer Society. 2013; PP(99) 1-9; Ayday E, Rasisaro JL, McLaren PJ *et al*. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. Proceedings of the 2013 USENIX conference on Safety, Security, Privacy and Interoperability of Health Information Technologies. 2015; 1-1; Baldi P, Baronio R, Cristofaro E *et al*. Countering GATTACA - efficient and secure testing of fully-sequenced human genomes. Proceedings of the 18th ACM conference on Computer and communications security. 2011. Oct 17-21; 691-702.

5. Ayday E, Cristofaro E, Jean-Pierre H *et al*. Whole genome sequencing: revolutionary medicine or privacy nightmare? IEEE Computer. 2015 Feb 16; 48(2) 58-66; Naveed M. Hurdles for genomic data usage management (position paper). IEEE Workshop on Data Usage Management (DUMA). May 2014; Danezis G and Cristofaro E. Fast and private genomic testing for disease susceptibility. Proceedings of the 13th Workshop on Privacy in the Electronic Society. 2014; 31-4; Heeney C, Hawkins N, dr Vries J *et al*. Assessing the privacy risks of data sharing in genomics. Public Health Genomics. 2011 Mar 29; 14(1): 17-25; Ayday E, Cristofaro E, Hubaux J-P *et al*. The chills and thrills of whole genome sequencing. IEEE Computer Society. 2013; PP(99) 1-9; Humbert M, Ayday E, Hubaux J-P *et al*. On non-cooperative genomic privacy. Proceedings of 19th International Conference on Financial Cryptography and Data Security. 2015 Jan 26-30.

6. Naveed M. Hurdles for genomic data usage management (position paper). IEEE Workshop on Data Usage Management (DUMA). May 2014.

7. Ayday E, Cristofaro E, Hubaux J-P *et al*. The chills and thrills of whole genome sequencing. IEEE Computer Society. 2013; PP(99) 1-9.

8. Naveed M, Ayday E, Clayton EW *et al*. Privacy in the genomic era. ACM Comput Surv. 2015 Sep 1; 48(1): Article No. 6.

9. Cristofaro E. Genomic privacy and the rise of a new research community. IEEE Security and Privacy. 2014 April/May; 12(2) 80-3; Humbert M, Ayday E, Hubaux J-P *et al*. On non-cooperative genomic privacy. Proceedings of 19th International Conference on Financial Cryptography and Data Security. 2015 Jan; 26-30; Ayday E, Cristofaro E, Hubaux J-P *et al*. The chills and thrills of whole genome sequencing. IEEE Computer Society. 2013; PP(99) 1-9; Ayday E, Rasisaro JL, McLaren PJ *et al*. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. Proceedings of the 2013 USENIX conference on Safety, Security, Privacy and Interoperability of Health Information Technologies. 2015; 1-1; Naveed M, Ayday E, Clayton EW *et al*. Privacy in the genomic era. ACM Comput Surv. 2015 Sep 1; 48(1): Article No. 6.

10. Cristofaro E. Genomic privacy and the rise of a new research community. IEEE Security and Privacy. 2014 April/May; 12(2) 80-3; Humbert M, Ayday E, Hubaux J-P *et al*. On non-cooperative genomic privacy. Proceedings of 19th International Conference on Financial Cryptography and Data Security. 2015 Jan; 26-30.

11. Heeney C, Hawkins N, de Vries J *et al*. Assessing the privacy risks of data sharing in genomics. Public Health Genomics. 2011 Mar 29; 14(1): 17-25.

12. Ohm P. Broken promises of privacy: responding to the surprising failure of anonymization U.C.L.A.L.Rev. 2009 Aug 13; 57: 1701-77.

13. Ohm P. Broken promises of privacy: responding to the surprising failure of anonymization U.C.L.A.L.Rev. 2009 Aug 13; 57: 1701-77; Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: ethical issues. 2015.

14. Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: ethical issues. 2015.

15. National Telecommunications & Information Administration, United States Department of Commerce (2016). Lack of Trust in Internet Privacy and Security May Deter Economic and Other Online Activities.

16. Lowrance W and Collins FS. Identifiability in genomic research. Science. 2007 Aug 3; 317(5838): 600-2.

17. Kaye J, Kanellopoulou N, Hawkins N *et al*. Can I access my personal genome – the current legal position in the UK. Med L Rev. 2014 Mar 22; 22(1): 64–86.

18. Hansson MG, Hanns L, Olaf R *et al*. The risk of re-identification versus the need to identify individuals in rare disease research. Eur J Hum Genet. 2016 May 25.

19. Kaye J. The tension between data sharing and the protection of privacy in genomics research. Annu Rev Genomics Hum Genet. 2012 Sep; 13: 415-31.

20. Curren L, Boddington P, Gowans H *et al*. Identifiability, genomics and U.K. data protection law. Eur J Health Law. 2010 Sep; 17(4): 329-44.

