Response to public consultation on EDPB Guidelines 01/2025 on Pseudonymisation

General assessment of the Guidelines

I felt the guidelines were very much written by lawyers for lawyers rather than for governance or IT practitioners. Rarely was a 'guideline' (some were simply explanatory paragraphs) written to express something that could or should be done, followed by an explanation of why. Often sentences started with a long conditional clause, so that the reader might easily lose track of what was actually being recommended (especially as so many paragraphs were explanatory).

Difficulties of the GDPR definition

The GDPR definition of 'pseudonymisation' conflates two different concepts: 1) the creation of a 'pseudonym' (usually unique for each individual concerned, though may only be unique for each record-set/case), and 2) general techniques to obscure a person's identity (which I would call 'de-identification' without any indication that the resultant data is wholly anonymous and outside the requirements of GDPR and to be freely used as non-personal data, only that the risk of re-identification has been reduced not necessarily entirely eliminated).

This means that data may be obscured but without creation of a pseudonym – but still qualifies as 'pseudonymisation' as defined in GDPR Article 4. This is recognised in Guideline 8: '*it [pseudonymisation] does not even explicitly require the replacement of direct identifiers by pseudonyms*'.

Equally, a pseudonym may be created but without any reduction in identifiability at all. Indeed, as Example 4 shows it may make a system more robust rather than simply relying on name/address combinations – except most readers would recognise this as the creation of an 'identifier' (national, regional, or clinic-specific) rather than a 'pseudonym' as the purpose is to assure proper identification of related records.

Guidelines 6-9 actually recognise this distinction, though perhaps not as starkly as might be helpful. Much of the technical discussion presumes a use of pseudonyms but without being clear on this point – mainly as there is no clear distinction made between 1) using pseudonyms, and 2) using other de-identification techniques throughout the document.

It would be helpful generally if the EDPB guidance could elaborate on the different aspects of 'pseudonymisation', perhaps by providing suitable terms, so that practitioners can better understand the issues and the guidance being given.

Status of pseudonymised data

GDPR Recital 26 includes:

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.

The Executive Summary paraphrases this in its penultimate paragraph as:

Controllers must always bear in mind that pseudonymised data, which could be attributed to a natural person by the use of additional information, remains information related to an identifiable natural person, and thus is personal data[1] *(Rec. 26 GDPR)*.

Guideline 76 notes: '*Since pseudonymised data, which could be attributed to a natural person by the use of additional information, is personal data …*'.

The use of commas around the phrase *which could be attributed to a natural person by the use of additional information* makes it a parenthetic statement[2] rather than a conditional clause (which it would be without the commas). So effectively it asserts '… pseudonymised data **is** personal data …' This appears to omit the possibility, which Recital 26 recognises, that pseudonymised data may have been sufficiently de-identified that it is, in fact, 'anonymous'. It effectively asserts that there can be no mechanism by which 'personal data' can ever be rendered non-personal, which is clearly false (see counter-example below).

Ideally, this should be re-phrased so that it is clear what is meant without relying on the presence or absence of commas. Perhaps: any of:

- *Controllers must always bear in mind that, if pseudonymised data could be attributed to a natural person by the use of additional information, then it remains information related to an identifiable natural person, and thus is personal data*.
- *Where pseudonymised data could be attributed to a natural person by the use of additional information, then it would be personal data* under GDPR …'

which are unambiguous.

It is perfectly reasonable to assert that the data is 'likely to be' personal data and that any processing of the data should comply with the GDPR, unless a DPIA (or similar risk

[1] Recital 26 is a little more nuanced and uses the form 'should be considered to be' rather than the absolute 'is' used here. Admittedly, its punctuation is wrong as well having only one comma rather than none or two.

[2] So would be read as 'Since pseudonymised data is personal data (because it could be attributed to a natural person by the use of additional information) …' rather than 'Where pseudonymised data could be attributed to a natural person by the use of additional information, then it would be personal data under GDPR'.

assessment) has established that the intended processing poses no risks to the rights and freedoms of the individuals concerned.

