phg foundation
making science work for health

Genomic medicine and research: when does the GDPR apply?

Authors

Colin Mitchell, Johan Ordish and Alison Hall

Acknowledgements

This work was funded by the Information Commissioner's Office Grants Programme

URLs in this report were correct as of July 2020

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

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

+44 (0)1223 761900

January 2020

Correspondence to:
intelligence@phgfoundation.org

How to reference this report:
Genomic medicine and research: when does the GDPR apply? (2019)

Summary

The EU's General Data Protection Regulation (GDPR) gives rise to significant uncertainty for those working in genomic medicine and research in Europe and those in collaborations involving EU citizens' data. In this discussion paper we address the challenge of understanding whether the GDPR applies to the data and the stakeholders involved in genomics projects. We outline one approach to this question and highlight the definitions of 'personal data', 'pseudonymisation' and (joint) 'control' which will influence when uses of genetic and clinical data will be governed by the GDPR. If this is the case, researchers and health care professionals are required to support individuals rights and fulfil the obligations set out in the GDPR.

Key points

- Whether the GDPR applies depends on the nature of the data and the context: if there are means to identify an individual which are reasonably likely to be used by legitimate users or third parties, then the GDPR will apply
- It is not yet clear if 'pseudonymised' data are always 'personal data' and if the possibility that they will be combined with a key to identify an individual will be enough to constitute 'personal data', or, whether pseudonymisation can lead to successful anonymisation if this is no longer reasonably likely to occur
- The GDPR does not apply to familial genetic or clinical information unless it is possible to distinguish the individual family member it relates to
- The GDPR may still apply to an organisation, professional or researcher that does not deal with 'personal data' if they help determine the purposes and means of processing personal data.

Introduction

The General Data Protection Regulation (GDPR) updated the legal framework for the processing of personal data across the EU. It established: new rights (such as the right to erasure), new obligations for data controllers and processors, new governance requirements (e.g. for a data protection officer reporting directly to the highest management level), and greatly enhanced potential fines of up to 4% global annual turnover. At the same time, new and improved genome-sequencing technologies are driving the ever increasing generation, collection and sharing of genetic and clinical data for healthcare and medical research. Taken together, these developments have generated significant uncertainty for professionals about whether, and how, data protection law applies in genomic medicine and research.

In this discussion paper we outline a way to determine whether the GDPR applies. First, by asking if data are 'personal data' or anonymous? Second, are they 'pseudonymised' data governed by the GDPR? Third, are they 'genetic data'? Fourth, even if you do not directly use personal data, could you be a 'joint controller' under the GDPR? (Figure 1)

The GDPR applies to the 'processing' of personal data. This is defined so broadly that it includes almost any operation that can be performed with data (Art 4(2)). Moreover, although it clearly applies to professionals based in an EU Member State, the transfer of personal data outside the EU is governed by the GDPR and subject to special controls (Art 44) so many of the following questions are relevant for those working with the genetic data of EU citizens.

Definitions:

'Processing' means any operation or set of operations which is performed on personal data … such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction (Art 4 (2))

'Controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data … (Art 4 (7))

'Processor' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller (Art 4 (9))

Figure 1: determing whether the GDPR applies in genomic medicine and research

Are data 'personal data' or 'anonymous'?

The GDPR only applies to 'personal data'.

> 'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. (Art 4 (1))

This makes clear that the starting point is an analysis to determine if an individual can be directly, or indirectly, identified from the data. Conversely, the GDPR does not apply to anonymous information.

> Namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. (Recital 26)

This recital to the Regulation explains the approach that is required to determine whether individuals are identifiable.

> To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. (Recital 26)

This means that unless clear direct identifiers (e.g. a name) are part of the data, the key question is whether an individual is indirectly identifiable based on the real-world context. In this analysis, the standard of risk of identification is whether means are reasonably likely to be used to identify an individual (i.e. the means must be both reasonably likely to be used, and to succeed). This has been contrasted with a 'mere hypothetical possibility',[1] or a 'remote' chance,[2] of identification, which would be insufficient.

As the recital states, 'all objective factors' should be taken into account, but some factors are more relevant in certain contexts. We have identified four related factors as particularly relevant to genomic medicine and research:

- The 'richness' of the data
- The people who may be able to identify an individual
- Whether there is other available information that could be used to help identify an individual
- The presence of safeguards against identification

How 'rich' are the data?

