European Legal Requirements for Use of Anonymized Health Data for Research Purposes by a Data Controller with Access to the Original (Identified) Data Sets

Mike Hintze
Khaled El Emam

17th August 2017
Contents

1. Introduction .................................................................................................................................................. 3
2. Can a Data Set be Anonymous if Held by a Data Controller That Also Holds the Original Data Set?... 4
3. Research Uses of De-Identified Health Data Where the Data May Not be Considered Fully Anonymous ................................................................................................................................. 6
   3.1. Consent and the Legal Basis for Processing.................................................................................. 7
   3.2. Other Compliance Benefits of Strong De-Identification .................................................................... 8
4. Conclusion ................................................................................................................................................... 9
5. About the Authors ....................................................................................................................................... 10

Disclaimer: This White Paper is for informational purposes only and is not intended to, nor shall it be construed as, providing any legal opinion or conclusion; does not constitute legal advice; and is not a substitute for obtaining professional legal counsel from a qualified attorney on your specific matter.
1. Introduction

The use of health information for research purposes is widespread and leads to innumerable benefits for society. Medical breakthroughs, the development of new and more effective pharmaceuticals, greater efficiencies and cost reductions in the delivery of health care, and providing better and more effective information to patients and consumers are among the valuable results of such research. And in an age of “big data,” the positive potential of such research and data analytics increases dramatically.

Researchers handling sensitive health data face the challenge of maximizing these beneficial results while protecting the privacy of individual patients. Recent developments in European data protection law, including the adoption of the General Data Protection Regulation (GDPR), highlight the need to address this challenge in a way that achieves both of these important objectives. A key to such a “win-win” outcome is the use of de-identification or anonymization.¹

Current EU law sets a high bar for what data can be considered fully anonymous. The GDPR, which comes into effect in May 2018, appears to retain a high bar for anonymity, but also creates the foundation for a more nuanced and flexible approach. In either case, data that meets the “anonymity bar” is no longer subject to data protection law. In determining whether a data set can be considered anonymous, several factors must be taken into account, including the anonymization methodology employed and the context of the anonymization. For example, under some conditions, data can be considered anonymous if held by one party, but non-anonymous if held by a party with a greater practical ability to re-identify the data.

One common scenario in the field of health research occurs when an entity collects health information in an identified state as part of providing health services to patients. It then seeks to use that data for research purposes. And it wishes to carry out that research in a responsible way that protects individual patient privacy, complies with legal obligations, and maintains the utility of the data to preserve the value and integrity of the research.

For example, a hospital or academic medical center may seek to extract data from patient records, apply an anonymization technique to the data, and use the resulting data set to conduct research aimed at improving patient care. Or an electronic medical records (EMR) vendor may wish to analyze anonymized patient data from clinics using the EMR service to improve the service or develop better solutions (such as predictive models aimed at improving diagnostic tests or drug responses).

In addition to using a valid anonymization technique that has been vetted and accepted by experts in the field, this scenario assumes the data controller will put in place strict controls to prevent the re-identification of the anonymized data set. Such controls would typically include policies against any attempt to re-identify data subjects from the anonymized data set, access controls on both the anonymized data set and the original data set (such that researchers accessing the anonymized data set would not have access to the original data set), and monitoring and auditing to ensure the policies are followed and the controls are effective. Such a data flow could look something like the diagram in Figure 1.

¹ For the purposes of this paper, we use “de-identification” as a general term that includes the full spectrum of methods, encompassing both pseudonymization and anonymization. “Pseudonymisation” and “anonymisation” are used as they are used in the GDPR, with anonymisation indicating the strongest form of de-identification such that fully anonymised data is no longer “personal data” subject to data protection law.
Figure 1: Illustration of how a controller can hold identifiable and anonymized data while maintaining a firewall between the two types of data.

This scenario presents two key questions.

1. Can an anonymized data set still be considered anonymous when held by a data controller with access to the original (identified) data set?

2. If there are circumstances in which the data cannot be considered anonymous in this context, then what is required to enable the use of the data for research purposes in a compliant manner?

The sections below analyze these two questions in the context of this scenario.

2. Can a Data Set be Anonymous if Held by a Data Controller That Also Holds the Original Data Set?

European data protection law applies to “personal data,” which is defined, in part, as “any information relating to an identified or identifiable natural person.” Data which has been anonymized is no longer “personal data” and is therefore not subject to the requirements of data protection law. Regulators and
courts interpreting these terms have set a high bar for what qualifies as fully anonymized data under current data protection law based on the 1995 Data Protection Directive.