It would also be reasonable to state that it is very unlikely that pseudonymised data could be sufficiently de-identified so as to be made publicly available and so robust against all possible attacks or future advances in technology. Further, it would be reasonable to suggest that the same levels of security should be applied to 'pseudonymised data' as to 'personal data', apart from recognising that it may need different levels of user authorisation to permit access or requests for processing.

The encouragement to support 'transparency' concerning the processing of pseudonymised data, regardless of its status as 'personal data', is to be welcomed.

Counter-example 1 – how pseudonymised data can be anonymous

A small regional health authority holds details of its local population and financial details of claims from care providers, so is able to create a file noting which patients have been treated for a particular condition.

Let us presume that the condition is fairly common, say, 10% of the total population and is not itself directly sex-related (e.g. prostate cancer). It seeks to extract just the sex (excluding and inter-sex or trans cases) and whether the person has been treated for the condition or not, so just M/F and Y/N. A pretty minimal dataset, but sufficient to check whether the condition is more common in males than females.

However, conscious that the sequencing of records might hold some significance, it first assigns a random number to each record (a pseudonym if you like, but not recorded elsewhere than in the new file), then it sequences the file by this value and then creates a new file of just the [M/F] and [Y/N] fields, so there can be no meaning assigned to the sequence of records. This is important in case the default sequence of records had been, say, the postcode so the sequence of records might allow some inference of where the patient associated with the instance actually lived and hence to their identity – quite how this inference might have been done in practice is not at issue here. We presume any intermediate files are thoroughly deleted, so there is no way of back-tracking to the original dataset.

The resultant file, of tens of thousands of records, where each record is effectively one of four possibilities is clearly anonymous even if held by the original controller.

It is fair to say that this dataset is of only marginal clinical usefulness and it would have been easier for the health authority to just generate the aggregate statistics required for the analysis.

However, it does disprove the assertion that 'pseudonymised data is personal data', and always so if held by the original controller.

It is reasonable to say that, in general, 'pseudonymised data is very likely to be personal data', or even that 'pseudonymised data should be treated as though it were personal data, until assessed otherwise'. These would both be sensible precautionary guidelines.

Note that Guideline 22 provides a more measured statement: *If pseudonymised data and additional information could be combined having regard to the means reasonably likely to be used by the controller or by another person, then the pseudonymised data is personal*. This concurs with Recital 26.

Guideline 22 ends with ' ... *the pseudonymised data becomes anonymous only if the conditions for anonymity are met*' but nowhere in the document is it made clear what these may be, so leaves practitioners struggling to work out what criteria supervisory authorities might use against them in any judgement.

This is an important issue – in part because earlier guidance (usually in executive summaries) has stated unambiguously that 'pseudonymised data is personal data' with the result that many governance experts are still quoting this as established legal fact (despite it being easily proven wrong as the counter-example above shows).

'Additional information' and 'Pseudonymisation secrets'

The phrase 'additional information' appears repeatedly in the GDPR, but in two distinct senses: 1) 'any additional information' that might permit the attribution of data to a person (Recital 26) for the purposes of determining whether data is 'personal data' or not, and 2) the 'additional information for attributing the personal data to a specific data subject' (Recital 29) referring to the 'pseudonymisation secrets' which may be generated by the pseudonymisation process (as made clear later but not until Section 3.1.1).

Without this clarity, Section 2.1 reads rather oddly, e.g. Guideline 19: '*The generation, or use of additional information is an inherent part of the pseudonymising transformation*'. This is confusing at first reading as one naturally assumes one is transforming the data to reduce its information content – once it is realised that the specific additional information being referred to relates to the pseudonym key or table (depending on technique being used) then it becomes clearer.

Guideline 76 poses a particular problem here: '*Since pseudonymised data, which could be attributed to a natural person by the use of additional information, is personal data, the rights of the data subject according to Chapter 3 GDPR apply*'. As noted previously, the phrasing is unfortunately ambiguous. Certainly, if the 'additional information' is the 'pseudonymisation secrets', so that there should be a clear reversal mechanism available, then this might be reasonable.

Guideline 77 recognises that Article 11 may apply, though Guideline 76 on its own does not admit that possibility, nor does it note the exception that Guideline 77 allows (admittedly in highly convoluted language). However, Guideline 77 also notes: '... *and is demonstrably unable to reverse the pseudonymisation with the assistance of another controller*' which seems extraordinary to suggest that one controller should be obliged

to help fulfil another controller's legal obligations. I wonder if the wording should have been '**without** the assistance of another controller' making it clear that on its own (or with its processors) it could not reasonably fulfil the request.

