DataTools4Heart

A European Health Data Toolbox for Enhancing Cardiology Data Interoperability, Reusability and Privacy

Clear policy and guidance on pseudonymised and synthetic data processing in the clinical sector

Reference	**MS2_ DataTools4Heart_PAN_29092023**
Lead Beneficiary	PANETTA
Author(s)	Lorenzo Cristofaro
Dissemination level	Public
Type	Report
Official Delivery Date	29/09/2023
Date of validation of the WP leader	29/09/2023
Date of validation by the Project Coordinator	29/09/2023
Project Coordinator Signature	

DataTools4Heart is funded by the European Union's Horizon Europe Framework Under Grant Agreement No. 101057849.

Version Log

Issue Date	Version	Involved	Comments
29/09/2023	1	Lorenzo Cristofaro	Revised and corrected final version

Executive Summary

Notwithstanding the pivotal importance of the distinction between personal and non-personal data according to the applicable laws, it is extremely burdensome, in most of the cases, to differentiate between these categories. This is because there is no precise indication in the legislation, or by competent regulators, as to what is the correct legal test to apply to correctly categorise data as anonymous or not. Practically, the EU regulatory scenario has so far been affected by a very rigid interpretation of data anonymisation, deriving from an important opinion adopted by the Article 29 Working Party in 2014. Moreover, the different stances taken by national Supervisory Authority have made reliance on anonymisation techniques even more complex to achieve and riskier, since the degree of irreversibility that individual de-identification must achieve so that data can be deemed anonymous varies from a member State to another. In parallel, the requirements to be abided in connection with the processing and even more the reuse of health data for scientific research are not homogeneous at European level, in that every member State is empowered to adopt its own limitations or conditions which apply in addition to or in lieu of the General Data Protection Legislation. DataTools4Heart aims to set, among others, new regulatory standards for ensuring that all types of health data, both in structured and unstructured format, can securely and lawfully undergo a secondary processing for the purpose of medical research in the cardiology sector. For this reason, this report starts with a detailed analysis of the current state-of-the-art regarding pseudonymisation, for then going in-depth into the benefits that can stem from some specific Privacy-Enhancing Technologies, focusing on Federated Learning, Differential Privacy, Secure Multi-Party Computation and, particularly, on Synthetic Data, with a view to overcoming the hurdles that to date prevent the implementation of the European Health Data Space and the progress of the EU Research area. The analysis goes focusing on the legal nature of synthetic data in the light of the Artificial Intelligence Act, and evaluating in detail the promising interpretative evolutions of what constitutes pseudonymous and anonymous data based on the crucial decision issued by the EU General Court in April 2023 in relation to the Case T-557/20. Finally, conclusions are drawn regarding the robustness and the reliability of the innovative solutions put forward in the project, to enhance the protection of personal data and patients' privacy, while achieving strong accountability and enabling a concrete progress of medical research thanks to compliant reuse of health data.

Table of Contents

List of tables

List of figures

1 The concept of ‘pseudonymisation’

Even though pseudonymisation represents a crucial tool to ensure privacy-by-design and reinforce accountability, the complexity of its practical interpretation, with particular regard to its distinction from anonymisation, made the application of pseudonymisation quite tricky – especially in some contexts – and therefore much less frequent than expected.

While no exact definition of ‘anonymisation’ or ‘anonymous data’ is provided for by Regulation (UE) of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, hereinafter ‘**GDPR**’ or ‘**Regulation**’), ‘pseudonymisation’ is described 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*” (Art. 4(5) GDPR).

In brief, pseudonymisation aims at protecting personal data by hiding the identities of individuals in a dataset, e.g. by replacing one or more personal identifiers with the so-called pseudonyms, and appropriately protecting the link between such pseudonyms and the initial identifiers.[1]

At a very basic level, pseudonymisation starts with a single input (the original data) and ends with two outputs (the pseudonymised dataset and the additional information). Together, these can reconstruct the original data. However, with respect to the individuals concerned (the data subject), each output has meaning only in combination with the other.

Therefore, pseudonymisation refers to techniques that replace, remove or transform information that identifies individuals, and keep that information separate and secure.

[1] The report on ‘Pseudonymisation techniques and best practices’ by the European Union Agency for Cybersecurity (ENISA), dated November 2019 (link), provides, *inter alia*, the following definitions: (a) “*Identifier is a value that identifies an element within an identification scheme7. A unique identifier is associated to only one element. It is often assumed in this report that unique identifiers are used, which are associated to personal data*”; (b) “*Pseudonym, also known as cryptonym or just nym, is a piece of information associated to an identifier of an individual or any other kind of personal data (e.g. location data). Pseudonyms may have different degrees of linkability (to the original identifiers). The degree of linkability of different pseudonym types is important to consider for evaluating the strength of pseudonyms but also for the design of pseudonymous systems, where a certain degree of linkability may be desired (e.g. when analysing pseudonymous log files or for reputation systems)*”; (c) “*Pseudonymisation entity is the entity responsible of processing identifiers into pseudonyms using the pseudonymisation function. It can be a data controller, a data processor (performing pseudonymisation on behalf of a controller), a trusted third party or a data subject, depending on the pseudonymisation scenario. It should be stressed that, following this definition, the role of the pseudonymisation entity is strictly relevant to the practical implementation of pseudonymisation under a specific scenario*”.
More technically, i) “*pseudonymisation function, denoted $\mathcal{P}$, is a function that substitutes an identifier Id by a pseudonym $pseudo$*”; ii) “*Pseudonymisation secret, denoted s _is an (optional) parameter of a pseudonymisation function $\mathcal{P}$. The function P _cannot be evaluated/computed if s is unknown*”; *iii)* “*Recovery function, denoted $\mathcal{R}$, is a function that substitutes a pseudonym $pseudo$ by the identifier Id using the pseudonymisation secret s. It inverts the pseudonymisation function $\mathcal{P}$*”; iv) “*Pseudonymisation mapping table is a representation of the action of the pseudonymisation function. It associates each identifier to its corresponding pseudonym. Depending on the pseudonymisation function $\mathcal{P}$, the pseudonymisation mapping table may be the pseudonymisation secret or part of it*”.

It has a prominent role in the GDPR both as a security measure (Art. 32) and as a tool to achieve privacy-by-design (Art. 25) and data minimisation (Art. 5.1(c))[2]: this technique can go beyond hiding real identities into supporting the data protection goal of unlinkability, i.e., reducing the risk that privacy-relevant data can be linked across different data processing domains, as well as contribute towards the key principle of data minimisation under the GDPR, for example in cases where the controller does not need to access personal data relating to the data subjects, but only to their pseudonyms.

For this reason, "*pseudonymisation can motivate the relaxation, to a certain degree, of data controllers' legal obligations if properly applied*".[3]

Without any prejudice to the aforementioned benefits, a key concept to keep in mind is that "*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*" (Recital 26 GDPR). In other words, pseudonymous data are still personal data, falling into the scope of application of the GDPR and any other applicable data protection legislation.

Practically, when implementing pseudonymisation, it is important to clarify the application scenario and the different actors involved and their roles, with particular respect to that of the pseudonymisation entity,[4] which can be attributed to different entities (e.g. a data controller, a data processor, a Trusted Third Party or the data subject) depending on the case. Under each specific scenario, it is then required to consider the best possible pseudonymisation technique and policy that can be applied, given the benefits and pitfalls entailed.

Obviously, there is not a one-size-fits-all approach and a case-by-case data protection risk analysis remains crucial, to consider all relevant aspects and variables (e.g. privacy protection, utility, scalability, etc).

1.1 Pseudonymisation scenarios in practice

The defining difference between the various scenarios is firstly the actor who takes the role of pseudonymisation entity and secondly the other potential actors that may be involved (and their roles).

The figures below – taken from the 'Pseudonymisation Advanced Techniques and Use Cases' report published by the European Union Agency for Cybersecurity ('**ENISA**') in January 2021[5] – outline six

[2] Not by chance, Art. 89 of GDPR ('Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes') specifically mentions pseudonymisation as a suitable technique to implement the minimisation of the processing operations ("*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*").

[3] ENISA's report on 'Pseudonymisation techniques and best practices' referred to in Note no. 1.

[4] See the definition in Note no. 1.

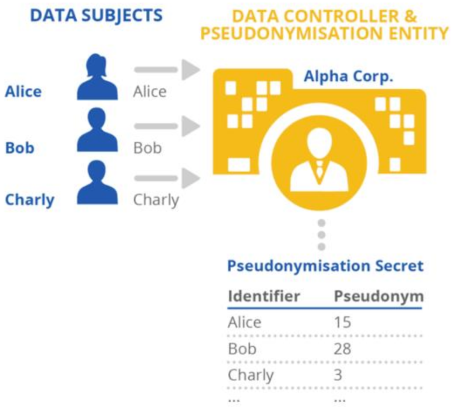
[5] This report can be read here.

different pseudonymisation scenarios that can be found in practice, listing the various actors and the specific goals of each case.