21. Cook AF. The truth about the truth - what matters when privacy and anonymity can no longer be promised to those who participate in clinical trial research. Research Ethics. 2013 Sep; 9(3): 97-108.

22. Ohm P. Broken promises of privacy: responding to the surprising failure of anonymization U.C.L.A.L.Rev. 2009 Aug 13; 57: 1701-77: "*With this provision, EU lawmakers sought to preserve space in society for the storage and transfer of anonymized data, thereby providing room for unencumbered innovation and free expression*"; "*Almost every single privacy statute and regulation ever written in the U.S. and the EU embraces – implicitly or explicitly, pervasively or only incidentally – the assumption that anonymization protects privacy, most often by extending safe harbors from penalty to those who anonymize their data.*"

23. Raza S, Hall A, Rands C *et al*. Data sharing to support UK clinical genetics and genomics services. PHG Foundation UK. 2015.

24. Curren L, Boddington P, Gowans H *et al*. Identifiability, genomics and U.K. data protection law. Eur J Health Law. 2010 Sep; 17(4): 329-44; Wallace SE, Gaye A, Shoush O *et al*. Protecting personal data in epidemiological research – dataSHIELD and UK law. Public Health Genomics. 2014 Jun; 17:149-57.

25. Erlich Y, Williams JB, Glazer D *et al*. Redefining genomic privacy – trust and empowerment. PloS Biol. 2014 Nov; 12(11): e1001983.

26. Greenbaum D. Genomic data disclosure – time to reassess the realities. Am J Bioeth. 2013 May; 13(5): 47-50.

27. Equifax hack: 44 million Britons' personal details feared stolen in major US data breach. The Telegraph. September 2017.

28. Ayday E, Raisaro J, Hubaux J-P *et al*. Protecting and evaluating genomic privacy in medical tests and personalized medicine. Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society. 2015 Nov 4; 95-106; Laurie G, Stevens L, Jones K *et al*. A review of evidence relating to harm resulting from uses of health and biomedical data. 2014 Jun 30.

29. Ayday E, Raisaro J, Hubaux J-P *et al*. Protecting and evaluating genomic privacy in medical tests and personalized medicine. Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society. 2015 Nov 4; 95-106.

30. Raisaro JL, Ayday E, Hubaux J-P. Patient privacy in the genomic era. PRAXIS 2014 May 7; 103(10): 579-86.

31. Rodriguez LL, Brooks LD, Greenberg JH *et al*. The complexities of genomic identifiability. Science. 2013 Jan 18; 339(6117): 275-6.

32. Walker N. All or Nothing: The False Promise of Anonymity. Data Science Journal. 2017; 16:24:1-7.

33. Department of Health. Information: to share or not to share? The information governance review. 2015.

34. Wallace SE, Gaye A, Shoush O *et al*. Protecting personal data in epidemiological research – dataSHIELD and UK law. Public Health Genomics. 2014 Jun; 17:149-57.

35. Global Alliance for Genomics and Health (2016) Data sharing lexicon.

36. Humbert M, Ayday E, Hubaux J-P *et al*. Reconciling utility with privacy in genomics. Proceedings of the 13th Workshop on Privacy in the Electronic Society. 2014 Nov 3; 11-20.

37. Raisaro JL, Ayday E, Hubaux J-P. Patient privacy in the genomic era. PRAXIS 2014 May 7; 103(10): 579-86; Humbert M, Ayday E, Hubaux J-P *et al*. Reconciling utility with privacy in genomics. Proceedings of the 13th Workshop on Privacy in the Electronic Society. 2014 Nov 3; 11-20; Erlich Y, Williams JB, Glazer D *et al*. Redefining genomic privacy – trust and empowerment. PloS Biol. 2014 Nov; 12(11): e1001983.

38. The Information Commissioner's Office. Anonymisation – managing data protection risk code of practice. 2012 Nov; Roebuck C. HSCIC data pseudonymisation review – interim report. Health and Social Care Information Centre. 2014.

39. Cristofaro E. Genomic privacy and the rise of a new research community. IEEE Security and Privacy. 2014 April/May; 12(2) 80-3; Ayday E, Cristofaro E, Hubaux J-P *et al*. Whole genome sequencing: revolutionary medicine or privacy nightmare?. IEEE Computer. 2015 Feb 15; 48(2): 58-66.

40. Humbert M. When others impinge upon your privacy – interdependent risks and protection in a connected world. Thesis - École Polytechnique Fédérale de Lausanne. 2015 Mar 13.