Identifiability is particularly challenging in genomic medicine and research because 'richer' (in scale and quality) genetic and associated phenotypic data is generally of greater utility. However, richer data has greater potential for identification of an individual, through increased capacity for linkage with other data sources, inferences, and singling-out of an individual within a dataset.[3] Though such data will not always be personal data it does require careful consideration of the safeguards that can be put in place.

It is best practice, and in accordance with the data protection principle of data minimisation (Art 5(1)(c)) that data are as limited as practically possible. This requires an audit of all the genetic and clinical data that are being processed or stored and consideration of whether all categories are necessary for the research or health purposes being pursued. For many research and healthcare purposes, rich genetic data (e.g. whole-exomes or genomes) and clinical data are necessary. This increases the likelihood of such data being defined as 'personal data' and the level of protection required.

It should also be kept in mind that even limited genetic or phenotypic data may be highly identifying,[4] for example, very rare variants or facial images, so there must also be scrutiny of the identifiability of individual categories of data.

Can I or my colleagues identify an individual?

Those working with genomic data should begin with a consideration of whether they, or those they work with, may be able to identify individuals. In the healthcare context this is likely to be the case, especially if professionals have access to information systems that could be used to cross-reference information. By contrast, in research, or where the data are processed by professionals with no access to information systems, it is possible that the data are no longer identifiable. For example, when a clinical centre has anonymised the data before it is transferred to researchers.

Researchers and downstream recipients of data still need to be cautious in projects which have explicitly incorporated mechanisms for recontact (something being increasingly recommended in genetics research[5]). If there are mechanisms available for researchers to obtain identifying information from clinical colleagues, or if they are reasonably likely to obtain this information, it is possible that data are personal data,[6] even if they are not likely to be used often. [7] We discuss the challenges of coded, pseudonymised data further below.

Could someone else identify an individual?

The recital makes clear that determining whether data are personal data requires an assessment of the likelihood that either the controller or another person could identify an individual. Depending on the context, this could include recipients of research or published data, those with remote access to a database (e.g. registered users), as well as potential third parties. The ICO advises data controllers to consider the possibility of identification by a determined person with a particular reason to want to identify individuals, and because some data is more attractive than others, there should be consideration of potential motivations and how this may alter the level of technical skill and knowledge involved.[8]

The more accessible data are, the harder it becomes to consider all the potential skills and motivations of third parties. For example, if data are openly available online, there are people around the world with the investigative and technical abilities to attempt identification of an individual from even a limited dataset. Limiting access to data where possible, and close consideration of the safeguards that can be adopted may help to reduce the risk of identification. There should also be careful consideration of whether other information is available which could help to identify an individual.

Is there additional information which could help identify an individual?

The availability of additional information is a key factor in assessing whether there are 'means reasonably likely to be used to identify an individual' but it is also an increasingly challenging question in the rapidly developing genomics context. For instance, the explosion in popularity of ancestry genetic testing has shown in the USA that even limited genetic data[9] can be used to search public genetic databases such as GEDMatch for familial matches and make inferences about the identity of an individual. It has already been demonstrated that the identity of individuals can be inferred from incomplete datasets – including genomic datasets[10] – using other publicly available additional information, such as census data. Though some argue that such inferences are often too uncertain[11] for there to be a reasonable likelihood of accurate identification, this is now being challenged; Rocher and colleagues have recently demonstrated that the likelihood of correctly re-identifying a specific individual can be estimated with high accuracy even when an 'anonymised' dataset is substantially incomplete. They argue that 'even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR'.[12]

On a more mundane but no less important level, the specific knowledge that clinicians, family members or others have about a particular patient or individual could be sufficient additional information to make identification more than reasonably likely. There has to be an ongoing assessment of the characteristics that have been included in any released datasets, and how they could be threatened by the availability of additional information like genealogy databases. If the intended purpose demands a relatively rich dataset then the risk of identification may still be reduced using safeguards.

Are there safeguards which protect against identification?

Another key factor which can influence whether data are identifiable, is whether sufficient safeguards are in place to protect them.[13] Safeguards will frequently be in place in healthcare and medical research, in particular: legal obligations of confidentiality, restricted access processes, and legal restrictions on re-use or re-identification of data. The intention behind such safeguards is that such a disproportionate effort is required that risk of identification becomes insignificant.[14]

However, there must always be careful consideration of the means available to recipients of data – or third parties who could obtain data through them – to identify individuals despite the presence of safeguards. A key safeguard that is heavily promoted in the GDPR is pseudonymisation. Its inclusion in the Regulation has, however, led to some confusion.