The Article 29 Working Party (WP29) 2014 Opinion on Anonymization Techniques states that, taking into account all means “likely reasonably” to be used to re-identify the data, anonymization must be “irreversible” and “as permanent as erasure.” The opinion provides the example that “when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data.” This example suggests that in the scenario raised in this paper, the original data controller could not consider an anonymized data set to be truly anonymous when that controller retains the original data set.

More recent guidance from the Irish Data Protection Commissioner, however, specifically recognizes and addresses this scenario in a way that appears to offer some flexibility. It states:

If the data controller retains the raw data, or any key or other information which can be used to reverse the anonymisation process and to identify a data subject, identification by the data controller must still be considered possible in most cases. Therefore, the anonymised data must normally still be considered personal data, and should only be processed in accordance with the Data Protection Acts. Where data has been anonymised to such an extent that it would not be possible to identify an individual in the anonymised data even with the aid of the original data, the anonymised data is not considered personal data. This might occur where the data is in an aggregated statistical format, or where random noise added to the data is such as to completely prevent a linkage between the original data and the anonymised data from being made.

Thus, the Irish guidance does not set out an absolute rule that an anonymized data set will always be personal data if in the hands of a data controller that also has the original source data. Rather, it says re-identification by the data controller must be considered possible in “most” cases, and the data must “normally” be considered personal data in such cases. Further, it specifies a contrary example in which the data can be considered anonymous – i.e., where the anonymization method used would prevent the identification or singling out of an individual “even to someone in possession of the source data.”

The guidance on this issue from regulators is based on current European privacy law, and the GDPR provides an opportunity to examine this scenario anew. While the GDPR appears to retain a similarly high standard for anonymity, it also suggests an openness to a more flexible approach that puts more focus on context and reasonableness.

The GDPR provides some additional guidance in its recitals. For instance, Recital 26 in the GDPR is more expansive than the equivalent recital in the 1995 Data Protection Directive. It reads, in part:

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.

---

2 Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 0829/14/EN (WP216), at 6.
3 Id., at 9.
4 https://dataprotection.ie/viewdoc.asp?DocID=1594&ad=1
The second sentence in this language is new and suggests that many factors must be considered in determining the likelihood that an anonymization method would be reversed. In particular, the reference to “all objective factors” must be read as including the context of the processing. And in real-world scenarios, that context necessarily includes factors such as the methodology employed, whether the data is closely held within a data controller or is publicly released, and the additional safeguards designed to prevent identification of individuals from the anonymized data set. Collectively, the consideration of “all” such factors suggests a “reasonableness standard” rather than the “impossibility standard” that seems to have taken hold under current law.

Further, the GDPR contains new provisions that recognize differing intermediate levels of de-identification. Several provisions include an explicit recognition of pseudonymization as a method of reducing risk. Additionally, Articles 11 and 12 refer to a level of de-identification that falls short of full anonymization, but enables the data controller to “demonstrate that it is not in a position to identify the data subject.” Collectively, the provisions of the GDPR reflect a recognition that there is a spectrum of de-identification.

These updates to the law provide an opportunity for a more flexible and nuanced approach across the full spectrum of de-identification, including where to draw the line between personal data and anonymous data, taking onto account context and safeguards. Under such an approach, it should be possible to conclude that in at least some contexts, data anonymized and used for research purposes can still be considered anonymous even when the controller retains the original data set. The scenario discussed in this paper provides what is perhaps the strongest case for such a conclusion. The anonymized data set is not released publicly or widely shared, a robust anonymization method is used that has been vetted by experts in the field, and strong safeguards are in place to keep the data set separate and otherwise prevent the identification of data subjects from the anonymized data set.

Such an interpretation and approach will encourage research that will inevitably result in enormous benefits to public health and welfare. And a fallback safeguard is always in place – if the data does become re-identified, it will come back within the scope of the GDPR and all the appropriate protections of data subjects’ rights and freedoms will apply.

It is foreseeable that different Data Protection Authorities will view this scenario with differing levels of pragmatism and flexibility.

3. Research Uses of De-Identified Health Data Where the Data May Not be Considered Fully Anonymous

Despite the circumstances discussed above under which data may be considered fully anonymized even if the data controller also retains the original data set, in some cases the data controller may conclude that it should treat the “anonymized” data as personal data. In such cases, the data can nevertheless still be used for research purposes.

The difference is that when the data is still considered personal data under the circumstances the data controller will need to meet certain legal obligations to use the data. And the strong de-identification and other safeguards described in the scenario above are still important because they will go a long way toward meeting those obligations. Some key GDPR obligations are discussed below.
3.1. Consent and the Legal Basis for Processing

Under European data protection law, processing personal data requires a legal basis, such as the explicit consent of the data subject or the “legitimate interests” of the data controller. In the context of research, obtaining explicit consent from each individual data subject is often impractical and could undermine the statistical validity of outcomes. Thus, establishing an alternative legal basis is often necessary in the context of research.

