Comparing the benefits of pseudonymisation and anonymisation under the GDPR

Received: 7th October, 2018

Mike Hintze
is a partner at Hintze Law PLLC. As a recognised leader in the field, he advises companies, industry associations and other organisations 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 adviser 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 organisations. 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.

Hintze Law, 505 Broadway E. #151, Seattle, WA 98102, USA
E-mail: mike@hintzelaw.com

Khaled El Emam
is the founder and President of Privacy Analytics, an IQVIA company. As an entrepreneur, Khaled founded or co-founded five companies involved with data management and data analytics. He has worked in technical and management positions in academic and business settings in the UK, Germany, Japan and Canada. Khaled is also a senior scientist at the Children's Hospital of Eastern Ontario (CHEO) Research Institute and director of the multidisciplinary Electronic Health Information Laboratory (EHIL) team, conducting academic research on de-identification and re-identification risk. He is a recognised expert in statistical de-identification and re-identification risk measurement. Khaled was one of the first Privacy by Design Ambassadors recognised by the Ontario Information and Privacy Commissioner. 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. He previously held the Canada Research Chair in Electronic Health Information at the University of Ottawa and is a professor in the Faculty of Medicine (Pediatrics) at the university. He has a PhD from the Department of Electrical and Electronics Engineering, King's College, at the University of London, UK.

Privacy Analytics, 251 Laurier Avenue W, Suite 200, Ottawa, Ontario, Canada, K1P 5J6
E-mail: kelemam@privacy-analytics.com

Abstract Many organisations are trying to obtain more value from their data to improve their products and services, offer new ones and optimise their own internal operations. For example, more chief data officers, or similar roles, are being created to drive such data-enabled transitions. With the General Data Protection Regulation (GDPR) in place, these organisations need to determine the lawful basis for such activities. De-identification techniques, such as pseudonymisation and anonymisation, can play an important role in facilitating such secondary uses and disclosures of data. In regard to de-identification, the GDPR introduces nuances that have not previously been seen, recognising the existence of different levels of de-identification and explicitly adding references to pseudonymisation as an intermediate form of de-identification. This paper explores the nuances introduced by the GDPR, compares the benefits of the different levels of de-identification found in the regulation, and provides practical guidance for using de-identification as a tool for addressing different GDPR compliance obligations.

KEYWORDS: anonymisation, pseudonymisation, de-identification, masking, secondary uses

INTRODUCTION

Organisations are examining ways to obtain more value from their data to optimise their operations, and develop or improve their products and services. This is evidenced by the rise in the new chief data officer role, tasked with enabling this type of transformation.[1] With the General Data Protection Regulation (GDPR)[2] in place since May 2018, these organisations need to determine the lawful basis for such activities.

De-identification techniques, such as pseudonymisation and anonymisation, can play an important role in facilitating such secondary uses and disclosures of data. The GDPR addresses de-identification in a nuanced way: while it maintains a high standard for achieving anonymisation, it recognises the existence of different levels of de-identification and it explicitly adds references to an intermediate form of de-identification — namely, pseudonymisation.

We use the terms ‘‘de-identification’, ‘pseudonymisation’ and ‘anonymisation’. For the purposes of this discussion, we use ‘‘de-identification’’ as a general term that includes the full spectrum of methods, from simple pseudonymisation to full anonymisation.

In this paper, we explore the nuances introduced by the GDPR, compare the benefits of the different levels of de-identification found in the regulation, and provide practical guidance for using de-identification as a tool for addressing different GDPR compliance obligations.

Psuedonymisation under the GDPR

‘Pseudonymisation’ commonly refers to a de-identification method that removes or replaces direct identifiers (eg, names, phone numbers, government-issued ID numbers, etc.) from a data set, but may leave in place data that could indirectly identify a person (often referred to as quasi-identifiers or indirect identifiers).[3] This replacement can be done by assigning random values or by using cryptographic techniques. Applying transformations to direct identifiers, and nothing else, is the elemental requirement for any pseudonymisation method.

Pseudonymisation is defined in the GDPR as:

> 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.[4]

Recital 26 of the GDPR notes that pseudonymised information ‘could be attributed to a natural person by the use of additional information’. This re-enforces the interpretation that indirectly identifying information can remain in such data.

To meet the GDPR definition of pseudonymisation, additional controls or ‘technical and organizational measures’ referred to above must also be implemented in addition to the replacement of the direct identifiers.

Because of the range of different technical and organisational measures that could be employed, however, pseudonymisation is itself a broad concept that encompasses a range of methods and strengths. Further, the GDPR recognises an intermediate threshold of de-identification such that the data controller ‘is not in a position to identify the data subject.’[5] Thus, it is possible to read into the GDPR at least two levels of pseudonymisation:

- *Basic pseudonymisation*. Under basic pseudonymisation, the direct identifiers are transformed and appropriate controls are put in place to ensure that cryptographic keys are stored and handled appropriately.
- *Strong pseudonymisation*. This is a superset of basic pseudonymisation where some indirect identifiers in the data are also perturbed[6] and any cryptographic keys are destroyed, making pseudonymisation irreversible. With strong pseudonymisation it is harder (than with basic pseudonymisation) to attribute data to a natural person.