Are 'pseudonymised' data personal data?

Pseudonymisation is generally understood to involve the removal and replacement of real-world identifiers with a key, cipher or code so that an individual cannot be easily identified from the data without the key or code. The inclusion of 'pseudonymisation' in the GDPR has caused some uncertainty about whether pseudonymised data are now always to be treated as 'personal data' and so governed by the GDPR. The GDPR refers to 'pseudonymisation' as a measure to reduce risks to data subjects (recital 28) and defines it as:

> 'Pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person. (Art 4(5))

In guidance on the previous data protection law, the ICO suggested pseudonymised data could be sufficiently well separated – using technical and organisational safeguards – from the 'key' ('additional information' in the GDPR definition) needed to link these to particular individuals, that it would amount to effective anonymisation.[15] This is the position under the US Common Rule (which governs research in the U.S.)[16] It is also the logical conclusion of applying the test for 'personal data' to data which have undergone pseudonymisation, so that it is no longer reasonably likely that an individual could be identified from those data.

However, there is now uncertainty about the status of pseudonymised data.[17] Part of recital 26 may suggest that a mere possibility of re-combining data with additional information means it will be '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.

And in their latest online guidance, the ICO state that:

> Pseudonymisation is effectively only a security measure. It does not change the status of the data as personal data. Recital 26 makes it clear that pseudonymised personal data remains personal data and within the scope of the GDPR.[18]

If pseudonymised data are to be treated as personal data in almost all cases, it means a significant difference between the risk assessment required of data more generally (that it must be at least reasonably likely that an individual can be identified), and the risk assessment for pseudonymised data and its identifying key. With the consequence that those working with pseudonymised genetic or clinical data would be likely to be processing 'personal data' if there is a key in existence which can be used to re-identify an individual.

We suggest that all data should be subject to the same standard of identifiability regardless of whether it has been pseudonymised or subjected to another process of de-identification. Mourby and colleagues draw on a 2016 judgment of the European Court of Justice (ECJ)[19] to argue that the conventional test of a reasonable likelihood of identification will be applied to data that are separate from additional identifying information.[20] This is compelling but unfortunately we cannot yet be certain that this approach will be followed because the data in that case had not actually undergone pseudonymisation and because the case was governed by the previous Directive, not the GDPR.

There are good reasons to doubt that a different standard of risk is applied to pseudonymised data: pseudonymisation is primarily a safeguard for data in the Regulation, it is referred to as a process not a category of personal data and no rationale is provided to justify an exceptional standard for this form of de-identified data versus another (e.g. aggregated data). In the absence of certainty around the status of pseudonymised data, professionals who wish to adopt a precautionary approach could treat the data as 'personal data' unless technical and organisational safeguards make it highly unlikely for the key and the data to be recombined.

Are they 'genetic data' under the GDPR?

The GDPR has explicitly incorporated a definition of 'genetic data' as a 'special category' of data meriting higher protection (Art 9). However, to fall within the scope of the GDPR, it is a prerequisite that data are 'personal data', meaning not all data commonly referred to as genetic data are governed by the GDPR.

> 'Genetic data' means personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question (Art 4 (13))

In some ways this is an expansive definition of genetic data: it extends to information about the physiology – not just health – of a person (see Figure 2) and the explanatory text adds that this not only includes the results of chromosomal, DNA or RNA analysis, but also 'equivalent information' obtained from the analysis of 'another element' which relate to the 'inherited or acquired genetic characteristics' of a natural person (Recital 34).

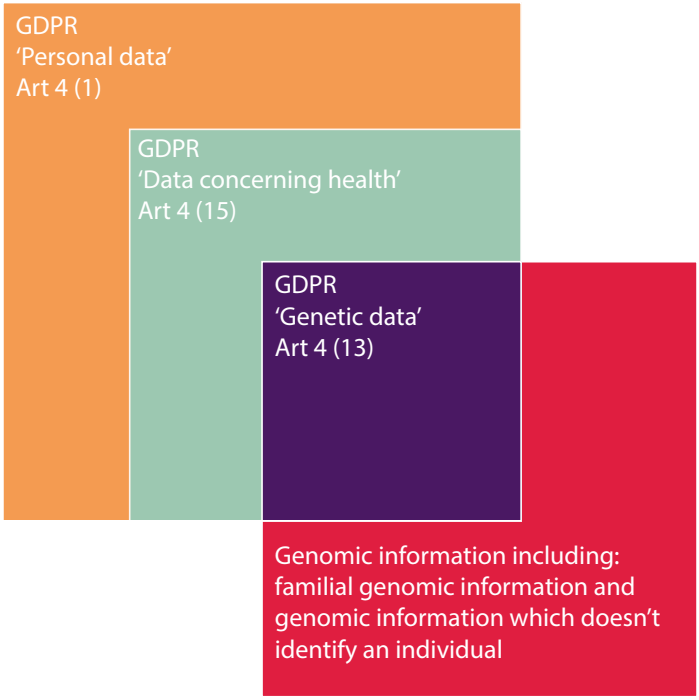
The scope of this definition is not clear but it could mean that information derived from the analysis of a biological sample, such as a blood test, which reveal 'genetic characteristics' also constitute 'genetic data' even if it has not been derived from analysis of DNA or RNA (for example, diagnosis of a genetic disorder such as sickle cell disease by inspection of the blood cells).