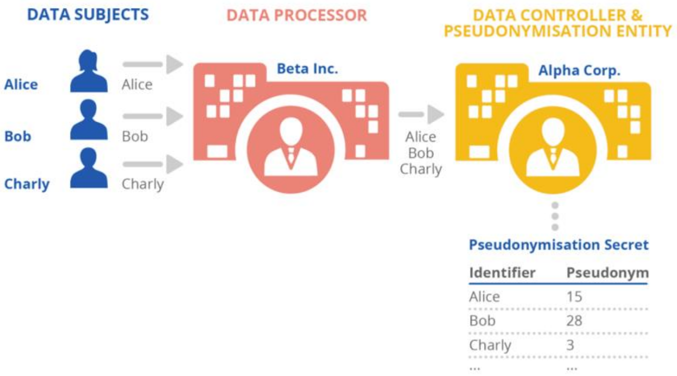
In all three first scenarios above, the data controller is the pseudonymisation entity (‘**PE**’), either acting alone (scenario 1) or involving a processor before (scenario 2) or after pseudonymisation (scenario 3). in scenario 4, the PE is the processor that applies pseudonymisation on behalf of the controller, while scenario 5 assigns the role of PE to a Trusted Third Party, outside the control of the data controller, therefore involving an intermediary to safeguard the pseudonymisation process. Lastly, scenario 6 provides for data subjects to be the PE and, thus, control an important part of the pseudonymisation process.

In more details:

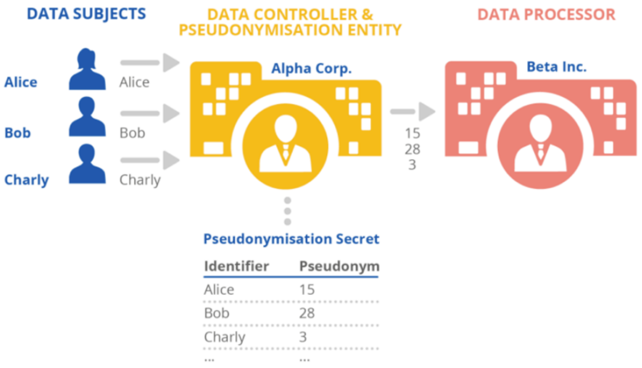
1) Scenario 1: the data controller (Alpha Corp.) has the role of pseudonymisation entity, as it performs the selection and assignment of pseudonyms to identifiers. It must be pointed out that the data subjects do not necessarily know nor learn their particular pseudonym, as the pseudonymisation secret (e.g. the pseudonymisation mapping table in this example), is known only to Alpha Corp. The role of pseudonymisation in this case is to enhance the security of personal data either for internal use (e.g. sharing between different departments within the controller’s organisation) or in the case of a security incident;



2) Scenario 2: this is a variation of scenario 1, where a dedicated data processor (Beta Inc.) is given the task to collect the identifiers from the data subjects and forward this information to a subsequent data controller (Alpha Corp.), which finally performs the pseudonymisation. The controller is then, again, the pseudonymisation entity. An example for such a scenario might be a cloud service provider that hosts data collection services on behalf of the data controller. Then, the controller still is in charge of applying data pseudonymisation prior to any subsequent processing activities;



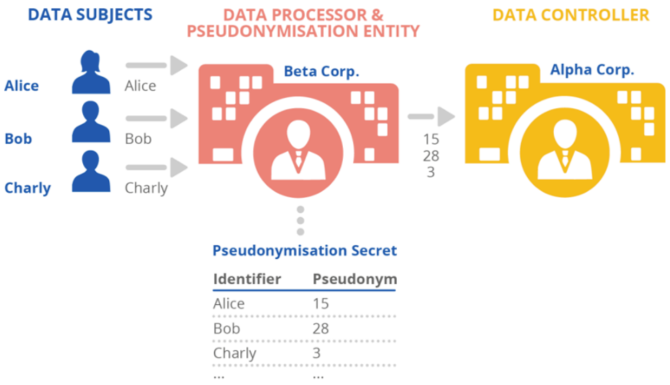
3) Scenario 3: contrary to the previous cases, here the data controller (Alpha Corp.) still directly collects the personal data and the applies pseudonymisation (in its role as PE). The data processor (Beta Inc.) only receives the already pseudonymised dataset, e.g. for statistical analysis, or persistent data storage. In this scenario, as Beta Inc. cannot learn the identifiers of the data subjects, it is not directly able to re-identify them (assuming that no other attribute that could lead to re-identification is available to Beta Inc.).[6] In this way, pseudonymisation protects the security of the data with regard to the processor (a variation of this scenario could be the case where the pseudonymised data is not sent to a data processor but to another data controller);



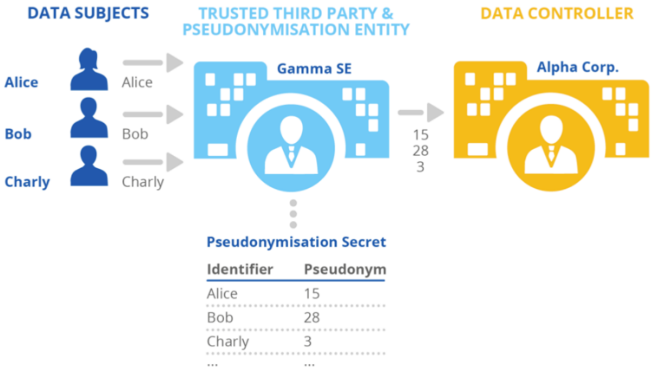
4) Scenario 4: in this case, personal data are provided by the data subjects to a data processor (Beta Inc), which subsequently performs the pseudonymisation, thus acting as the PE on behalf of the controller (Alpha Corp). The pseudonymised dataset is then forwarded to the latter. In this particular scenario, only the pseudonymised data are stored on the controller's side, so that the security level is enhanced through data de-identification (e.g. in case of data breach). Still, since

[6] As it will be better detailed below, the qualification of the data as pseudonymous or anonymous under this scenario is the subject of a heated debate in the EU.

the controller remains in condition to re-identify the data subjects through the data processor, security at processor's side becomes of crucial importance;[7]



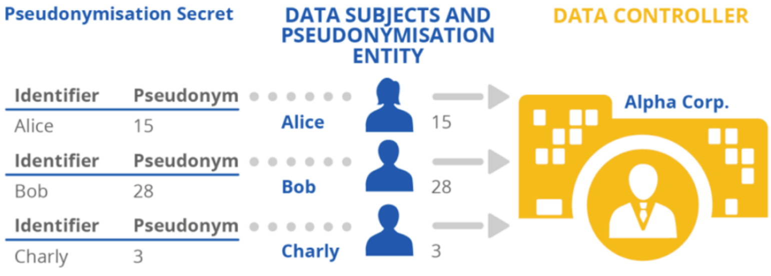
5) Scenario 5: this is the hypothesis when the pseudonymisation is entrusted to and applied by a third party (not a processor), that subsequently forwards the data to the controller. Contrary to the Scenario 4, the controller (Alpha Corp) in this scenario does not have access to the data subjects' identifiers, as the third party (Gamma SE) – acting in quality as PE – does not operate on behalf and upon instructions of the data controller. As a consequence, the latter cannot directly or indirectly link individual data records to data subjects itself, in such a way that security and data protection at controller's level are enhanced in accordance with the principle of data minimisation;



6) Scenario 6: this is a special case where the pseudonyms are created directly by the data subjects themselves, as part of the overall pseudonymisation process. An example could be the use of the public key of a key pair in blockchain systems, to produce the pseudonym. The goal in this case is that the controller does not learn the identifiers of the data subjects, it being understood that the responsibility of the overall pseudonymisation scheme still rests with the data controller.

[7] A variation of this scenario could be a case where several different processors are involved in the pseudonymisation, as a sequence of pseudonymisation entities (*i.e.,* a chain of processors).

Again, this is in line with the principle of data minimisation and can be applied in cases where the controller does not need to have access to the original identifiers (i.e., the pseudonyms are sufficient for the specific data processing operation).



1.2 Traditional and advanced pseudonymisation techniques

1.2.1 Basic techniques

Subject to the pseudonimisation policies above, based on the more recent studies carried out by the ENISA[8], the most frequently applied basic techniques are as follows:

a. **Counter:** the simplest pseudonymisation function, where the identifiers are substituted by a number chosen by a monotonic counter (e.g. 110, 111, 112, 113 and so on). Its advantages rest with its simplicity, which make it a good candidate for small and not complex datasets. It provides for pseudonyms with no connection to the initial identifiers (although the sequential character of the counter can still provide information on the order of the data within a dataset). However, the solution may have implementation and scalability issues in cases of large and more sophisticated datasets;

b. **Random Number Generator (RNG):** a similar approach to the counter, with the difference that a random number is assigned to the identifier (e.g. 110; 319; 818; 196 and so on). It provides stronger data protection (as, contrary to the counter, a random number is used to create each pseudonym, thus it is difficult to extract information regarding the initial identifier, unless the mapping table is compromised). However, collisions (namely the case of two identifiers being associated to the same pseudonym) may be an issue[9], as well as scalability, depending on the implementation scenario, especially in cases of large datasets;

c. **Cryptographic hash function:** directly applied to an identifier to obtain the corresponding pseudonym with the properties of being a) one-way, meaning that it is computationally

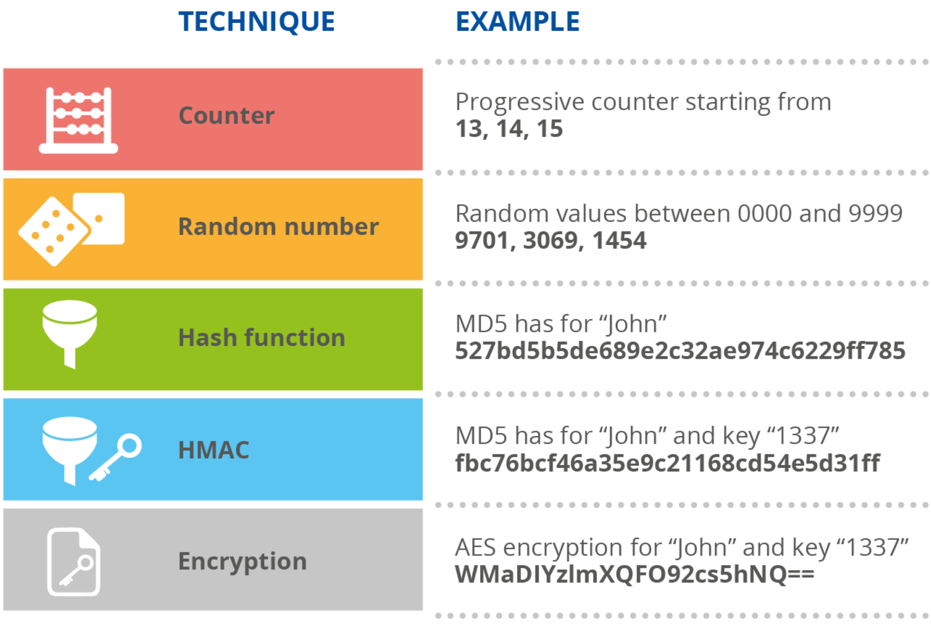
[8] Reference is made, particularly, to the following reports by the ENISA (i) 'Pseudonymisation techniques and best practices', dated November 2019; (ii) 'Pseudonymisation Advanced Techniques and Use Cases', dated January 2021; (iii) 'Deploying pseudonymisation techniques – The case of the health sector', dated March 2022 (link).