Pseudonymised data remains 'personal data' and is therefore subject to the requirements of the GDPR and the appropriate security and privacy controls suitable for handling personal data. But the GDPR provides some regulatory incentives to adopt pseudonymisation and there are therefore some significant benefits to employing it. Specifically, pseudonymising data can help an organisation meet some of the GDPR requirements, but it does not fully release the organisation from them. Strong pseudonymisation helps an organisation meet more of the GDPR requirements than basic pseudonymisation.

Anonymisation under the GDPR

By contrast, 'anonymisation' as used in this paper refers to an even stronger form of de-identification. For the purposes of this paper, we will assume that strong anonymisation methods are being used and that these methods would be considered acceptable by European data protection authorities (DPAs). In Box 1, 'Basic principles of risk-based anonymisation', we provide more details about what such methods entail and specifically risk-based anonymisation methods that balance protecting individual identity against maintaining data utility. Under a risk-based anonymisation approach, data is claimed to be anonymous as a function of *both* data perturbations *and* additional technical and administrative controls that are put in place.

Fully anonymised data that meets the legal bar set by European data protection law is no longer 'personal data' and is therefore not subject to the obligations of the GDPR at all. Thus, the benefits of pseudonymisation pale in comparison to the benefits of full anonymisation.

The identifiability spectrum

When identifiability is viewed as a spectrum,[7] with one end signifying identified data and the other end signifying anonymised data, then the distinctions noted above suggest four different regions on this spectrum, as illustrated in Figure 1.

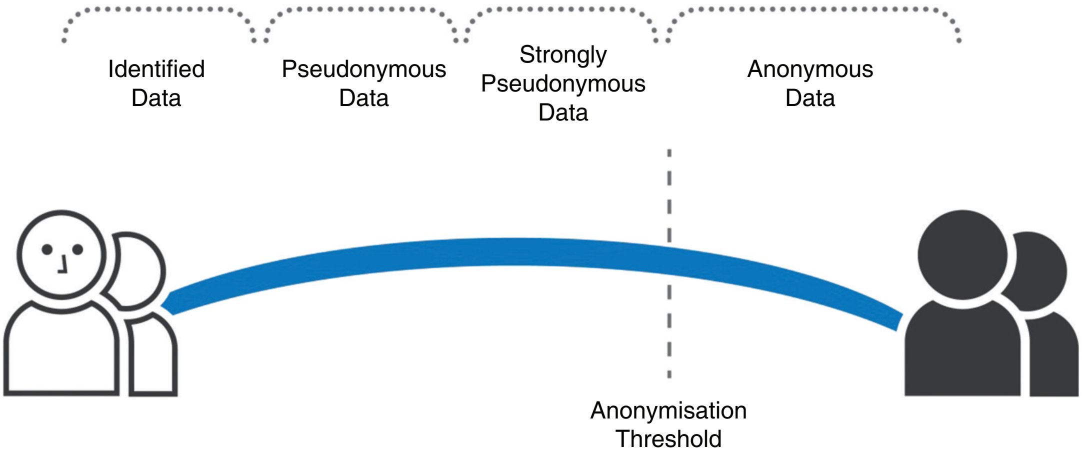

Figure 1: The four different regions on the identifiability spectrum

BOX 1: BASIC PRINCIPLES OF RISK-BASED ANONYMISATION

Risk-based anonymisation methods are consistent with recommendations from the Information Commissioner's Office in the UK,[8] anonymisation guidance from the European Medicines Agency,[9] the privacy commissioner of Ontario,[10] the expert determination methods under the HIPAA Privacy Rule in the USA[11] and other governmental, academic, and professional associations and groups globally.[12]

There are three fundamental concepts underpinning risk-based anonymisation methods.

The first concept differentiates among the different types of information in data. Data may have direct identifiers, which are things like a data subject's government-issued identification number. Direct identifiers are assumed to directly identify a patient with a high probability of success. These are typically pseudonymised or removed to create pseudonymous data. This type of data is still considered personal information.[13] Another type of information would be quasi- (or indirect) identifiers. These are things like demographics (eg, age, gender and race), socioeconomic information (eg, income and years of education) and important life events (eg, marriage, births, rare diagnoses and hospital visits). There is evidence that quasi-identifiers can still identify individuals.[14] Transforming this kind of information can produce anonymous data.

Therefore, the act of anonymisation is focused on the quasi-identifiers only. The assumption is that pseudonymisation has already been applied to address re-identification risks from direct identifiers.

The second is that risk-based methods are quantitative. The quantity that is being measured is the risk of re-identification of an individual in the data. The initial step is to set an acceptable threshold for this risk. This means that the acceptable risk of re-identification is going to be some value larger than zero. Setting the threshold at zero risk means that no useful data will be retained because the level of perturbation of the data will be quite high.

The actual risk of re-identification is then measured on the data. This measured value is compared to the threshold. If the measured risk is above the threshold then the data is not considered anonymous. If the measured risk is below the threshold then the data is considered anonymous. If the data is not anonymous then various transformations can be applied to bring the measured risk below the threshold. These transformations may include generalising certain values in the data (eg, generalising a date of birth to a year of birth) or suppressing certain values in the data that make individuals stand out.