Use of pseudonyms to improve 'accuracy'

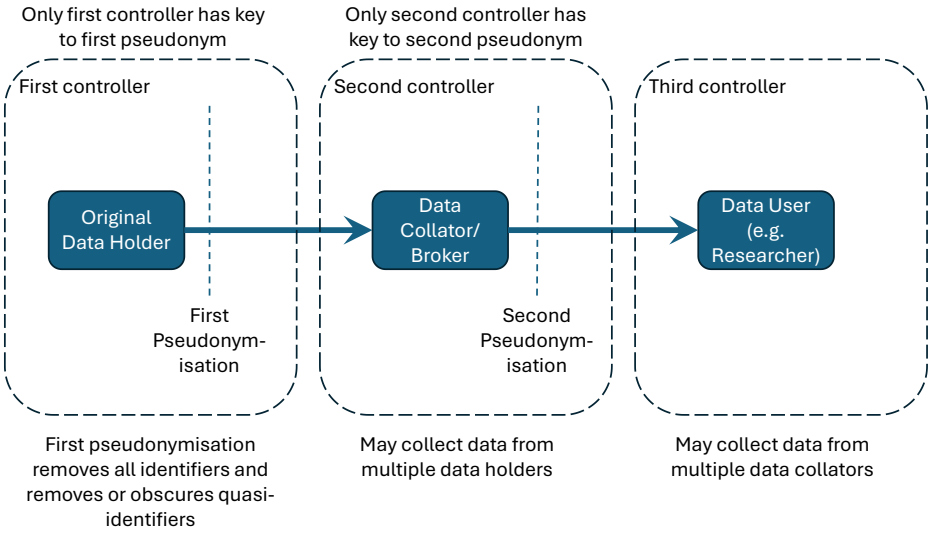
Guideline 58 suggests on first reading that pseudonymisation may improve accuracy, which cannot be the case in general. Admittedly, the clause does state 'an appropriate pseudonymisation procedure' which makes it then correct when read by a lawyer, but not when read by a lay or simply technical reader. Moreover, this so glosses over the technical and administrative work involved in creating such a reliable process as to be misleading – and hence not a good 'guideline'.

I believe this is an attempt to note that having an effective, unique, and secure ID mechanism will help ensure that records are reliably connected to an individual, but this is about effective identity management, not the use of 'pseudonyms', which are about obscuring the identity of an individual.

I think that this would be better addressed in a pre-amble to Example 4 rather than as explicit 'guideline' which may confuse rather than aid a practitioner.

Use of 'double pseudonymisation'

Some practitioners use and would recommend a practice of 'double pseudonymisation' by different controllers in a data pipeline, for example:



No single controller can re-identify an individual case (assuming adequate de-identification at source); further, the data collection step will reduce attribution risk by increasing range of records and cases – it may also apply additional obscuration techniques (e.g. limiting outliers, providing only a random sample, reducing geographic and other attributes).

While this is clearly 'good practice' and it is not possible in general to say that this will allow data to be anonymised, guidelines 76 and 77 together (as written) suggest that the final data user must hold 'personal data', despite the recent opinion by the Attorney-General on SRB vs EDPS supporting the General Court decision.

Data linkage issues

Section 3.3 is useful, particularly the distinction between the three types of pseudonym – though these do not appear in the Glossary. Real-world examples might help readers envisage why they might wish to use one rather than another – just a simple text-box rather than a fully worked up example as in the Annex.

Section 3.3.2 would be more helpful with some diagrams (c.f. the Examples) to help compare and contrast the three different approaches suggested.

Use of bullet-points (as per #64) would help readers identify more readily the three different types of pseudonym and the three different approaches suggested in 3.3.2.

Risk assessment methods

The Executive Summary notes:

Controllers should establish and precisely define the risks they intend to address with pseudonymisation. The intended reduction of those risks constitutes the objective of pseudonymisation within the concrete processing activity. Controllers should shape pseudonymisation in a way that guarantees that it is effective in reaching this objective.