[9] The risk of collisions can be made negligible if large pseudo numbers are generated (e.g. of 100-digit length).

infeasible to find any input that maps to any pre-specified output, and b) collision free, as it is computationally infeasible to find any two distinct inputs that map to the same output. While a hash function can significantly contribute towards data integrity, it is generally considered weak as a pseudonymisation technique, as it is prone to brute force and dictionary attacks;

d. **Hash-based Message authentication Code (HMAC)**: similar to a cryptographic hash function, except that a secret key is introduced to generate the pseudonym. Without the knowledge of this key, it is not possible to map the identifiers and the pseudonyms. MAC is generally considered as a robust pseudonymisation technique from a data protection point of view. Recovery might be an issue in some cases (i.e., if the original identifiers are not being stored). Different variations of the method may apply with different utility and scalability requirements;

e. **Symmetric encryption:** two-way (and so reversible) cryptographic function transforming an input personal data in values that can be re-transformed in its original format using a key. The block cipher (such as AES) is used to encrypt an identifier using a secret key, which is both the pseudonymisation secret and the recovery secret. Using block ciphers for pseudonymisation requires to deal with the block size. Symmetric encryption is a robust pseudonymisation technique, with several properties being similar to HMAC (i.e., the aforementioned properties of the secret key). One possible issue in terms of data minimisation is that the PE can always reverse the pseudonyms, even if there is no need to store the initial individuals' identifiers.



TECHNIQUE	EXAMPLE
Counter	Progressive counter starting from **13, 14, 15**
Random number	Random values between 0000 and 9999 **9701, 3069, 1454**
Hash function	MD5 has for “John” **527bd5b5de689e2c32ae974c6229ff785**
HMAC	MD5 has for “John” and key “1337” **fbc76bcf46a35e9c21168cd54e5d31ff**
Encryption	AES encryption for “John” and key “1337” **WMaDIYzlmXQFO92cs5hNQ==**

Figure 1 – Application of basic pseudonymisation techniques

Regardless of the choice as to which specific pseudonymisation technique is applied, the policy (or mode) of implementation is also critical. Three main different pseudonymisation policies can be listed (considering an identifier Id which appears several times in two datasets A and B):

- ✓ **Deterministic pseudonymisation**: in all the databases and each time it appears, Id is always replaced by the same pseudonym $pseudo$;
- ✓ **Document randomised pseudonymisation:** each time Id appears in a database, it is substituted with a different pseudonym ($pseudo1$, $pseudo2$, etc.); however, Id is always mapped to the same collection of ($pseudo1$, $pseudo2$) in the datasets A and B;
- ✓ **Fully randomised pseudonymisation:** for any occurrences of Id within a database A or B, Id is replaced by a different pseudonym ($pseudo1$, $pseudo2$).

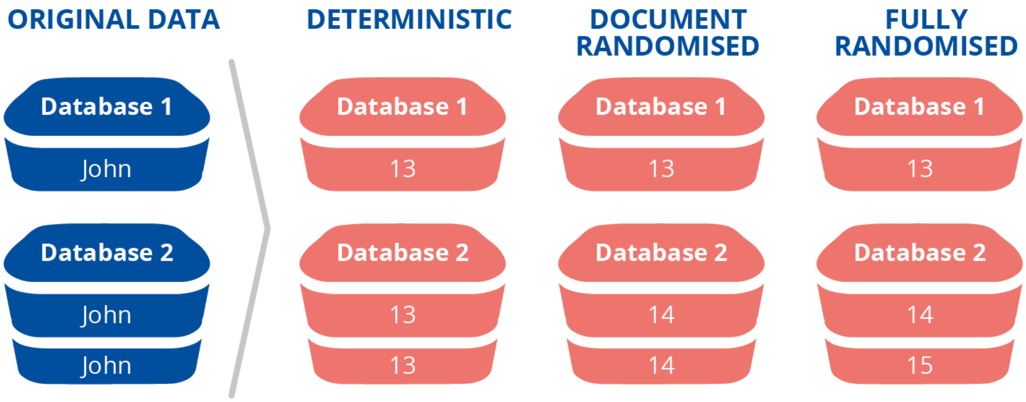


Figure 2 – Application of pseudonymisation policies

1.2.2 Advanced techniques

The basic techniques described above, alongside with relevant policies and scenarios, can improve the level of protection of personal data, provided that the pseudonymisation secrets used to create the pseudonyms are not exposed. However, in order to address some specific data protection challenges – even more in the healthcare domain – such typical solutions may not always suffice.

In this case, it is possible to address more complex situations, ensuring that the level of security is enhanced and that the risks of a personal data breach are properly minimised, thanks to more complex pseudonymisation techniques, such as asymmetric and homomorphic encryption; ring signatures and group pseudonyms; chaining mode, pseudonyms based on multiple identifiers or attributes; pseudonyms with proof of ownership; secret sharing schemes.

Some advanced pseudonymisation solutions, based on cryptographic techniques, also qualify as Privacy-Enhancing Technologies (‘**PETs**’) which will be further discussed below, aiming to enforce "*privacy principles in order to protect and enhance the privacy of users of information technology (IT) and/or of individuals about whom personal data are processed*".[10]

Among the many cutting-edge pseudonymisation applications, it is worth focusing:

a. Asymmetric encryption

This option enables the possibility to have two different entities involved during the pseudonymisation process: (i) a first entity can create the pseudonyms from the identifiers using the Public pseudonymisation Key (PK), and (ii) another entity is able to resolve the pseudonyms to the identifiers using the Secret (private) pseudonymisation Key (SK). The entity who applies the pseudonymisation function and the entity who can resolve the pseudonyms into the original identifiers do not have to share the same knowledge.

For example, a data controller can make its public key available to its data processors. The data processors can collect and pseudonymise the personal data using the PK, but the data controller is the only entity which can later compute the initial data from the pseudonyms, thanks to its SK. Such a scenario is strongly related to the generic scenario of a data processor being the Pseudonymisation Entity, with the additional advantage, in terms of protecting individuals' identities, that the processors do not have the pseudonymisation secret (and that they do not store the mapping between original identifiers and the derived pseudonyms).

Similarly, a Trusted Third Party (TTP) may share its public key with one or more data controllers, remaining nonetheless the only one to be able to resolve any pseudonym created by any data controller using its private SK (e.g. at the request of a data subject). Such scenario may also be relevant to cases of joint controllership, where a controller is performing the pseudonymisation and another controller only receives the pseudonymised data for any admitted further processing.[11]

b. Homomorphic Encryption

Certain asymmetric encryption schemes support homomorphic operations, i.e., a specific type of encryption allowing a third party (e.g. a cloud service provider) to perform specific computations on the ciphertexts without having knowledge of the relevant decryption key. In other words, homomorphic encryption (HE) allows performing computations on encrypted data without first decrypting it: the computations themselves are also cyphered and, once decrypted, the outputs

[10] Fischer-Hübner, S. (2009). *Privacy-Enhancing Technologies*. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems, pp. 2142–2147. Springer, Boston, MA.

[11] As pointed out by the ENISA in the aforementioned 'Pseudonymisation Advanced Techniques and Use Cases' report, "*Several pseudonymisation schemes based on asymmetric encryption have already been proposed. A typical application is to make available healthcare data to research groups; more precisely, by using fully randomised pseudonymisation schemes based on asymmetric cryptography (…), we may ensure that the identifiers (e.g. social security number or medical registration number or any other identifier) of a given patient are not linkable. For instance, a participant may have different local pseudonyms at doctors X, Y, Z, and at medical research groups U, V, W – thus providing domain-specific pseudonyms to ensure unlinkability between these different domains; by these means, doctors will store both the real name/identity of their patients and their local pseudonyms, but researchers will only have (their own) local pseudonyms*".

are identical to what would have been produced if the computations would have been performed on the original plaintext data.

HE uses a public key-generation algorithm to generate a pair of private and public keys, and an evaluation key which is needed to perform computations on the encrypted information when it is shared with the entity that will perform them. This entity does not need access to the private key to perform the analysis: the client (e.g. a data controller), who retains the private key, can then decrypt the output to obtain the results it requires. Any entity that has only the public and the evaluation keys cannot learn anything about the encrypted data in isolation.

c. Zero-Knowledge Proof

A well-known cryptographic primitive is the so-called Zero-Knowledge Proof (ZKP), i.e., any protocol by which a party (prover) is able to prove to another party (verifier) that he/she is in the possession of a secret, without revealing any information about the secret itself. More generally, ZKP can be leveraged to prove that a statement is true, without revealing any details of the statement.