The third concept pertains to the context of the data. The actual risk of re-identification is a function of both the data and the context. The context represents the security, privacy and contractual controls that are in place. For example, one context can be a public data release (eg, an open data initiative). Another context would be a researcher who analyses the data in a very secure enclave. These are two very different contexts and the risk of re-identification is different in each of these, even for the same data. Therefore, the context consists of characteristics of the data recipient/holder, the contract or data use agreement and the data itself.[15]

The overall risk is a function of both the data risk and the context risk. When expressed as probabilities, the overall risk of re-identification is the multiplication of these two numbers.

This means that the same data can have different levels of risk if it is processed in different contexts. But it also means that the same data can have different risk levels as it moves from one organisation to another in the same data flow (ie, over time). For example, if the data moves from an organisation performing analytics to a say, a researcher, the risk may be low in the first instance but increase in the second instance after the transfer.

To illustrate this point, we will look at 11 key GDPR obligations to see how the different types of pseudonymisation and anonymisation affect their applicability and provide practical guidance for compliance. The next section begins with a summary table listing these obligations, followed by a discussion of each.

DIFFERENT OBLIGATIONS UNDER THE GDPR

Table 1 summarises the benefits of each type of de-identification and the extent to which each obligation applies.

ANALYSIS OF OBLIGATIONS

Notice to data subjects

A basic principle of data protection law is transparency and the obligation to provide notice to data subjects regarding the collection, use and disclosure of personal information. Under the GDPR, data controllers must provide extensive details to data subjects whenever they process personal data (see Box 2, 'Notice required to data subjects').[16] The text of the GDPR makes no distinction between fully identified personal data and pseudonymised personal data. Thus, the full range of mandated disclosures apply to any use or other processing of pseudonymised data. By contrast, because anonymised data is no longer considered personal data, none of the notice obligations apply to uses of data that has been fully anonymised.

Lawful basis for processing

The GDPR requires there to be a lawful basis to process personal data.[34] The most well-known basis is the explicit consent of the data subject;[35] however, under the GDPR, obtaining explicit consent can be difficult and in some scenarios, such as research, big data analytics and machine-learning projects, obtaining explicit consent may be impractical or impossible. Furthermore, there is evidence

Table 1: GDPR obligations for different types of pseudonymised and anonymised data

GDPR obligation	Type of data			
	Identified	**Pseudonymised (basic)**	**Strongly pseudonymised**	**Anonymised**
1. Provide notice to data subject	Required	Required	Required	Not required
2. Legal basis for processing (legitimate interests, consent)	Required	Stronger case for legitimate interests	Much stronger case	Not required
3. Data subject rights (access, portability, rectification)	Required	Required	Not required	Not required
4. Give right to erasure/right to be forgotten	Required	Required	May not be required	Not required
5. Basis for cross-border transfers	Required	Required	Required	Not required
6. Data protection by design	Required	Partially met	Strengthens the ability to meet this obligation	Not required
7. Data security	Required	Partially met	Strengthens the ability to meet this obligation	Not required
8. Data breach notification	Likely to be required	Less likely to be required	Strengthens the case that notification is not required	Not required
9. Data retention limitations	Required	Required	Required	Not required
10. Documentation/recordkeeping obligations	Required	Required	Required	Not required
11. Vendor/sub-processor management	Required	Required	Required	Not required

BOX 2: NOTICE REQUIRED TO DATA SUBJECTS

Articles 13, 14 and 15 of the GDPR set out a long list of items that organisations must include in privacy notices provided to individual data subjects. In this respect, the GDPR represents a significant change from the 1995 Data Protection Directive that it replaced, which specified a much more limited set of information that must be included in a privacy notice. Under the GDPR, notices must include:

- The identity and the contact details of the controller and, where applicable, of the controller's representative.[17]
- The contact details of the data protection officer, where applicable.[18]
- Where personal data is obtained from a source other than the data subject:
    - the types of personal data obtained;[19] and
    - the source(s) 'from which the personal data originate, and if applicable, whether it came from publicly accessible sources'.[20]
- Where the personal data is collected from the data subject, whether providing the data is required, including:
    - whether it is a requirement necessary to enter into a contract;
    - whether it is otherwise required by statute or contract; and
    - the possible consequences of the failure to provide such data.[21]
- The intended purposes of processing the personal data.[22]
- The legal basis for the processing.[23]
- Where the legal basis for processing is 'the legitimate interests pursued by the controller or a third party under Article 6(1)(f)', a description of those interests.[24]
- Where the legal basis for processing is 'the consent of the data subject under Articles 6(1)(a) or 9(2)(a),' the existence of the right to withdraw such consent at any time (which will not affect the lawfulness of any processing that occurred before such consent is withdrawn).[25]
- Where personal data is used for automated decision making, including profiling, referred to in Article 22(1) and (4), the existence of such processing, meaningful information about the logic involved, and the significance of the processing and any anticipated consequences for the data subject.[26]
- 'The recipients or categories of recipients of the personal data, if any'.[27]
- The period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period.[28]
- The existence of the right of a data subject to:
    - request from the controller 'access to and rectification or erasure of personal data'; or
    - object to the processing of personal data or obtain a restriction of such processing under certain circumstances.[30]
- Receive data he or she has provided to the controller in a structured, commonly used and machine-readable format, and transmit that data to another controller (data portability).[31]
- The right to lodge a complaint with a supervisory authority.[32]
- Where the controller intends to transfer personal data to a third country or international organisation, the fact of such transfer and either:
    - the existence or absence of an adequacy decision by the [European] Commission; or
    - in the case of transfers based on 'suitable safeguards' under Articles 46, 47 or 49(1)(b) (such as contractual provisions or binding corporate rules), a description of such safeguards and how to obtain a copy of them.[33]