How genetic data are governed by the GDPR is limited in one significant respect: the definition makes clear, that 'genetic data' must first and foremost be 'personal data'. The approach the law takes to 'personal data' is that an individual must be capable of being uniquely singled-out and distinguished from others.[21] Isolating the data to more than one person is insufficient for it to constitute 'personal data'. This means that the GDPR does not apply to familial genetic information unless it is possible to distinguish the individual member of the family it relates to. This is also a potential reading of the word 'unique' in the definition:[22] that the information must be unique to one specific person (i.e. not 'shared' by that person and their relatives with nothing more to identify the specific data subject).

However, there should be caution when deciding that health and genetic data are familial as opposed to personal because, as we have outlined, if the data are reasonably likely to be combined with other available information or background knowledge (e.g. of family members) to identify an individual data subject, the data will fall within the scope of the GDPR.

The approach to genetic data under the GDPR is clear even if it doesn't correspond to more common understandings of genetic data: First, are the data 'personal data', i.e. can an individual be directly or indirectly identified and distinguished from the data? Second, and only if the answer is yes, do the data fall within the definition of genetic data? This is important for a number of reasons. For example, the processing of 'special categories' of personal data is prohibited without specific justification and Member States are allowed to introduce their own conditions for the processing of genetic data (Art 9(4)). However, there are broad and overlapping special categories of data, such as 'data concerning health' (Art 4 (15)) which means in practice most data used in genomic medicine and research is likely to be subject to higher protection.

Figure 2: The relationship between GDPR 'personal data', 'data concerning health', 'genetic data', and genomic information.



Even if I do not deal with personal data, am I a joint 'controller' under the GDPR?

As well as considering what constitutes 'personal data' the GDPR assigns different responsibilities to those processing data depending on whether they are a data processor or controller. As outlined above, 'processor' refers to anyone performing almost any imaginable operation on personal data. However, a controller does not have to access or ever deal with 'personal data':

> 'Controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data (Art 4 (7))