In the context of pseudonymisation, if an individual associated with a pseudonym needs to prove that he/she is the owner of that pseudonym, without revealing his or her exact identity, a ZKP may provide the solution[12].

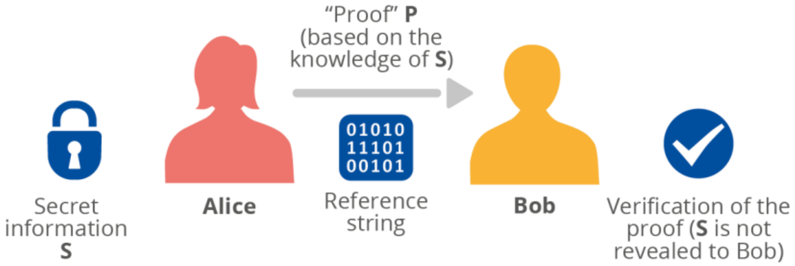


Figure 3 – Zero-knowledge proof for pseudonymisation

d. Secure Multi-Party Computation

A Secure Multiparty Computation (MPC) protocol allows the participating parties to jointly compute a function over their input data while keeping those input data private. Essentially, it removes the need for a trusted third party to view and manage the data. All parties (or a subset of the parties) may learn the result, depending on the nature of the processing and how the protocol is configured. SMPC uses a cryptographic technique called 'secret sharing', referring to the division of a secret and its distribution among each of the parties. This means that each

[12] An example of this scenario is provided by 'anonymous transactions' in cryptocurrencies: ZKP is used to allow verification of the transactions without the verifiers (miners) knowing anything about the transactions' contents (and the senders and the receivers of the transactions are concealed).

participating party’s information is split into fragments to be shared with other parties. Secret sharing is not the only way to perform SMPC, but it the most common approach used in practice. Each party’s information cannot be revealed to the others unless some proportion of fragments of it from each of the parties are combined (see Figure 4 below, based on a highly simplified example made by the UK Information Commissioner’s Office – ‘**UK ICO**’ – in its guidance on ‘Privacy-enhancing technologies’, dated June 2023). As this would involve compromising the information security of a number of different parties, in practice it is unlikely to occur. This limits the risks of exposure through accidental error or malicious compromise and helps to mitigate the risk of insider attacks and other types of data breaches.

Example

Three organisations (Party A, Party B and Party C) want to use SMPC to calculate their average expenditure. Each party provides information about their own expenditure – this is the “input” that will be used for the calculation.

SMPC splits each party's information into three randomly generated ‘secret shares’. For example, Party A’s input – its own total expenditure – is €10.000. This is split into secret shares of €5.000, €2.000 and €3.000. Party A keeps one of these shares, distributes the second to Party B and the third to Party C. Parties B and C do the same with their input data.

Party	Input data	Secret share 1 (to be kept)	Secret share 2 (to be distributed)	Secret share 3 (to be distributed)
A	€10,000	€5,000	€2,000	€3,000
B	€15,000	€2,000	€8,000	€5,000
C	€20,000	€7,000	€4,000	€9,000

When this process is complete, each party has three secret shares. For example, Party A has the secret share it retained from its own input, along with a secret share from Party B and another from Party C. The secret shares cannot reveal what each party's input was (i.e., Party A does not learn the total expenditure of Parties B or C), and so on.

Each party then adds together their secret shares. This calculates a partial result both for each party and the total expenditure of all three. The SMPC protocol then divides the total by the number of parties – three, in this case – giving the average expenditure of each: €15,000, as based on the original amounts, but no single party is able to learn what the other's actual expenditure is.

Party	Input data	Secret share kept	Secret share Received	Secret share Received	Partial Sum
A	€10,000	€5,000	€4,000	€5,000	€14,000
B	€15,000	€2,000	€2,000	€9,000	€13,000
C	€20,000	€7,000	€8,000	€3,000	€18,000

Figure 4 - Simplified application of SMPC

SMPC can be used to enable privacy in both the inference and training phases of machine learning systems. Oblivious model inference allows a client to submit a request to a server holding a pre-trained model, keeping the request private from the server S and the model private from the client C. In this setting, the inputs to the SMPC are the private model from S, and the private test input from C, and the output (decoded only for C) is the model's prediction.

Therefore, SMPC can be leveraged to secure the data and protect the patients' privacy while allowing technical partners to train the envisaged DataTools4Heart models, based on the clinical partners' combined datasets, without exposing such data and so ensuring privacy-by-design.

2 Privacy-Enhancing Technologies

Although the concept of PETs is far from new and their use is spreading, it has never had a universally accepted definition. As pointed out by the Organisation for Economic Co-operation and Development ('**OECD**') in its recent and detailed report on 'Emerging Privacy Enhancing Technologies - Current Regulatory and Policy Approaches', "*[t]he absence of a stable definition in this field can hinder a concerted analysis by policy makers, and privacy enforcement authorities (PEAs) in particular, of the potential impacts of PETs on data protection and privacy assessments*" (§2.1).

The ENISA refers to PETs as "*software and hardware solutions, i.e. systems encompassing technical processes, methods, or knowledge to achieve specific privacy or data protection functionality or to protect against risks to privacy of an individual or a group of natural persons*"[13].

According to the UK Information Commissioner's Office, in its turn, "*PETs are technologies that embody fundamental data protection principles by minimising personal data use, maximising data security and/or empowering individuals*"[14].

Like most of the major technological innovations, PETs are also subject to a substantial underlying ambiguity. On the one hand, they offer new functionalities and solutions that can assist with the implementation of the basic privacy principles, such as data minimisation, purpose limitation, privacy-by-design and by-default and data security, guaranteeing the accountability of both data controllers and processors. On the other hand, PETs can also challenge the implementation of other principles and privacy obligations, as in the case when a data controller using encrypted data processing tools may lose the ability to "see" – and so control – data feeding into their models. This would lead, under certain circumstances, to the impossibility to follow-up on any requests to exercise specific rights (e.g. of access or portability, if applicable) made by the data subjects, as well as to ensure that the data undergoing processing are accurate, complete and kept updated.

In this light, PETs should not be regarded as "silver bullet" solutions. They cannot substitute legal frameworks, but operate within and according to them, so that their applications will need to be

[13] ENISA's 'Readiness Analysis for the Adoption and Evolution of Privacy Enhancing Technologies' dated 31 March 2016.

[14] The already mentioned and more recent update of the ICO's guidance on 'Privacy-enhancing technologies', dated June 2023.

combined with legally binding and enforceable obligations to protect privacy and data protection rights, in the event that no specific derogation exists in predetermined context (e.g. scientific research).

Combining a number of studies and research and observing major developments in the private sector, including by academic institutions such as the Massachusetts Institute of Technology and specialised agencies such as the US National Institute of Standards and Technology, the OECD proposed to divide PETs into the following four categories[15]:

- ✓ **Data obfuscation tools** include zero-knowledge proofs, differential privacy, synthetic data and anonymisation and pseudonymisation tools. These tools increase privacy protections by altering the data, by adding "noise" or by removing directly or indirectly identifying details. Obfuscating data enables privacy-preserving machine learning and information verification, without requiring personal data collection and processing. These tools can leak information if not implemented carefully, e.g. (allegedly) anonymous data can still allow singling-out, with the help of data analytics and complementary data sets.

- ✓ **Encrypted data processing tools** include homomorphic encryption, secure multi-party computation (including private set intersection), as well as trusted execution environments. Encrypted data processing PETs allow data to remain encrypted while in use (so called 'in-use encryption'), so avoiding the need to decipher them before processing or computing. For This kind of PETs were widely deployed in COVID19 tracing applications, but their average computation costs tend to be high.

- ✓ **Federated and distributed analytics** allows executing analytical tasks upon data that are not visible or accessible to those executing the tasks. In federated learning, for example, data are pre-processed at the data source, as in the case of DataTools4Heart clinical partners: this means that the personal data don't have to leave their repositories and systems, because the computation takes place locally and the model is then trained at central level thanks to already aggregated inputs. In this way, only the summary statistics/results are transferred to the technical partners executing the tasks (in quality as data processors, in DataTools4Heart). Federated and distributed analytics requires reliable connectivity to properly operate.

- ✓ **Data accountability tools** include accountable systems, threshold secret sharing, and personal data stores. These tools do not primarily aim to protect the confidentiality of personal data at a technical level and are therefore often not considered as PETs in the strict sense. However, these tools seek to enhance privacy and data protection by enabling data subjects' control over their own data, and to set and enforce rules for when data can be accessed. Most tools are in their early stages of development, have narrow sets of use cases and lack stand-alone applications.

[15] See the aforementioned report on 'Emerging Privacy Enhancing Technologies - Current Regulatory and Policy Approaches'.

PETs can be seen as the underpinnings of a new evolving paradigm of privacy and data protection, as they provide more control to data subjects and enhance social trust in the processing of big data, implementing data minimisation.

In its data protection glossary, the European Data Protection Supervisor ('**EDPS**')[16] describes PETs as "*a coherent system of ICT measures that protects privacy by eliminating or reducing personal data or by preventing unnecessary and/or undesired processing of personal data, all without losing the functionality of the information system*"[17].

From a technical perspective, they can be perceived as building blocks towards meeting data protection principles and the obligations on privacy-by-design set out by Art. 25 of GDPR.