that there are systematic differences between consenters and non-consenters in some domains, which could have an impact on the interpretability of the data.[36]

Nevertheless, there is language in the GDPR that 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 — in particular where the 'processing for another purpose is compatible with the purpose for which the personal data are initially collected.'[37] One of the key criteria to be used in determining whether such processing can proceed is 'the existence of appropriate safeguards, which may include encryption or pseudonymisation.'[38] Thus, the use of pseudonymisation, at least in some circumstances, can help enable data processing for secondary purposes without the need to obtain the explicit consent of the data subjects.

Further, 'legitimate interests' is a lawful basis frequently relied on as an alternative to consent. Under the GDPR, this basis may apply where the 'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.'[39] Inherent in these criteria is a balancing test between the interests of the data controller and the rights and freedoms of the data subject. Because pseudonymisation can, in some cases, reduce the risks to the rights and freedoms of data subjects, its use may help support a case for relying on legitimate interests as a basis for processing.

If the controller is seeking to rely on a lawful basis other than consent, such as legitimate interests, arguably strong pseudonymisation would strengthen the case for relying on such a basis (compared to basic pseudonymisation).

The Article 29 Working Party has made clear that the processing of personal data to fully anonymise such data is 'compatible with the purpose for which the personal data are initially collected' and therefore does not require an additional lawful basis.[40] And once the data is fully anonymised, it is outside the scope of data protection law and data controllers have no need to have or articulate a lawful basis for further processing.

Data subject rights

The GDPR gives data subjects a number of rights to their data — including the rights of access, rectification and data portability, as well as rights to object to the processing of personal data or obtain a restriction of such processing under certain circumstances.[41]

Here, the implementation challenges can be profound. But Articles 11 and 12(2) of the GDPR specify that if the data controller can demonstrate that it is not in a position to identify the data subject from the data it holds, it need not comply with the articles setting out these data subject rights. This provision reflects the commonsense proposition that if the data controller cannot reliably tie the data it holds to the individual asserting this right, it will be unable to comply.

Given that the nature of the data subject access requirement creates a risk of disclosing personal data to the wrong individual (including malicious actors seeking to steal personal data), data controllers must identify the data subject with certainty before providing access.

Thus, in practice, to exercise these rights the data subject would have to provide some information to the controller to allow them to find the data subject's records. A relevant question, therefore, is what level of detail can a data subject reasonably provide to help the controller find the data subject's records?

The text of Article 11 of the GDPR on processing which does not require identification states:

(1) If the purposes for which a controller processes personal data do not or do no longer

require the identification of a data subject by the controller, the controller shall not be obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with this Regulation.

(2) Where, in cases referred to in paragraph 1 of this Article, the controller is able to demonstrate that it is not in a position to identify the data subject, the controller shall inform the data subject accordingly, if possible. In such cases, Articles 15 to 20 shall not apply except where the data subject, for the purpose of exercising his or her rights under those articles, provides additional information enabling his or her identification.

Furthermore, in the recitals of the GDPR, recital 57 states:

> the controller should not refuse to take additional information provided by the data subject in order to support the exercise of his or her rights. Identification should include the digital identification of a data subject, for example through authentication mechanism such as the same credentials, used by the data subject to log-in to the on-line service offered by the data controller.

The last part of Article 11, paragraph 2 implies that the data subject can provide the controller unlimited direct and indirect identifiers to allow the identification of their records. And the first quoted sentence of recital 57 suggests that a data controller cannot place a limit on the information provided by a data subject. But the second sentence suggests that a data controller can nevertheless refuse to grant a data subject access based on a type of information provided, and that sentence further suggests that the type of information contemplated by this provision is the traditional types of credential or authentication information.

In reality, the controller would need to limit the amount and type of information it can accept to enable identification of a data subject. The controller will likely not ask for very detailed transactional data from the data subject to identify their records because then the controller would be accused of making the process too complicated and consequently impeding the data subject's rights. Further, accepting an unlimited amount and scope of data, even if the data subject provides it, is unworkable practically and unduly puts a burden on data controllers. Such detailed information may be more likely to contain errors and many erroneous pieces of information can increase the risk of misidentification.

Consequently, the case can be made that the data subject would provide their direct and demographic indirect identifiers (ie, credentials used to log-in to an online service) as the primary information to identify their records.