The GDPR provides alternatives to obtaining consent that can apply in the context of research – particularly where the data is protected by strong de-identification. It sets out criteria for when a secondary use of data (such as for research or analysis) can proceed on a basis other than the consent of the data subject.

First, Recital 50 of the GDPR notes that:

The processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected. In such a case, no legal basis separate from that which allowed the collection of the personal data is required. . . . Further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be considered to be compatible lawful processing operations.

This language is reflected in Article 5(1)(b) which provides that:

[Personal data shall be] collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes (‘purpose limitation’).

Article 89(1) states:

Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.

And finally, Article 6(4) provides:

Where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject’s consent . . . the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia . . . the possible consequences of the intended further processing for data subjects [and] the existence of appropriate safeguards, which may include encryption or pseudonymisation.

Based on these provisions, scientific research is likely to be considered a purpose “compatible” with the purpose(s) for which the data was originally collected. And where strong de-identification is applied to the data sets used for research purposes, the data controller can demonstrate that it has applied “appropriate safeguards” and that the likelihood of negative consequences on the data subject is
exceedingly low. The last two sentences of Article 89(1) suggest that the appropriate safeguards employed by the data controller should include the strongest form of de-identification that is compatible with the research purpose.

Questions may arise regarding the applicability of these provisions based on the purposes and nature of the research being conducted. Health data may be used for a variety of research purposes. Some research may be focused on promoting public health. Some may be academic research aimed at advancing scientific knowledge. Some may be designed to develop drugs or medical devices in the life sciences industry. Some may be to monitor the safety of a drug or device after it has been approved and marketed. Some may focus on developing commercial health applications or services. And some may be aimed at improving the effectiveness of information or marketing messages provided to consumers. Common sense and practical experience may lead data controllers to conclude that regulators are likely to look more favorably upon research purposes that are closer to the “purely academic” end of the spectrum, or where a strong public benefit to the research can be demonstrated. And there is some basis in the text of the GDPR for concluding there is a preference for research in the public interest.  

However, the lines between academic or public-interest research and commercial research are not clear or obvious. Much research performed by commercial entities promotes public interests, such as advancing scientific knowledge and furthering public health. Significantly, the GDPR itself suggests the lines between academic and commercial research are not determinative, with Recital 159 stating: “For the purposes of this Regulation, the processing of personal data for scientific research purposes should be interpreted in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research.” Thus, the conclusions may not be fundamentally different for commercial research vs. purely academic research.

In sum, data controllers conducting research on data that has been strongly de-identified have a strong case under the GDPR for relying on a legal basis other than consent (such as legitimate interests), or no additional legal basis at all.

3.2. Other Compliance Benefits of Strong De-Identification

The GDPR includes a number of obligations that require data controllers and data processors to implement technical and organizational measures designed to protect data subjects and reduce risk. In each of these cases, the use of strong de-identification can be an important part of meeting these obligations.

One key example is the new set of requirement introduced by the GDPR which are referred to as “data protection by design and by default.” These new rules require data controllers to “implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to

For instance, under Article 9(2)(i), the higher level of restrictions on the processing of special categories of sensitive personal data (including health data) don’t apply where the “processing is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health or ensuring high standards of quality and safety of health care and of medicinal products or medical devices.”

It is worth noting, however, that many uses of data flowing from research for commercial purposes will raise additional legal obligations. For example, if an output of research is a better algorithm for tailoring marketing messages to consumers, the company wishing to send those tailored marketing messages will need to comply with all the legal obligations that apply to direct marketing, including initial consent, providing users the ability to stop receiving such messages at any time, etc. Thus, organizations need to be aware of the regulatory obligations applicable to all subsequent data uses.
implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.” In the scenario discussed in this paper, applying de-identification to a data set used for research, along with safeguards against linking back to the original data set or other means of re-identifying the data, provide a textbook example of data protection by design and by default. And the stronger the method of de-identification, the stronger the case that this obligation has been fulfilled.

Similarly, strong de-identification with additional safeguards can be seen as meeting the data security obligations as well. As under current data protection law, controllers and processors handling personal data are obligated under the GDPR to implement measures sufficient “to ensure a level of security appropriate to the risk.” And the strength of de-identification applied will clearly be a relevant factor in evaluating the level of risk posed by personal data. In fact, the Article 29 Working Party has described de-identification as a security precaution. In the scenario discussed in this paper, the safeguards against re-identification of data should consider both well-intentioned researchers and malicious actors, as well as threats both inside and external to the organization.