41. Naveed M, Ayday E, Clayton EW *et al*. Privacy in the genomic era. ACM Comput Surv. 2015 Sep 1; 48(1): Article No. 6; Humbert M, Ayday E, Hubaux J-P *et al*. Reconciling utility with privacy in genomics. Proceedings of the 13th Workshop on Privacy in the Electronic Society. 2014 Nov 3; 11-20; Naveed M. Hurdles for genomic data usage management (position paper). IEEE Workshop on Data Usage Management (DUMA). May 2014; Raisaro JL, Ayday E, Hubaux J-P. Patient privacy in the genomic era. PRAXIS 2014 May 7; 103(10): 579-86.

42. Budin-Ljøsne I, Burton P, Isaeva J *et al*. DataSHIELD – an ethically robust solution to multiple-site individual-level data analysis. Public Health Genomics. 2015 Feb; 18: 87-96.

43. See for example: Global Alliance for Genomics and Health. Data sharing lexicon. 2016.

44. See, generally: Global Alliance for Genomics and Health. Data sharing lexicon. 2016; Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: ethical issues. 2015; Department of Health. Information: to share or not to share? The information governance review. 2015.; Roebuck C. HSCIC data pseudonymisation review – interim report. Health and Social Care Information Centre. 2014.

45. Skopek JM (2015). Reasonable expectations of anonymity. Va L Rev. 2015 May; 101(3): 691-762.

46. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

47. Lynskey O. Deconstructing data protection: the 'added-value' of a right to data protection in the EU legal order. Int'l & Comp LQ. 2014 Jul; 63(3): 569-97.

48. Article 29 Data Protection Working Party. Letter to DG Connect – Annex: health data in apps and devices. 2015.

49. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

50. HM Government. 2017. The exchange and protection of personal data. A future partnership paper.

51. Duhigg C. How companies learn your secrets. The New York Times. 2012 Feb 16.

52. Data Protection Act 1998, s.1(1).

53. The Information Commissioner's Office. Anonymisation – managing data protection risk code of practice. 2012.

54. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. [Internet]. Department of Health & Human Services (USA); 2016.

55. Oswold M. Anonymisation Standard for Publishing Health and Social Care Data Specification (Process Standard). Information Standards Board for Health and Social Care. 2013; The Information Commissioner's Office. Anonymisation - managing data protection risk code of practice. 2012. Information Commissioner's Office. Big data, artificial intelligence, machine learning and data protection. 20170904. Version: 2.2.

56. This is the ICO interpretation of *The Queen on the application of the Department of Health v Information Commissioner* [2011] EWHC 1430 (Admin). See: The Information Commissioner's Office. Anonymisation – managing data protection risk code of practice. 2012.

57. *The Queen on the application of the Department of Health v Information Commissioner* [2011] EWHC 1430 (Admin).

58. *APG v Information Commissioner & The Ministry of Defence* [2011] UKUT 153 (AAC) (15 April 2011).

59. Suver C, Wilbanks J, Friend SH. US-EU scientific research collaborations – Sage bionetworks' experience navigating the complex regulatory landscape. Sage Bionetworks. 2013.

60. Erlich Y and Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014 May 8; 15: 409-421.

61. Naveed M, Ayday E, Clayton EW *et al*. Privacy in the genomic era. ACM Comput Surv. 2015 Sep 1; 48(1): Article No. 6; Lowrance W and Collins FS. Identifiability in genomic research. Science. 2007 Aug 3; 317(5838): 600-2; Pereira S, Gibbs RA, McGuire AL. Open access data sharing in genomic research. Genes (Basel). 2014 Sep; 5(3): 739-47.

62. Zhen L, Owen AB, Altman RB. Genomic research and human subject privacy. Science. 2004 Jul 9; 305(5681):183.

63. Pakstis AJ, Speed WC, Fang R *et al*. SNPs for a universal individual identification panel. Hum Genet. 2010 Mar; 127(3): 315-24.

64. Im HK, Gamazon ER, Nicolae DL *et al*. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am. J. Hum. Genet. 2012 Apr 6; 90(4): 591-8; Homer N, Szelinger S, Redman M *et al*. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PloS Genet. 2008 Aug 29; 4(8): e1000167.

65. Im HK, Gamazon ER, Nicolae DL *et al*. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am. J. Hum. Genet. 2012 Apr 6.

66. Wang R, Li YF, Wang X *et al*. Learning your identity and disease from research papers – information leaks in genome wide association study. Proceedings of the 16th ACM conference on computer and communications security. 2009 Nov 9-13; 534-44.

67. Pereira S, Gibbs RA, McGuire AL. Open access data sharing in genomic research. Genes (Basel). 2014 Sep; 5(3): 739-47; Kaye J. The tension between data sharing and the protection of privacy in genomics research. Annu Rev Genomics Hum Genet. 2012 Sep; 13: 415-31.

68. Masca N, Burton PR, Sheehan NA. Participant identification in genetic association studies – improved methods and practical implications. Int J Epidemiol. 2011 Dec 7; 40(6): 1629-42.