In such a case, basic pseudonymisation would still enable the data controller to identify the appropriate records of the data subject. With basic pseudonymisation the indirect identifiers remain intact in the data. For example, if the data subject provides certain direct identifiers, along with their date of birth, gender, and postal code, that would allow the data controller to identify the pseudonymised record with certainty. If the data controller retains the cryptographic keys, then that can be used to encrypt (or hash) the credit card number or health insurance number that can be provided by the data subject and use it to match on the pseudonymised direct identifiers.

For strong pseudonymisation, the indirect identifiers that cover the demographics and other information that the data subject is likely to provide to exercise these rights would be de-identified. This means that there would be a strong potential for error in the matching process. Furthermore, the controller would not retain the cryptographic keys to allow matching on the direct identifiers. In such a case, a strong argument can be made that matching is not likely to work. And thus, the data controller would be much more likely to be able to demonstrate that they are not in a position to identify the data subject, as stated in Articles 11 and 12(2).

If de-identification is even stronger, such that the data is fully anonymised, it will be outside the scope of the GDPR and therefore free of these obligations.

Right to erasure/right to be forgotten

Article 17 of the GDPR creates a new right for data subjects to request that personal data about them be deleted. This right is referred to as the right to erasure, or the 'right to be forgotten.' If certain criteria are met, data controllers are required to respond to such requests and to erase the personal data 'without undue delay.' Further, if the controller has made the personal data public, it may be obligated to take 'reasonable steps' to inform other controllers that may be processing the personal data of the erasure request.

Many organisations are finding that implementing this so-called 'right to be forgotten' to be among the most onerous legal, technical and operational challenges in their GDPR compliance efforts. Locating all copies of such personal data in all systems requires detailed data mapping efforts. Creating a scalable ability to granularly delete data related to a particular individual often require re-architecture of data systems and the development of new tools. Determining when data must be deleted under the GDPR can require case-by-case review of the facts and a legal determination based on those facts. And operationalising the ability to receive erasure requests, authenticate the individuals making the requests, determining how to respond based on different criteria, locating and purging all copies of such data, and being able to demonstrate and document that such data has been fully erased will normally require a number of new resources and processes.

Nevertheless, Article 12(2) of the GDPR specifies that if the data controller can demonstrate that they are not in a position to identify the data subject from the data they hold, they need not comply with a request to erase data.

For the same reasons noted above, however, basic pseudonymisation methods are not likely to result in this exemption applying. Many implementations of basic pseudonymisation are readily reversible by a data controller and are likely to contain indirect identifiers that could allow the data controller to match the data if the data subject making the request supplies sufficient demographic data to allow the match. Therefore, basic pseudonymisation in such cases would not meet the standard reflected in Article 12(2).

On the other hand, strong pseudonymisation would make it very difficult for the data controller to identify the records that belong to the subject making the request.

In situations where deleting the data would not adversely affect any particular individual if the wrong person's data was inadvertently deleted, a case can be made that an attempt should be made to find a data subject's records even if strong pseudonymisation is in place. The chance and consequences of deletion of a wrong person's records would have to be assessed on a case-by-case basis, however.

Thus, whether pseudonymisation will result in relief from these GDPR obligations will depend on the strength of the method and implementation employed. And organisations asserting that the exemption applies will need to demonstrate or prove that their pseudonymisation methods meet the standard.

Of course, de-identification that is even stronger, such that it meets the bar for full anonymisation, will mean that the data will also be free of the obligation to delete.

Cross-border data transfers

Both existing European privacy law and the GDPR restrict the transfer of personal data outside the European Economic Area, except under certain conditions.[42] For example, personal data may be transferred to the small number of jurisdictions that the European Commission has found to have 'adequate' data protection regimes in place, transfers are allowed if subject to contracts that contain the model clauses approved by the Commission or transfers may take place if the recipient is part of the EU–U.S. Privacy Shield. Both fully identified and pseudonymised personal data (irrespective of the level of pseudonymisation) are equally subject to these restrictions. By contrast, these cross-border transfer restrictions do not apply to fully anonymised data.

Data protection by design and by default

A new requirement imposed by the GDPR is referred to as 'data protection by design and by default'. This new set of rules requires data controllers to 'implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to 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.'[43]

This provision applies to both identified and pseudonymised personal data, but given that pseudonymisation is called out as a key example of the types of measures required under this provision, data controllers can conclude that pseudonymising data will at least partially satisfy this requirement and that the number of additional measures to protect the data typically will be lower for pseudonymised data than for fully identified data. Strong pseudonymisation could contribute to a significantly better compliance story than basic pseudonymisation. For fully anonymised data, no such measures are required because the data is no longer subject to this requirement.

Data security

Controllers and processers handling personal data are obligated under the GDPR to implement measures sufficient 'to ensure a level of security appropriate to the risk.'[44] In gauging the level of risk posed by personal data, the level of de-identification applied is certainly a relevant factor. Thus, pseudonymised data will typically pose a lower risk than fully identified data, and therefore the level of security measures required will normally be reduced for pseudonymised data. In fact, the text of the GDPR suggests that pseudonymisation itself can be thought of as a security measure that safeguards personal data.[45] Strong pseudonymisation provides stronger protections than basic pseudonymisation and thus can be seen as a more robust security measure. Here too, fully anonymised data is different in that it is no longer subject to the GDPR requirements and therefore the security obligations do not apply.