This again rather overstates what is possible in practice – especially given that these guidelines are themselves remarkably weak in 'precisely defining the risks' or the methodology or even criteria by which risks should be assessed and found sufficient.

It is unlikely that any process can 'guarantee' that it will achieve its objectives in all circumstances, particularly when we are addressing risk minimisation rather than risk elimination (which might just be achievable).

Perhaps all that is meant is 'clearly define' as in 'be specific' and that 'Controllers should shape pseudonymisation so that it is effective in reaching this objective' (no 'guarantees' involved) – part of 'privacy by design'!

Example 10 – respecting data subject rights

This example is not as well developed as the previous ones (excepting Example 4 as weak as noted above) which is surprising given the complexity of the point being made.

An example occurs in clinical trials whereby a pharmaceutical company as sponsor is generally considered to be the controller of the data collected as part of the clinical trial

(though not the data routinely collected in any electronic health record (EHR) maintained by the hospital or clinic).

However, the pharmaceutical company is at great pains to avoid knowing the actual identity of the individuals, needing only anonymised data (using the term loosely here) for analysis and presentation to regulators as part of the regulation and approval of medicinal products.

As part of this 'anonymisation', it may be possible to share the study data with other researchers in order to validate the study results – part of the 'Open Data' initiatives. Clearly, not 'public domain' data-sharing but only with recognised research institutes.

However, the thrust of the example, as described, implies that the sponsor must seek to re-identify the individual by seeking more information from them rather than the more straightforward position of recommending that the data subject contact the relevant institution(s) running the study, who can provide a complete response about data held and also the information that would have been passed to the sponsor as a result of the study.

This approach meets both the individual's needs and rights as well as avoiding the 'contamination' of the 'anonymised' extract by re-identification of certain records – leaving the sponsor in the quandary of whether they now need to remove these records from the study (even though the individual is happy to remain part of the study) so that they continue to only share 'anonymised' data for research purposes (compatible with any consent or Article 89(1) requirements) where appropriate – though now no longer representative of the original study because records have necessarily been removed.

It is a long-established principle in medical research that only the treating clinicians know the identity of the individual – the 'scientific' side of research only knows about 'cases' with the relevant details to preform research to the benefit of all. This is to avoid potential bias in the interpretation of results.

I think this is an area where Article 15(4) comes into play: '*The right to obtain a copy referred to in paragraph 3 shall not adversely affect the rights and freedoms of others*'. The privacy interests of other participants is at risk, as well as their altruism in taking part in medical studies as the study may be invalidated or compromised if the 'anonymised' dataset has to have instances removed. There is also the wider public interest in supporting medical research and maintaining medical confidentiality.

The example given does not offer the obvious approach and will be interpreted further as reason not to pseudonymise data as it will need to be re-identified anyway if any data subject seeks to assert their data subject rights.

It may also be worth noting that if controllers seek to direct data subjects to request data from their processor, then they need to commission a 'data subject rights service

fulfilment' from that processor as well, so that the data subject's rights are protected as far as possible.

Terminology & Glossary

Finally, I note that the Glossary makes use of an arrow symbol ('→') which is wholly unexplained, but is apparently an indicator that the following word or phrase is also defined in the Glossary, though the actual extent of the phrase is not clearly indicated. So does '→ pseudonymised data holding' refer to an entry 'pseudonymised', 'pseudonymised data', or 'pseudonymised data holding'? The reader has to deduce what the symbol means and what entry it may refer to. At least at other points in the document, new phrases are emboldened or italicised to make clear what the actual phrase is.

The normal convention (in my experience) is to highlight the phrase, probably using italics or apostrophes, followed by '(q.v.)' so readers know to look it up.

Perhaps an explanation of the intended practice for this convention could appear at the beginning of the Glossary section. I think it would also be helpful to have the glossary after the Executive Summary rather than at the end – it would save some of the more convoluted phrasing in the earlier section before the new phrases are defined in Section 3.

See also #105 where terms are introduced but not reflected in the Glossary.

Peter Singleton

Director

Cambridge Health Informatics Limited

18 Wordsworth Grove

Cambridge

CB3 9KK

UK