69. Naveed M, Ayday E, Clayton EW *et al*. Privacy in the genomic era. ACM Comput Surv. 2015 Sep 1; 48(1): Article No. 6; Conley JM, Doerr AK, Vorhaus DB. Enabling responsible public genomics. Health Matrix. 2010; 20(2): 325-85.

70. Gymrek M, McGuire AL, Golan D *et al*. Identifying personal genomes by surname inference. Science. 2013 Jan 18; 339(6117): 321-4; Philibert RA, Terry N, Erwin C *et al*. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. Clin Epigenetics. 2014 6(1):28; Pereira S, Gibbs RA, McGuire AL. Open access data sharing in genomic research. Genes (Basel). 2014 Sep; 5(3): 739-47.

71. Lunshof JE and Ball MP. Our genomes today – time to be clear. Genome Medicine. 2013 Jun 27; 5(6):52.

72. Gymrek M, McGuire AL, Golan D *et al*. Identifying personal genomes by surname inference. Science. 2013 Jan 18; 339(6117): 321-4; Erlich Y and Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014 May 8; 15: 409-421.

73. Rodriguez LL, Brooks LD, Greenberg JH *et al*. The complexities of genomic identifiability. Science. 2013 Jan 18; 339(6117): 275-6; Lunshof JE and Ball MP. Our genomes today – time to be clear. Genome Medicine. 2013 Jun 27; 5(6):52; Hayden EC. Privacy protections: the genome hacker. Nature. 2013 May 8.

74. Gymrek M, McGuire AL, Golan D *et al*. Identifying personal genomes by surname inference. Science. 2013 Jan 18; 339(6117): 321-4.

75. Lippert C, Sabatini R, Maher MC *et al*. Identification of individuals by trait prediction using whole-genome sequencing data. PNAS 2017 vol. 114 no. 38 10166–10171; doi: 10.1073/pnas.1711125114

76. Humbet M, Huguenin K, Hugonot J *et al*. De-anonymizing genomic databases using phenotypic traits. Proceedings on Privacy Enhancing Technologies. 2015 Jun; 2:99-114.

77. Erlyich Y. Major flaws in "Identification of individuals by trait prediction using whole-genome sequencing data". Preprint. BioRXiv.

78. Berger KM, Roderick J. National and transnational security implications of big data in the life sciences. American Association for the Advancement of Sciences. November 10 2014.

79. Rodriguez LL, Brooks LD, Greenberg JH *et al*. The complexities of genomic identifiability. Science. 2013 Jan 18; 339(6117): 275-6.

80. Ipsos MORI. The one-way mirror – public attitudes to commercial access to health data. The Wellcome Trust. 2016.

81. McEwen JE, Boyer JT, Sun KY. Evolving approaches to the ethical management of genomic data. Trends Genet. 2013 Jun; 29(6): 375–82.

82. Rose H. The commodification of bioinformation: the Icelandic health sector database. The Wellcome Trust. 2001.

83. Ipsos MORI. Wellcome Trust Monitor Report Wave 3: Tracking public views on science and biomedical research. Wellcome Trust. 2016.

84. Pereira S, Gibbs RA, McGuire AL. Open access data sharing in genomic research. Genes (Basel). 2014 Sep; 5(3): 739-47.

85. National Data Guardian for Health and Care. 2016. Review of Data Security, Consent and Opt-outs.

86. O'Doherty K, Christofides E, Yen J *et al*. If you build it, they will come: unintended future uses of organised health data collections. BMC Medical Ethics (2016) 17:54.

87. London: Department of Health. 2017. Annual Report of the Chief Medical Officer. Generation Genome.

88. Greenberg A. The WannaCry Ransomware Hackers Made Some Real Amateur Mistakes. Wired. 2017.

89. House of Commons Science and Technology Committee. The big data dilemma: Fourth Report of Session 2015–16. 2016.

90. UK Government (2017) Data Protection Bill, paragraph 162

phg foundation
making science work for health

About the PHG Foundation

The PHG Foundation is a pioneering independent think-tank with a special focus on genomics and other emerging health technologies that can provide more accurate and effective personalised medicine. Our mission is to make science work for health. Established in 1997 as the founding UK centre for public health genomics, we are now an acknowledged world leader in the effective and responsible translation and application of genomic technologies for health.

We create robust policy solutions to problems and barriers relating to implementation of science in health services, and provide knowledge, evidence and ideas to stimulate and direct well-informed discussion and debate on the potential and pitfalls of key biomedical developments, and to inform and educate stakeholders. We also provide expert research, analysis, health services planning and consultancy services for governments, health systems, and other non-profit organisations.

ISBN: 978-1-907198-22-9

CAMBRIDGE UNIVERSITY
Health Partners

PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
+44 (0) 1223 761 900
www.phgfoundation.org

phg
foundation
making science
work for health