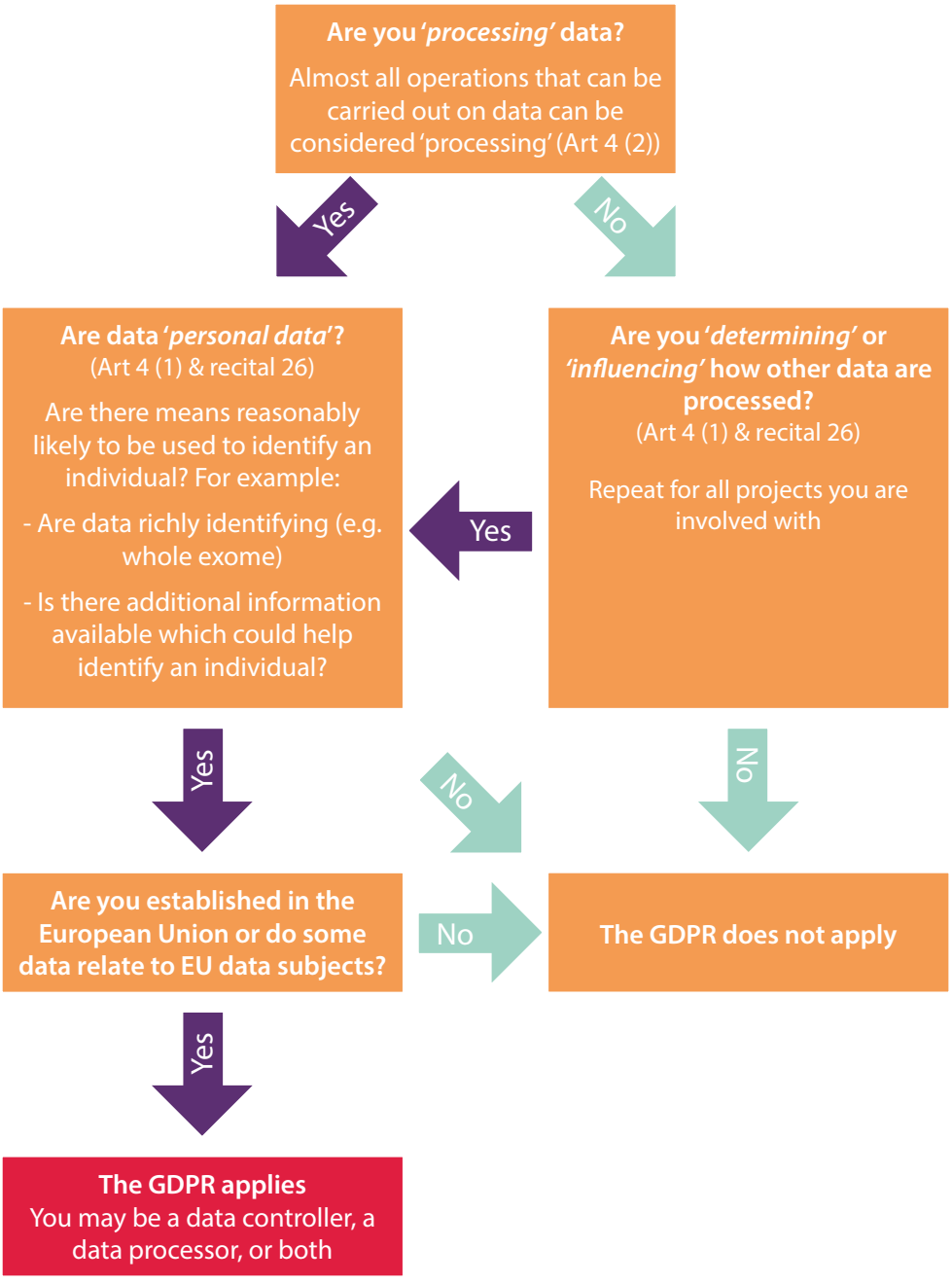
As under the previous law, it is clear that there can be more than one data controller and in recent years, the ECJ has taken an increasingly expansive approach to determining that an individual or organisation is jointly 'determining the purposes and means of the processing' of personal data. This case law suggests that even very limited involvement or influence over the collection or processing of personal data may suffice.

In the Wirtschaftsakademie case[23] the ECJ found that administrators of a Facebook fan page were jointly responsible for processing of personal data by Facebook even though they never had access to any personal data themselves, and in Tietosuojavaltuutettu[24] the court found that the Jehovah's Witnesses religious community was jointly responsible for notes made by individual members in door-to-door preaching even though it neither required nor obtained those notes. Most recently, in Fashion ID, [25] the ECJ found that simply embedding a Facebook 'Like' button on a website, which automatically transmitted personal data to Facebook, was sufficient to be a controller of that data. It was enough that the website operator had 'made it possible' for personal data to be obtained.

The essence of these decisions is that if a person or organisation 'influences', creates an 'opportunity' for, 'encourages' or 'coordinates' the processing of personal data by another, there is a significant chance that they are jointly determining the purposes and means of that processing of personal data.

If applied to the genomics context, where there are frequently multiple collaborators dealing with data in different forms as part of a shared enterprise, it is likely that many professionals and institutions would meet this standard for 'control'. To determine if the GDPR applies to them, collaborators should assess whether they are processing personal data and/or acting as a data controller (Figure 3). If they are a data controller they must determine their respective responsibilities for compliance with the GDPR in a transparent manner but will all remain joint and severally liable for all the obligations owed under the GDPR (Art 26).

Figure 3: Determing whether the GDPR applies



Conclusions

This discussion paper considers how to determine if the GDPR applies in genomic medicine and research in the EU or when using data from EU data subjects. There are many specific factors that could be relevant to individual contexts but the general approach we propose is to begin by asking first if the data being processed are 'personal data'? Second, are they pseudonymised data (and therefore best treated as personal data)? Third are they genetic data under the GDPR? Fourth, even if you do not directly use personal data, could you be a 'joint controller' under the GDPR?

If it is determined that the GDPR applies then there are a wide range of rights and obligations that must be met. The challenge of meeting some of these in the genomics context is the subject of our further research.