Data breach notification

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.'[46] 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.'[47] Thus, where there is *some* risk, the supervisory authority must be notified; and where there is a *high* risk, both the supervisory authority and the affected data subjects must be notified.

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 pseudonymised. Thus, while the need for notification is likely in the event identified data is breached, it is less likely if pseudonymised data is breached. And the strength of pseudonymisation is relevant here too. In general, the stronger the pseudonymisation, the lower the risk will be.

As a practical matter, for the data involved in an incident not to be covered by notification obligations under the GDPR at all, all of the direct and quasi-identifiers would need to be addressed/transformed during de-identification. Partially de-identifying the information will not be likely to pass that test and therefore notification (at least to the supervisory authority) will likely have to be triggered.

By contrast, most US breach notification laws are triggered only if certain direct identifiers are present. For example, many state data breach notification laws are triggered by the presence of first and last name in combination with a social security number or other government-issued ID number, a financial account number, or (in some cases) health or medical information.[48] Under the HIPAA breach notification rules, health information that has been de-identified by removing specified direct and indirect identifiers will no longer trigger notification obligations. Thus, in some cases, the absence of certain direct identifiers is enough to completely release companies from any notification obligations at all. In other cases, the removal of certain direct and indirect identifiers will eliminate all notification obligations. Thus, it is likely that data controllers will point to that precedent as a basis for arguing that there should not be an obligation to notify data subjects when similar identifiers are removed as a means of pseudonymisation.

As with other provisions discussed in this paper, fully anonymised data, being outside the scope of the GDPR, does not trigger a breach notification obligation at all.

Data retention limitations

Data minimisation principles reflected in the GDPR require that personal data not be retained longer than necessary to carry out the legitimate purposes of processing. Thus, data controllers must evaluate their needs to retain data and establish appropriate retention schedules for the personal data they hold. Data controllers may argue that from a policy standpoint, they should have flexibility to retain pseudonymised data longer, particularly where the data is highly useful even if not strictly necessary. But those arguments to not relieve the data controller from the obligations to assess and establish data retention timeframes for all personal data they hold — whether identified or pseudonymous.

With respect to data retention, full anonymisation of data is considered the functional equivalent of deletion, and fully anonymised data may be kept indefinitely.

Documentation and recordkeeping obligations

The GDPR imposes far more documentation and recordkeeping obligations on data controllers and processers than is the case under current EU data protection law (see Box 3, 'Documentation and recordkeeping obligations'').[49] In almost all cases, these obligations apply equally to the processing of both fully identified and pseudonymised data.[50] By contrast, they do not apply to anonymised data.

Vendor or sub-processor management

A number of requirements in the GDPR apply to the use of vendors, processors, or sub-processors who handle or access personal data. Some dictate certain provisions that must be in a contract between a data controller and a data processor.[51] Others apply directly to data processors who are processing data on behalf of a controller.[52] Still others regulate the use of sub-processors and the obligation to pass through certain requirements to such entities.[53] Collectively,

BOX 3: DOCUMENTATION AND RECORDKEEPING OBLIGATIONS

Article 30 of the GDPR sets out specific records that must be retained by a data controller or processor with respect to the processing of personal data. For data controllers, these records include:

- the name and contact details of the controller and, where applicable, the joint controller, the controller's representative, and the data protection officer;
- the purposes of the processing;
- a description of the categories of data subjects and of the categories of personal data;
- the categories of recipients to whom the personal data have been or will be disclosed;
- where applicable, transfers of personal data to a third country or an international organisation, including the identification of that third country or international organisation, and the documentation of suitable safeguards as applicable;
- where possible, the envisaged time limits for erasure of the different categories of data; and
- where possible, a general description of the technical and organisational security measures referred to in Article 32(1).

Similar requirements are set out for data processors. Organisations that operate as both controllers and processors would need to maintain records for both types of processing activities.

Additionally, the requirements under Article 35 for 'data protection impact assessments' will result in additional mandated documentation of any data processing activities of the organisation that potentially involve a 'high risk' to the privacy rights of individuals. As a practical matter, organisations are likely to conclude that some form of documented assessment will need to be conducted for new technologies that involve the collection or use of personal data, for data analytics that involve large volumes of personal data or other data processing that is not obviously low risk.

Other GDPR obligations, such as the notice requirements discussed in earlier, will also require extensive documentation.

these requirements obligate organisations to have robust vendor management programmes in place to carefully manage vendors who may touch personal data and ensure such data is protected as it is passed from one entity to another. These provisions apply equally to identified and pseudonymised data; however, they do not apply to anonymised data.

SUPERVISORY AUTHORITY OVERSIGHT OF ANONYMISED DATA

One of the concerns with anonymised data that we have often heard from supervisory authorities is that information that is anonymised falls outside the regulation and falls outside their supervision. Under the risk-based approach to anonymisation that is described in this paper (see Box 1: 'Basic principles of risk-based anonymisation'), an argument can be made that this is not completely the case.