Closely related to the proactive steps of data security are the reactive steps organizations must take in the event of a data breach. The GDPR introduces new requirements to notify supervisory authorities and/or data subjects in the event of a breach of personal data. Supervisory authorities must be notified “unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons.” And data subjects must be notified if “the personal data breach is likely to result in a high risk to the rights and freedoms of natural persons.” As with data security, the risk assessment for these provisions will certainly take into account the level of de-identification of the data. In the event of a data breach, fully identified personal data will almost always pose a greater risk than if that data were de-identified. Thus, the need for notification in the event of a data breach is far less likely if the data is strongly de-identified.

In sum, if it is found that the data can no longer be considered anonymous in the scenario described above, and, therefore, the data is subject to GDPR obligations, the use of the strong de-identification along with the additional safeguards in place, will allow the organization to meet the key GDPR obligations.

In each of these cases, the stronger the de-identification method, the stronger the legal position of the data controller will be. For example, very strong de-identification methodology will be seen as more thoroughly meeting a data controller’s data security or “data protection by design and by default” obligations than will a relatively weak pseudonymization implementation.

Of course, the obligations discussed above are not the only obligations imposed by the GDPR. However, in the scenario discussed in this paper, where the data controller retains the original, identified data set, the controller will already be subject to those other obligations. For example, it likely will be required by the GDPR to maintain a publicly-facing privacy notice, or to appoint a Data Protection Officer (DPO). But with respect to such requirements, the research uses of this anonymized / de-identified data set may have a marginal impact on how those requirements are carried out, but they typically would not create new or additional compliance obligations.

4. Conclusion

This paper addresses the common scenario in which health data is anonymized and used for research purposes within a data controller that retains the original data set. In such cases, robust anonymization
combined with strong safeguards to protect the anonymized data from being re-associated with the original data or otherwise re-identified, creates a strong case under the GDPR that the data should still be considered fully anonymous and therefore outside the scope data protection law. But even where that is not the case, the same strong anonymization methodology and safeguards will enable the data controller to meet key GDPR obligations. In either case, strong de-identification is an essential tool for enabling the use of sensitive health data for research purposes.

5. About the Authors

**Mike Hintze**

Mike Hintze is a partner at Hintze Law PLLC. As a recognized leader in the field, he advises companies, industry associations, and other organizations on global privacy and data protection law, policy, and strategy. He was previously Chief Privacy Counsel at Microsoft, where, for over 18 years, he counselled on data protection compliance globally, and helped lead the company’s strategic initiatives on privacy differentiation and public policy. Mike also teaches privacy law at the University of Washington School of Law, serves as an advisor to the American Law Institute’s project on Information Privacy Principles, and has served on multiple advisory boards for the International Association of Privacy Professionals and other organizations. Mike has testified before Congress, state legislatures, and European regulators; and he is a sought-after speaker and regular writer on data protection issues.

Prior to joining Microsoft, Mike was an associate with Steptoe & Johnson LLP, which he joined following a judicial clerkship with the Washington State Supreme Court. Mike is a graduate of the University of Washington and the Columbia University School of Law.

Privacy Analytics Inc. is a client of Hintze Law PLLC, and support for Mike Hintze’s contribution to this White Paper has been provided by Privacy Analytics Inc. Neither Mike Hintze’s representation of Privacy Analytics Inc. nor his contribution to this White Paper serves as an endorsement of any technology, products or services of Privacy Analytics Inc.

**Khaled El Emam**

Dr. Khaled El Emam is the founder of Privacy Analytics Inc. and Director of Real World Evidence Solutions. As an entrepreneur, Khaled helped found five companies involved with data management and data analytics. He has worked in technical and management positions in academic and business settings in England, Scotland and Japan.

Khaled is also a senior scientist at the Children’s Hospital of Eastern Ontario (CHEO) Research Institute and Director of the multi-disciplinary Electronic Health Information Laboratory (EHIL) team, conducting academic research on de-identification and re-identification risk. He is a world-renowned expert in statistical de-identification and re-identification risk measurement. He is one of only a handful of individual experts in North America qualified to anonymize Protected Health Information under the HIPAA Privacy Rule.

In 2003 and 2004, Khaled was ranked as the top systems and software engineering scholar worldwide by the Journal of Systems and Software based on his research on measurement and quality evaluation and improvement.

Previously, Khaled was a Senior Research Officer at the National Research Council of Canada. He also served as the head of the Quantitative Methods Group at the Fraunhofer Institute in Kaiserslautern, Germany.
Khaled was one of the first Privacy by Design Ambassadors recognized by the Ontario Information and Privacy Commissioner. He previously held the Canada Research Chair in Electronic Health Information at the University of Ottawa and is an Associate Professor in the Faculty of Medicine at the University. He has a PhD from the Department of Electrical and Electronics Engineering, King’s College, at the University of London, England.