For this reason, a growing number of policy makers and supervisory authorities are considering how to incorporate PETs in their domestic legislations. However, the highly technical and fast evolving nature of these technologies often presents a barrier to implementation by organisations and to their consideration in policy and legal frameworks applicable to personal data and, more in general, to new technologies.[18]

Though, a crucial regulatory change could be on the horizon.

The European Data Protection Board ('**EDPB**')[19] announced, with a view to reinforcing the application of fundamental data protection principles and individual rights and to establishing common positions and guidance in the context of new technologies, in its official 'Work Programme 2023/2024' adopted on 14 February 2023, that it will lay down (i) new 'Guidelines on anonymisation' and (ii) new 'Guidelines on pseudonymisation'[20].

Such a perspective, combined with a game-changing judgement adopted by the Court of Justice of the European Union on 26 April 2023 – better examined below (§3.1) – which clarified when and by whom data can be considered anonymous[21], may help overcoming the highly restrictive and patchy interpretation of anonymisation and pseudonymisation that many Data Protection authorities have traditionally embraced and implemented so far.

Trying to be one step ahead of the times, DataTools4Heart will leverage a combination of advanced pseudonymisation techniques and PETs (particularly, Federated Learning, Differential Privacy, SMPC and Synthetic Data), fine-tuning them with specific reference to the healthcare domain, so as to ensure

[16] The EDPS is the European Union's (EU) independent data protection authority.

[17] The glossary can be read here (link).

[18] This is one of the reasons why the ENISA suggested that "*Regulators (e.g. Data Protection Authorities and the European Data Protection Board) and the European Commission should promote the establishment of relevant certification schemes, under Article 42 GDPR, to ensure proper engineering of data protection*" ('Data Protection Engineering – From theory to practice' report, dated January 2022).

[19] The EDPB is an independent European body, composed of representatives of the national Supervisory Authorities and the European Data Protection Supervisor (the EU Commission participates in the activities and meetings of the Board without voting right). Its tasks consist primarily in providing general guidance on key concepts of the GDPR and the Law Enforcement Directive, advising the EU Commission on issues related to the protection of personal data and new proposed legislation in the European Union, and adopting binding decisions in disputes between national supervisory authorities.

[20] The document is available here (link).

[21] SRB v. EDPS (Case T-557/20).

both controllers' and processors' accountability when processing the data and to set a high threshold of data security, to the benefit of the patients involved, the research community and, more generally, the society as a whole.

In this sense, besides Secure Multi-Party Computaion – which was already described above (it allows computation or analysis on combined data without the different parties revealing their own private inputs to the computation) – DataTools4Heart technological and legal architecture is based on (i) a federated approach to machine learning; (ii) data synthesis as a new method of anonymisation; (iii) global differential privacy.

2.1 Federated Learning

Federated Learning ('**FL**') is a type of remote execution in which models are 'sent' to remote data-holding machines (e.g. servers) for local training. This allows researchers to process datasets residing – and remaining – at other sites for training models without accessing those data. Implementing this approach in DataTools4Heart results in enabling technical partners in different jurisdictions to train models on all clinical partners' data relating to patients with heart failure, even as that data remains 'invisible' to each of them.

In brief, FL is a data-private machine learning technique which allows to collaboratively train AI models across multiple hospitals without the need to exchange any local data. This technique enhances both data security and minimisation, in line with the GDPR and all data protection applicable laws. In contrast with traditional approaches that require to gather the patients' data to a centralised location to train machine learning models, FL ensures that all data will be kept at their respective clinical site's repository, thereby ensuring strong patients' privacy.

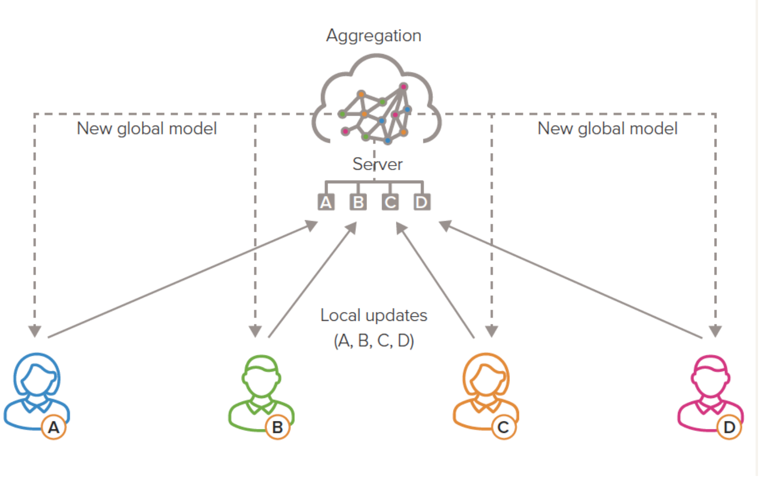


Figure 5 - Federated machine learning illustration[22]

[22] This figure is taken from the report '*From privacy to partnership: the role of Privacy Enhancing Technologies in data governance and collaborative analysis*', dated January 2023, prepared by The Royal Society in close collaboration with the Alan Turing Institute.

There are two approaches to accomplishing Federated Learning:

- ✓ Centralised: in this case, each site analyses its own data and builds a model, which is then shared to a remote, centralised location (a node) common to all researchers involved. This node then combines all models into a 'global' one and shares it back to each site, where researchers can use the new, improved model. Practically, a co-ordination server creates a model or algorithm and duplicate versions of that model are sent out to each distributed data source (i.e., the clinical partners). The duplicate model trains itself on each local data source and sends back the analysis generated, so that it can be synthesised with all the others coming from other data sources and integrated into the centralised model by the coordination server. This process repeats itself to constantly refine and improve the model.

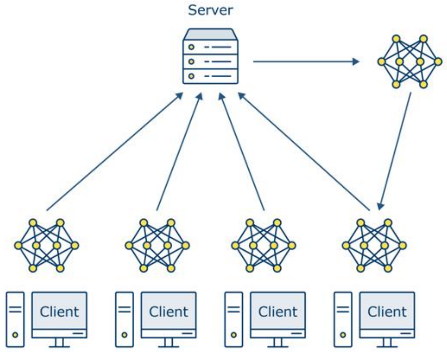


Figure 6 - Centralised FL in UK ICO's guidance on PETS

- ✓ Decentralised: in this different hypothesis, the model is built iteratively, as the remote node and local nodes take turns sending and returning information. Each participant sends a gradient on its dataset until the algorithm converges and the iterations use an optimisation routine (such as stochastic gradient descent). Hence, there is no central co-ordination server involved in decentralised FL: each clinical site communicates with each other, and they can all update the global model directly.

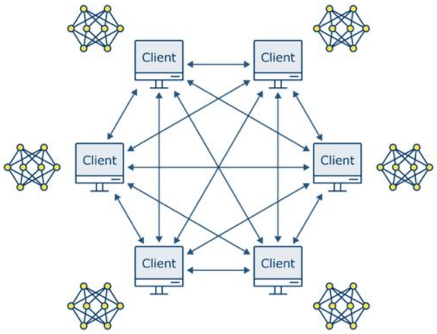


Figure 7 - Decentralised FL in UK ICO's guidance on PETS

In either approach, all the models which must be developed by technical partners are improved by 'learning' from remote datasets, which are never revealed. This is because FL ensures that raw identifiable data are never pulled out of the data controllers' own repositories and thus shared, preventing the most common issues associated with data protection, such as the risk of breaches.

Nonetheless, models are still vulnerable to some advanced attack. FL applications, for instance, can leak information in the parameters that are sent back to the central node, as local models may preserve features and correlations from the training data samples that could then be extracted or inferred by attackers.

This is why, in DataTools4Heart, the following additional PETs will be implemented:

- SMPC to address this issue by encrypting the shared model parameters, in order to ensure that they do not reveal their inputs;
- Differential Privacy to add sufficient noise so as to make reasonably impossible singling-out any individual whose data were used in connection with any FL-orchestrated training task.

2.2 Differential Privacy

Differential privacy ('**DP**') is a property of a dataset or database, based on the randomised injection of noise, providing a formal mathematical guarantee about people's unidentifiability.

A crucial characteristic of DP is the concept of "epsilon" or *ε* (also known as the "privacy budget" or "privacy parameter"), which determines the level of added noise. More precisely, it represents the worst-case amount of information inferable from the result by any third party about someone, including whether or not they were included in the input dataset.[23]

Therefore, noise allows for 'plausible deniability' of a specific person's personal information being in a record, implying that it is not possible to determine with reasonable confidence that data relating to that individual are included in a given set of information. That is to say that DP allows for risk to be quantified as the probability of reidentification, allowing the controller to 'dial up or down' and adjust for performance privacy trade-offs by referring to a set 'privacy budget'.

This could be exemplified thinking of a situation where someone asks someone else the question: "*Do you like icecream?*", which is a binary – yes or no – answer. However, it could be modified with the aid of a coin toss.[24] Prior to answering, a coin is tossed: if a head is the result, the person answering the question tells the truth; if not heads, the person will give a 'random' answer (which in this case is another coin toss with a predefined "yes" if heads and "no" if not). Notably, though it is possible to deduce the probability of people who like icecream, the individuals answering this question now have a so-called 'plausible deniability'. Indeed, although combining some basic facts about the independence of events may produce a probability distribution, because of the introduction of

[23] Lee, J., and Clifton, C. (2011). *How Much Is Enough? Choosing ε for Differential Privacy*. In: Lai, X., Zhou, J., Li, H. (eds) Information Security. ISC 2011. Lecture Notes in Computer Science, vol 7001. Springer, Berlin, Heidelberg (link).
[24] Dwork, C., and A. Roth. 2014. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9 (3–4): 211–407 (link).

randomness (i.e., a person's veracity depends on a coin toss) which produces deniability, the individuals may now say "*I may or may not be 'in' the database*". the individuals are now permitted to say "I may or may not have answered truthfully". And this is the gravitas of differential privacy[25].