Under a risk-based anonymisation approach, data is claimed to be anonymous as a function of both data perturbations and additional technical and administrative controls that are put in place. The data perturbations are applied once to the data; however, the controls need to be applied continuously to ensure that the data remains anonymous. If the appropriate controls are not in place, lapse or are not strong enough, then the data is no longer anonymous and would fall back within the scope of the regulation and the supervision of data protection authorities.

While the obligations of the GDPR do not apply to the data while it is anonymous, the data controller would need to ensure the controls are in place on an ongoing basis. This can be verified through audits, for example, or self-declarations as appropriate. Overseeing and creating accountability that these controls remain in place, such that the data remains anonymous, is an appropriate role for the supervisory authorities. In that manner, the supervisory authorities will have a continuing role to play even when data is anonymised to ensure that the totality of the risk-based anonymisation criteria are being met.

CONCLUSIONS

The above discussion and the summary in Table 1 make clear that pseudonymised data is far more similar to identified data than it is to anonymised data in terms of the GDPR obligations that apply to it. While pseudonymisation can form part of an overall GDPR compliance strategy in certain cases, and strong pseudonymisation provides greater compliance benefits, it does not result in complete relief from GDPR obligations in the way anonymisation does. Thus, organisations should not confuse the limited advantages of pseudonymisation with the far more sweeping advantages of anonymisation.

ACKNOWLEDGEMENTS

Support for Mike Hintze's contribution to this White Paper has been provided by Privacy Analytics Inc.

References and Notes

1. Logan, V., Moran, M., Richardson, J., Edjlali, R. and Faria, M. (2017) 'Survey Analysis: Third Gartner CDO Survey — How Chief Data Officers Are Driving Business Impact', Gartner, Stamford, CT.
2. Council Regulation 2016/679, 2016 O.J. (L 119) (EU) 1.
3. Another term that is sometimes used to mean the same thing as pseudonymisation is 'tokenisation'.
4. GDPR Article 4(5).
5. GDPR Articles 11(2) and 12(2), which state that if this threshold is met, certain data subject rights with respect to the data will not apply.
6. The level of perturbation ensures that individuals are not unique in the population on these indirect identifiers so that they cannot be 'singled out'.
7. Polonetsky, J., Tene, O. and Finch, K. (2016) 'Shades of gray: Seeing the full spectrum of practical data de-identification,' *Santa Clara Law Review*, Vol. 56, No. 3, pp. 593–629; El Emam, K., Gratton, E., Polonetsky, J. and Arbuckle, L. (2016) 'Seven states of data: When is pseudonymous data not personal information?,' in 'Policy and Practical Solutions for Anonymization and Pseudonymization', Brussels, Belgium, available at: https://fpf.org/brussels-privacy-symposium/.
8. Information Commissioner's Office (2012) 'Anonymisation: Managing Data Protection Risk Code of Practice', Information Commissioner's Office, Wilmslow, Cheshire.
9. European Medicines Agency (2017) 'External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use', available at: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2017/04/WC500225880.pdf (accessed 17th April, 2017).
10. Cavoukian, A. and Castro, D. (2014) 'Big data and innovation, setting the record straight: De-identification does work', Information & Privacy Commissioner of Ontario, available at: https://iapp.org/resources/article/big-data-and-innovation-setting-the-record-straight-de-identification-does-work/ (accessed 28th October, 2015); Cavoukian, A. and El Emam, K. (2011) 'Dispelling the myths surrounding de-identification: Anonymization remains a strong tool for protecting privacy', Information and Privacy Commissioner of Ontario, available at: https://www.ipc.on.ca/wp-content/uploads/2016/11/anonymization.pdf (accessed 2nd November 2018); Cavoukian, A. and El Emam, K. (2010) 'A positive-sum paradigm in action in the health sector', Office of the Information and Privacy Commissioner of Ontario; Cavoukian, A. and El Emam, K. (2014) 'De-identification protocols: Essential for protecting privacy', Office of the Information and Privacy Commissioner of Ontario; Information and Privacy Commissioner of Ontario (2016) 'De-identification guidelines for structured data'.
11. Office for Civil Rights (2012) 'Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule', Department of Health and Human Services, Washington, DC.
12. Institute of Medicine (2015) 'Sharing clinical trial data: Maximizing benefits, minimizing risk', Institute of Medicine, Washington, DC; The Expert Panel on

Timely Access to Health and Social Data for Health Research and Health System Innovation, (2015) ‘Accessing health and health-related data in Canada’, Council of Canadian Academies, Ottawa, Ontario; PhUSE De-Identification Working Group (2015) ‘De-identification standards for CDISC SDTM 3.2’, Broadstairs, Kent; Elliot, M., Mackey, E., O’Hara, K. and Tudor, C. (2016) ‘Anonymisation Decision-Making Framework’, UKAN Publications, Manchester, available at: http://ukanon.net/ukan-resources/ukan-decision-making-framework/ (accessed 2nd November, 2018).