Key messages

Are data 'personal data' or 'anonymous'?

- An assessment of whether it is reasonably likely that an individual may be identified should include: consideration of the richness of the data and how potentially identifying different categories are, whether users or third parties may be able to identify an individual, whether additional information could help to identify an individual, and whether sufficient safeguards are in place to reduce the risk.

- If data are released in 'anonymised' form, there should be an ongoing assessment of the characteristics that have been included, and how they could be threatened by the availability of additional information sources such as genealogy databases, and new technologies

Are they pseudonymised data?

- Given the uncertainty around the status of pseudonymised data, anyone wishing to adopt a precautionary approach should treat this data as 'personal data' unless technical and organisational safeguards make it effectively impossible for the key and the data to be recombined

Are they 'genetic data' under the GDPR?

- The GDPR only applies to 'personal data' which means that it does not apply to familial genetic information unless it is possible to distinguish the individual member of the family to whom it relates

- However, there should be caution in deciding that data are familial as opposed to personal if the data are reasonably likely to be combined with other available information or background knowledge (e.g. of family members) to identify an individual data subject

Even if I do not deal with personal data, am I a joint 'controller' under the GDPR?

- If an organisation or person influences, creates an opportunity for, encourages or coordinates the processing of personal data they may be considered a joint controller under the GDPR, even if they never receive or process personal data. Joint controllers are joint and severally liable for the obligations owed under the GDPR and must transparently determine their respective responsibilities for compliance with the GDPR

References

1. Article 29 Data Protection Working Party. Opinion 4/2007 on the Concept of Personal Data. 2007 Working Party Opinions; (Lx) 1–26.
2. R (on the application of the Department of Health) v Information Commissioner [2011] EWHC 1430 (Admin).
3. Finnegan T, Hall A. Identification and genomic data. PHG Foundation; 2017. p22-9.
4. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR. EMBO Reports; 2019: e48316.
5. Bombard Y, Brothers K B, Fitzgerald-Butt S, et al. The Responsibility to Recontact Research Participants after Reinterpretation of Genetic and Genomic Research Results. American Journal of Human Genetics. 2019; 104(4): 578–595.
6. Article 29 Data Protection Working Party. Opinion 4/2007 on the Concept of Personal Data. 2007 Working Party Opinions. p19-20.
7. Case C-582/14 Breyer v Bundesrepublik Deutschland, 19 October 2016.
8. Guide to the General Data Protection Regulation (GDPR). The Information Commissioner's Office.
9. Kim J, Edge MD, Algee-Hewitt BFB, et al. Statistical Detection of Relatives Typed with Disjoint Forensic and Biomedical Loci. Cell. 2018; 175(3): 848-58.e6.
10. Gymrek M, McGuire AL, Golan D, et al. Identifying personal genomes by surname inference. Science. 2013; 339: 321-24.
11. El Emam K. Arbuckle L. De-identification: a critical debate. Future of Privacy Forum. July 24 2014.
12. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications. 2019; 10(1): 3069.
13. Anonymisation: managing data protection risk code of practice. The Information Commissioner's Office. 2012.
14. Case C-582/14 Breyer v Bundesrepublik Deutschland, [46].
15. Anonymisation: managing data protection risk code of practice. The Information Commissioner's Office; 2012. p21.
16. Mark Phillips. Can Genomic Data Be Anonymised? GA4GH. 10th October 2018.
17. Shabani M, Borry P. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. European Journal of Human Genetics. 2018; 26(2): 149–56.
18. Is pseudonymised data still personal data? The Information Commissioner's Office.
19. Case C-582/14 Breyer v Bundesrepublik Deutschland. 19 October 2016.
20. Mourby M, Mackey E, Elliot M, *et al*. Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. Computer Law & Security Review. 2018; 34(2): 222–33.
21. Article 29 Data Protection Working Party. Opinion 4/2007 on the Concept of Personal Data: p12.
22. Dove ES. The EU General Data Protection Regulation: Implications for International Scientific Research in the Digital Era. The Journal of Law, Medicine & Ethics. 2019; 46(4): 1013–30.
23. Case C-210/16 Wirtschaftsakademie, 5 June 2018.
24. Case C-25/17 Tietosuojavaltuutettu, 10 July 2018.
25. Case C-40/17 Fashion ID, 29 July 2019.

phg foundation
making science work for health

PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN

+44 (0) 1223 761900

@phgfoundation

www.phgfoundation.org

UNIVERSITY OF CAMBRIDGE