In a clinical setting, DP algorithms "*can provide assurance that after analyzing a dataset of several individuals, the outcome of the analysis will not be affected and will remain the same, even if any individual's data (up to ε) was not included in the dataset*"[26]. Which means that it allows studying larger statistical trends in a dataset, while protecting the personal information referring to the individuals who participate in it.

There are two methods for the privacy budget to be enforced:

✓ interactive DP: where the noise is added to each query response and querying is terminated once the privacy parameter ε is met (i.e., when the information obtained from queries reaches a level where personal data may be inferred);

- non-interactive DP: where the privacy budget is set *a priori*, as a property of the dataset itself. Non-interactive mechanism computes some function from the original database and releases the output once and for all, so that it can then be used by anyone to compute the answer to a particular class of queries, without requiring any further interactions with the DP curator. In this case, privacy protection algorithms are processed and published to the database, and users can process this database for any operation.

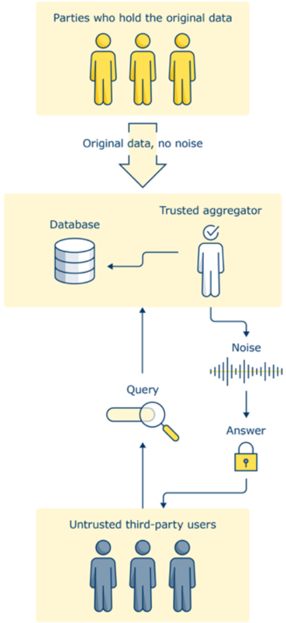


Figure 8 - Centralised DP in UK ICO's guidance on PETS

Noise can be added at the time of data collection (distributed DP), or at the central location before the data are released (centralised DP). In more detail:

a. Centralised (or global) DP: it involves an "aggregator" having access to the real data. Each user of the system sends information to the aggregator without prior adding noise. The central node then applies DP, by adding noise to the output during computation of the final result, before it is shared with any third party. As evident, the key requirement – and so main disadvantage – of this approach is that all users have to trust the aggregator to act appropriately and protect people's privacy, as it has to access the data in clear.

b. Distributed (or local) DP: in this scenario, each user of the system (or a trusted third party on its behalf) must apply the DP mechanism before sending the data to the aggregator, preventing the issue relating to the need for trust in the central node. Since noise is added at the

[25] This example is taken from Bellovin, Steven M., Preetam K. Dutta, and Nathan Reitinger. 2019. *Privacy and synthetic datasets*. Stanford Technology Law Review 22 (1): 2–52 (link).
[26] '*Data Protection Engineering – From theory to practice*' report by ENISA, dated January 2022.

individual input data level, the total ‘amount’ of noise required is (much) larger than in global differential privacy, entailing a decrease in data accuracy. However, proper implementation of secure aggregation techniques, including SMPC, can help appropriately addressing this hurdle.

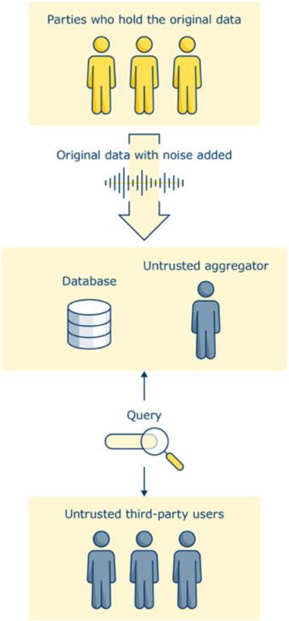

Parties who hold the original data
Original data with noise added
Database
Untrusted aggregator
Query
Untrusted third-party users

Figure 9 - Local DP in UK ICO’s guidance on PETS

Both methods of DP can help obtaining anonymous information as output, provided that a sufficient level of noise (privacy budget) is added to the data, it being understood that any original information kept by the aggregator in the global model, or by the individual parties in the local approach, represent personal information.

Another important propriety of DP is that post-processing is allowed, meaning that the result of the processing of differential private data through a fixed transformation remains differential private (e.g. if a generative model is trained using an algorithm satisfying DP, the samples obtained can be published and processed without further privacy implications).

Finally, it must be highlighted that DP and Federated Learning can be combined in two ways: output perturbation (where noise is added to the output of an optimisation algorithm) and objective perturbation (where noise is added at every step of the optimisation algorithm), achieving high security levels and privacy-by-design.

2.3 Synthetic data

2.3.1 Technical status of synthetic data

There are several types of synthetic data, but the term essentially refers to the generation of artificial data with the aim of reproducing the statistical properties of an original dataset. This is made by learning relevant distributions of real data using AI driven tools, for then mimicking and sampling them to produce realistic, but totally fake datasets having the very same statistical properties of the original ones, so enhancing the protection of personal data and patients' privacy, while maintaining the intrinsic utility of the novel dataset for statistical analysis and medical research.

A more methodological-oriented definition has been provided in a report specifically dedicated to synthetic data which was commissioned by The Royal Society and the Alan Turing Institute, whereby this type of data is described as "*data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)*" [27].

These data can be either partially (e.g. where only some of the variables have been generated by a ML-driven model) or fully synthetic, with the former containing generated data alongside original data, and the latter composed exclusively of algorithmically generated data. Also, hybrid dataset can be produced from both the real-world set and the fully synthetic one, to better represent the specifics of the original data. To do this, it is possible, for example, for each of the real points, to select the closest point in the synthetic set: this will make it possible to reproduce certain special cases of the source set (e.g. a specific clinical feature), without directly using the real data.

Many practical alternatives exist for generating synthetic data[28]. The easiest option is drawing samples from a known distribution. In this case, the outcome does not contain any original (and personal) data and re-identification is unlikely to occur, mainly due to randomness. More complex methods rely on mixing real data and fake ones (the latter being still sampled from known multivariate distributions, conditioned on the real observed data). In this case, some disclosure of personal data and re-identification is possible due to the presence of true values within the dataset, unless additional measures are implemented to prevent any reasonably foreseeable risk of singling-out.

The main difference compared to traditional anonymisation techniques is that data synthesis is based on the addition of statistically similar information to the original data, rather than on the stripping away of unique identifiers[29].

In any case, by preserving the overall statistical properties, analysing synthetic data can lead to the same research and mathematical conclusions as the analysis of the original data source.

[27] Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. Synthetic Data – What, why and how? arXiv:2205.03257, 2022 (link).

[28] The model can take many forms, from deep learning architectures such as the popular Generative Adversarial Networks (GANs), or Variational Auto-encoders (VAEs), through agent-based and econometric models, to a set of (stochastic) differential equations modeling a physical or economic system (see the report mentioned in Note 23).

[29] Bellovin, S.M., Dutta, P.K., Reitinger, N., *Privacy and synthetic datasets*. 2019, Stanford Technology Law Review 22 (1): 2–52.

As said, methods to produce this type of data vary, but the underlying principle is that some or all of the values in the original dataset are substituted, by specific algorithms (such as in particular, GANs – Generative Adversarial Networks)[30], with others taken from statistically equivalent distributions and structures, to create entirely new records with as little traceable relation to the originals as possible.

This technique can hence help researchers prototype data-driven models and verify and validate machine learning pipelines, providing some assurance of performance and enforcing privacy-by-design.

Among the various benefits associated with data synthesis, given the objectives of DataTools4Heart, it is worth mentioning the following:

i. **de-biasing**: given biased training data, a natural approach is to train models using available data; these biases can then be seen in the output of the trained models. Rather than attempting to debias each trained model individually, one could generate a de-biased synthetic dataset and use it to train each model, creating a unified approach for handling biases across an organisation. Such data can be used then for training 'black box' ML pipelines, while mitigating the risk of historical biases being amplified;

ii. **data augmentation and imputation**: synthetic data can be generated with the specific aim to enlarge datasets that are too small, e.g. to provide robustness against 'outlier' examples (when they are not needed for the purpose of a medical research). This is often referred to as semi-supervised learning. Synthetic data can both act as a regulariser, reducing variance in the learned downstream model, and be expanded for imputating (replacing missing values with substitutes), i.e., filling gaps, correcting skewed value distributions, or removing spurious values in the original data. This addresses collection, formatting or normalisation issues, which are pervasive especially in clinical datasets, and thus produces data that are actually more informative and realistic than the original ones[31]. In addition, training machine (and in particular deep) learning networks and models – even more in the scientific area – requires vast amounts of correctly labelled data, which is often costly to produce. Synthetically generated labelled data offer a cost-efficient solution to this challenge;

i. **data minimisation**: a large number of real-world examples demonstrate that high-dimensional, often sparse, datasets are inherently vulnerable to privacy attacks and that existing anonymisation techniques too often do not provide adequate protection. By using synthetic data, "*a controller will be respectful of individuals' confidentiality, since they differ from real data and the generation and processing of synthetic data does not invade the personal sphere of data subjects (in particular when real data refer to individuals' sensitive*

[30] A Generative Adversarial Network uses two models playing against each other: the 'Generator' learns to capture and recreate the data distribution, while the 'Discriminator' estimates the probability that a generated sample belongs to the original data distribution or rather has been created by the Generator, so determining whether the data is fake or not.