13. El Emam, K. (2014) ‘Pseudonymous data is not anonymous data’, BMJ Blog, 20th November, available at: http://blogs.bmj.com/bmj/2014/11/20/khaled-e-emam-pseudonymous-data-is-not-anonymous-data/ (accessed 2nd November, 2018).
14. El Emam, K. et al. (2011) ‘A systematic review of re-identification attacks on health data’ *PLoS ONE*, Vol. 6, No. 12, e28071, available at: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0028071 (accessed 2nd November, 2018).
15. El Emam, K. (2013) ‘Guide to the De-Identification of Personal Health Information’, CRC Press, Auerbach.
16. GDPR Articles 13, 14 and 15.
17. Id. arts. 13(1)(a), 14(1)(a).
18. Id. arts. 13(1)(b), 14(1)(b). See id. art. 37, for the requirements for designating a data protection officer.
19. Id. arts. 14(1)(d), 15(1)(b).
20. Id. arts. 14(2)(f), 15(1)(g); see also id. Recital 61 (‘Where the origin of the personal data cannot be provided to the data subject because various sources have been used, general information should be provided.’).
21. Id. art. 13(2)(e).
22. Id. arts. 13(1)(c), 14(1)(c), 15(1)(a). Note that ‘processing’ is defined broadly and includes any collection, use or sharing of personal data, see id. art. 4(2).
23. Id. arts. 13(1)(c), 14(1)(c); see also id. art. 6 (listing the legal bases for processing personal data).
24. Id. arts. 13(1)(d), 14(2)(b).
25. Id. arts. 13(2)(c), 14(2)(d).
26. Id. arts. 13(2)(f), 14(2)(g), 15(1)(h).
27. Id. arts. 13(1)(e), 14(1)(e), 15(1)(c).
28. Id. arts. 13(2)(a), 14(2)(a), 15(1)(d).
29. Id. arts. 13(2)(b), 14(2)(c), 15(1)(e); see also id. art. 15 (right of access), art. 16 (right to rectification), art. 17–44 (right to erasure).
30. Id. arts. 13–15. The right to object applies to processing based on Article 6(1)(e) (‘necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller’) or Article 6(1)(f) (‘necessary for the purposes of the legitimate interests pursued by the controller or by a third party’) or for the purposes of marketing. Id. arts. 6(1)(e)–(f), art. 21(1) and (2). The right to obtain a restriction on processing applies under four narrow circumstances described in Article 18(1). See id. An organization may choose to specify these circumstances in its privacy statement in order to avoid implying a broader right to object or restrict processing than is provided by the GDPR.
31. Id. art. 12(7). See also id. art. 20, for the scope of the data portability obligations.
32. Id. arts. 13(2)(d), 14(2)(e), 15(1)(f).
33. Id. arts. 13(1)(f), 14(1)(f); see also id. art. 15(2).
34. GDPR Article 6(1).
35. See Article 6(1)(a). Article 4(11) of the GDPR defines consent as being ‘freely given, specific, informed and unambiguous.’ Compared to the definition in the 1995 Data Protection Directive, the GDPR definition adds the requirements that consent be ‘unambiguous’, which could be interpreted as raising the bar on what may constitute valid consent.
36. El Emam, K., Jonker, E., Moher, E. and Arbuckle, L. (2013) ‘A review of evidence on consent bias in research’, *American Journal of Bioethics*, Vol. 13, No. 4, pp. 42–44.
37. See Article 6(4). See also Article 5(1)(b).
38. GDPR Article 6(4)(e).
39. GDPR Article 6(1)(f).
40. See Opinion 05/2014 on anonymisation techniques, at 7 (‘the Working Party considers that anonymisation as an instance of further processing of personal data and can be considered to be compatible with the original purposes of the processing but only on condition the anonymisation process is such as to reliably produce anonymised information in the sense described in this paper’).
41. See GDPR Articles 15(3) and (4) (right of access), 16 (right of rectification), 18 (right to restriction of processing), 20 (right to data portability) and 21 (right to object to processing).
42. See GDPR Articles 44–49.
43. GDPR Article 25(1).
44. GDPR Article 32(1).
45. See, for example, Article 32(1)(a), which lists pseudonymisation as one of the ‘appropriate technical organization measures’ and Article 6(4) which refers to ‘appropriate safeguards, which may include encryption or pseudonymisation.’
46. GDPR Article 33(1).
47. GDPR Article 34(1).
48. See, Congressional Research Service (2012) ‘Data security notification laws’ at 6, available at: https://fas.org/sgp/crs/misc/R42475.pdf (accessed 2nd November, 2018).
49. See, for example, Article 30.
50. There is a narrow exception under Article 30(5) that applies to small organisations (under 250 employees) that do not process sensitive personal data and only occasionally process personal data at all. In such a case, the organisation must to a risk assessment to determine whether the Article 30 documentation requirements apply, and the use of pseudonymisation could play a factor in such assessment.
51. See GDPR Article 28(3).
52. See, for example, Article 29 limiting the scope of data processors’ activities to that which is directed by the data controller or as required by law, and the recordkeeping obligations of processors under Article 30(2).
53. See GDPR Article 28(2) & (4).