[31] Morley-Fletcher, E., '*New Solutions to Biomedical Data Sharing: Secure Computation and Synthetic Data*', in *Personalized medicine in the making: philosophical perspectives from biology to healthcare*. Beneduce, C., Bertolaso M., Springer (2021), 173-189.

> *characteristics, or to rare attributes that may be difficult to retrieve or may have a significant power of identification)*"[32].

That being said, it is not hard to see why synthetic data are gaining exponential attention in these years. The EDPS itself included data synthesis in both the (2021-2022 and 2022-2023) releases of its Tech Sonar (a report aimed to anticipate emerging technology trends and better understand future developments in the technology sector from a data protection perspective)[33], endorsing the advantages described above.

After describing the pros, the EDPS also identifies the cons which must be properly addressed by data controllers (or by processors acting on their behalf):

- ✓ output control: especially in complex datasets, the best way to ensure the output is accurate and consistent is by comparing synthetic data with original data, or human-annotated data. However, for this comparison to be carried out, access to the original data is required;
- ✓ difficulty to map outliers: as synthetic data can only mimic, replicating specific properties of a phenomenon, but not duplicate real-world data, they may not cover some outliers in the original dataset, which though can sometimes be more important than regular data points (e.g. in clinical research)[34];
- ✓ quality of the model: quality of synthetic data is highly correlated with that of both the original dataset and the machine learning model used to generate the fake data, which thus may reflect and incorporate the same biases of the original data. Furthermore, the manipulation of datasets to create fair synthetic datasets might result in inaccurate data. Therefore, controllers will always need to reconcile a tension between different data protection principles, especially if the result of the processing entails consequences (e.g. legal or health implications) for data subjects.

All of these well-known challenges will be specifically addressed within the project.

2.3.1.1 Synthetic data (and other PETs) in DataTools4Heart

First of all, to prevent patients' reidentification in line with data minimisation and privacy-by-design obligations posed by applicable legislation, data synthesis tool will be coupled with Differential Privacy, so implying the generation of fake datasets which do not "*contain any information that can be traced back to specific individuals in the original data*"[35].

[32] '*Data Protection Engineering – From theory to practice*' report by ENISA.
[33] The last version of the Tech Sonar, published in November 2022, is available here.
[34] For example, it would be very difficult to 'hide' a multi-billionaire in a synthetic dataset containing information about average wealth. A synthetic data generator would either not accurately replicate statistics regarding very wealthy people or would otherwise reveal potentially private information about these individuals.
[35] National Institute of Standards and Technology, 'Pioneering Data Privacy Research & Resources', March 2021 (link).

As said above, a differentially private synthetic dataset looks like the original one – having the same schema and maintaining the same statistical properties (e.g., correlations between attributes) – but it provides a provable privacy guarantee for individuals in the original dataset[36].

The actual risk or re-identification can be effectively quantified in relation to the original data (DP parameter) and so be modulated, in the generative process, based on the intended use and distribution. The addition of well-calibrated noise ensures that the presence or absence of an individual in a dataset does not affect the query results.

In parallel, the flexible and modular Data Ingestion and Feature Extraction Suite which are being developed by technical partners, in connection with the project's cardiology Common Data Model based on the international, open health data exchange standard FHIR (HL7 Fast Healthcare Interoperability Resources)[37], will enable, *inter alia*: selecting data types; derivation of additional values; combining elements to create new cohorts; infer insight in the availability of data per each clinical center, as well as in patient trajectories; particularly, seamless evaluation of the extracted data, including: clustering methods, means, outlier assessment, clinical context.

Concretely, this means that all clinically relevant outliers – i.e., data points with some uniquely identifying features – will be appropriately captured to be then computed. However, to avoid singling-out and provide robustness against outliers examples, the differentially private data synthesis tool will dilute their intrinsic identifiable value, by enlarging relevant data through proper augmentation. Indeed, since singularities cannot be easily hidden in the masses, to prevent under-representation of minority groups in a context where outliers may be of the utmost importance, DP combined with synthetic data proves as the best state-of-the-art solution for data accuracy.

Regarding DP to secure the data synthesis pipeline, as participating hospitals can pose heterogeneous budget requirements on the final model – i.e., give different 'privacy parameter' (epsilon) values (larger epsilon means less privacy)[38] – multiple methods are being investigated to extract the best possible utility in all different scenarios.

The learned DP generative model can then be sampled through standard techniques (e.g., Markov Chain Monte Carlo, generative network forward computation, etc.) to generate synthetic data that is guaranteed to be differentially private. As an alternative approach, technical partners in charge will experiment with a novel idea, enforcing the DP budget at the sampling stage and not at the generative model training stage.

[36] Jordon J, Yoon J, van der Schaar, M. 2019 PATE-GAN: Generating synthetic data with differential privacy guarantees (link).

[37] This component is made of/based on the following tools/services: i. Data Ingestion Suite: an ETL engine and mapping definitions to convert source data into HL7 FHIR resources conforming to specified DT4H Common Data Model and store them into Health Data Repository; ii. Health Data Repository: a secure health data repository that provides FHIR compliant RESTful API to manage (create, read, update, delete) and search records. Its API is used by the Data Ingestion Suite and Feature Extraction Suite and expected to be used by the federated querying platform; iii. Feature Extraction Suite: a feature extraction engine and feature definitions that gets data from Health Data Repository and prepare datasets with the defined features that will be used for training Artificial Intelligence models.

[38] See § 2.2.

Summarising the above, security and privacy for the pseudonymous and synthetic health data will be attained through three key mechanisms:

- ✓ Federated Learning and analytics will allow running complex analysis and machine-learning pipelines locally, without exporting or publishing hospital data. The key idea is to bring analytics computation to each hospital's own node, thus requiring only the communication of secure aggregated statistics (e.g. model gradients) during the lifetime of FL pipelines;
- ✓ Secure Multi-Party Computation (SMPC) comprises a collection of powerful cryptographic primitives allowing the computation of data analytics over a collection of sites, ensuring that each of them learns nothing about the other sites' data (other than the final output of the analysis). SMPC protocols can offer strong, cryptographic guarantees on the privacy of data inputs and computations.
- ✓ Differential Privacy offers various mechanisms for computing analytics while offering strong, formal guarantees on individual privacy. Thus, DP effectively complements SMPC by guaranteeing that the outputs can no longer be reasonably identified.

2.3.2 Legal status of synthetic data

Both EDPS and EDPB recently devoted much attention to synthetic data, also organising a specific webinar (link)[39] and including this technology – as already said – in both the first and second release of *Tech Sonar*[40].

But more importantly, for the very first time, data synthesis finds its well-deserved place in the EU legislation, specifically in the 'Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence' (so-called Artificial Intelligence Act, '**AI Act**'), which is likely to be one of the most critical and impactful regulations ever adopted by the European Union[41].

Let's cut straight to the chase:

- ✓ **Article 10.5 of the AI Act** states that:

  «*To the extent that it is strictly necessary for the purposes of ensuring negative bias detection and correction in relation to the high-risk AI systems, the providers of such systems may exceptionally process special categories of personal data* [which include health data] *referred to in Article 9(1) of Regulation (EU) 2016/679 (…) subject to appropriate safeguards for the fundamental rights and freedoms of natural persons, including technical limitations on the re-use and use of state-of-the-art security and privacy-preserving. In particular, all the following conditions shall apply in order for this processing to occur:*

[39] Internet Privacy Engineering Network (IPEN) Webinar on the 16 June 2021, entitled "*Synthetic data: what use cases as a privacy enhancing technology?*".

[40] See Note 31.

[41] On 14 June 2023, the long-awaited EU Parliaments' negotiating position on the AI Act was adopted (link). The talks will now begin with EU countries in the Council, regarding the final form of this law. An agreement is expected by the end of this year. The full text of the draft adopted my MEPs can be read here.

a) *the bias detection and correction cannot be effectively fulfilled* ***by processing synthetic or anonymised data****;*

b) ***the data are pseudonymised****;*

c) *(…)*»

✓ **Article 54.1, (b) of the AI Act** establishes that:

«*In the AI regulatory sandbox personal data lawfully collected for other purposes may be processed for the purposes of developing, testing and training of innovative AI systems in the sandbox under the following cumulative conditions:*

a) *(…)*

b) *the data processed are necessary for complying with one or more of the requirements referred to in Title III, Chapter 2* [namely those applicable to high-risk AI systems] *where those requirements cannot be effectively fulfilled* ***by processing anonymised, synthetic or other non-personal data****;*

c) *(…)*»

In the world of law, these apparently modest words speak volumes regarding the legal statute of synthetic data.

Art. 10 pairs anonymised data and synthetic data, *de facto* considering both these types of data as equivalent and interchangeable alternatives, as further confirmed by: (i) the use of the word "or" in between, and (ii) the clear distinction marked with pseudonimised data, listed separately, and so qualified as a different category of data compared to the synthetic ones.

In its turn, Art. 54 provides a clear qualification of synthetic data, first grouping them with anonymous data and then explicitly equating synthetic data to non-personal data (the expression "*or other*" unquestionably signifies that both the categories of anonymous and synthetically generated data must be included in the same cluster of unidentifiable data, falling out of the scope of data protection legislation).

One of the first and fundamental rules of law is that legal provisions must be interpreted on the basis of their literal meaning, i.e., the exact words used by the Legislator. In brief, any free reading or understanding must be ruled out.

In this light, the intention of the EU Legislator to equalise anonymous and synthetic data – all the more in relation to crucial obligations imposed to AI developers – appears undeniable.

Notwithstanding this crucial indication in the AI Act, the need for market players to pass the 'reasonable non-identifiability' test that the law in force (Recital 26 of GDPR) establishes as a fundamental condition for considering the data as anonymous remains unchanged (see below for further details).

Practically, this means that a number of variables will have to be prudently evaluated by data controllers on a case-by-case basis – such as the type of generative model to be deployed, the context, means and purposes of the processing, the nature of the data, the category of data subjects, the use of Differential Privacy or other PETs as additional privacy-preserving layers, etc. – to comprehend whether any re-identifiability risk still remains notwithstanding the technical, organisational and legal measures implemented to prevent any reasonably foreseeable threats for the individuals and their data[42].

Furthermore, "*Because risk is likely to evolve over time depending on the context, events, time, or agents, the very nature of risk affects the legal determination of synthetic data in the way that, under certain circumstances, synthetic data will be considered anonymous data, while in others, this will not be the case*"[43].

Although such an assessment is inherently subjective and there is no single defined threshold of identifiability which can be used, best practice incorporates using quantifiable statistical assessments, where feasible, as well as conducting penetration or motivated intruder testing. Moreover, *ex-post* audit on re-identifiability and data protection impact assessment can assist in defining appropriate additional safeguards which may be required. Indeed, synthetic data requires validation of:

a. utility, the property that the performance and predictive power of algorithms based on ML-generated data is not substantially lower than performance on the original data;

b. obfuscation, the property that the synthetic data do not leak private information from the original data.

A thorough consideration of a range of factors relating to the data, the data environment and relevant mitigations and safeguards, will influence whether synthetic data are viewed as personal data or non-personal data. The question, as well as the judicial and supervisory authorities' divergent interpretations and academic debate as to what constitutes anonymisation, entirely turn on (i) whether the risk of re-identification must reach zero or not, and (ii) whether such risk must be evaluated in relation to anyone and everywhere (as it can be inferred by the Article 29 Working Party's 'Opinion 05/2014 on anonymisation techniques' – see below in §3), or only to some specific parties based on their functions and contractual agreements. Moreover, some more peculiar challenges should be borne in mind by regulators when dealing with synthetic data.

First, in case it is genuinely determined that synthetic datasets constitute 'personal data', this gives rise to significant complexity in determining how data protection rights and obligations should apply. For example, how can the principle of data accuracy and right to rectification be enforced to synthetic data, where it is not even clear if a relevant individual is the focus of the data? How does the right to object to processing apply if an individual's data has at some point been used to develop a model[44]?

[42] An article on synthetic data published by the EDPS and authored by Robert Riemann, reads that "*A privacy assurance assessment should be performed to ensure that the resulting synthetic data is not actual personal data. This privacy assurance evaluates the extent to which data subjects can be identified in the synthetic data and how much new data about those data subjects would be revealed upon successful identification*" (link).

[43] Fontanillo Lopez, C.A., Elbi, A., *On the legal nature of synthetic data*, Center for IT and IP Law, KU Leuven, NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research (link).

[44] Mitchell C., Redrup Hill, E., *Are synthetic health data 'personal data'?*; PHG Foundation, Cambridge University; 2023 (link).

Secondly, some practical and legal differences between pseudonymised or anonymised data and synthetic data substantiate the theory according to which this latter form of AI-generated information constitutes a third form of data, distinct from the previous ones. Indeed, the nature of some data synthesis techniques and models results by-definition in almost negligible identification risks.

Drawing a conclusion, the time has come – also following the path indicated by the aforementioned crucial ruling recently adopted by the CJEU (SRB v. EDPS, Case T-557/20) – for a clear change of approach by (all) data protection Supervisory Authorities and regulators regarding when, by whom and under what conditions, data can no longer be considered as identifying an individual (or a group of individuals), opening the door to state-of-the-art PETs, also combined together, and ML-driven synthetic data generation. If this step forward is not taken, without this evolution determining any limitation or detriment to data subjects' rights and privacy, the key objectives of Data Act and European Health Data Space will remain very hard to achieve[45].

3 Data anonymisation in the EU: the need for legal standardisation

In a well-known article published in December of 2020, two prominent researchers, Aloni Cohen and Kobbi Nissim, pointed out that "*There is a significant conceptual gap between legal and mathematical thinking around data privacy*"[46]. Nothing could be truer in relation to data anonymisation.

Notwithstanding the pivotal importance of the distinction between personal and non-personal data, it is extremely burdensome, in most of the cases, to differentiate between these categories. This difficulty is anchored in both technical and legal factors. From the first perspective, the increasing availability of data points and sources, as well as the continuing sophistication of data analysis algorithms – all the more in connection with machine learning – and performant hardware makes it easier to link datasets and infer personal information from ostensibly non-personal data.

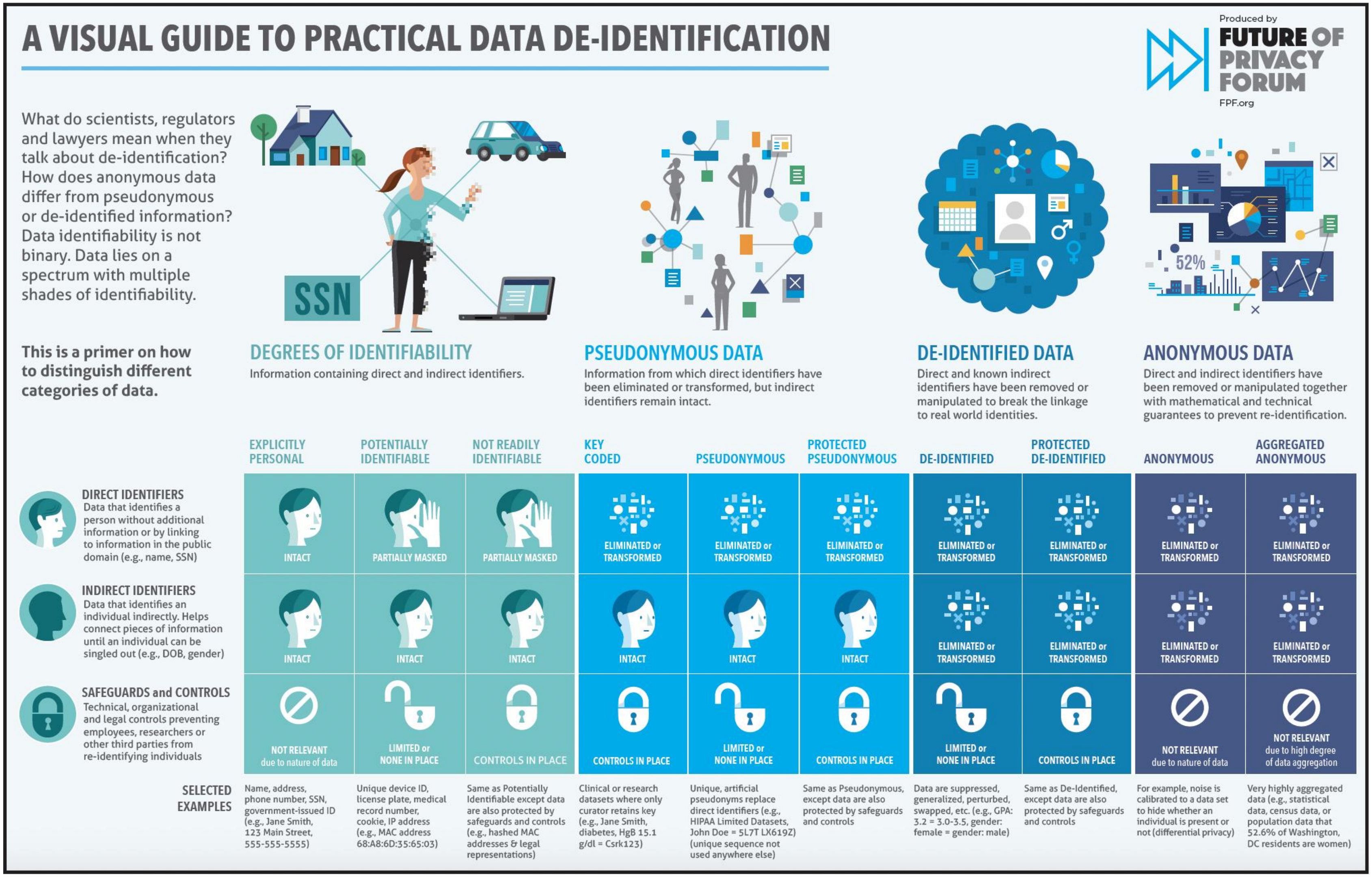
From a legal standpoint – even after 28 years of application of data protection legislation (including 5 under the GDPR)[47] – it is to date still not obvious what the correct legal test is that should be carried out to correctly categorise data as anonymous or not[48].

[45] Amongst many others, also BBMRI (Biobanking and Biomolecular Resources Research Infrastructure) stressed, in its official 'Statement on a European Health Data Space', dated 4 February 2021, that "*The existing regulatory framework seems insufficient to deliver on the promises of the EHDS. Health data governance remains fragmented at national and regional level, hindering any effort to scale up research and healthcare solutions. Most importantly, it is necessary to protect and promote the use of health data, defining clear pan-European rules to overcome the existing gaps in practice*" (link).

[46] Cohen A; Nissim K., *Towards Formalizing the GDPR's Notion of Singling Out*, 117, Proceedings of the National Academy of Sciences, 8344, 2020 (link).

[47] The Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data became applicable in December of 1995, while the GDPR entered into force in May of 2018.

[48] Finck, M., Pallas, F. *They who must not be identified - distinguishing personal from non-personal data under the GDPR*. International Data Privacy Law, 2020, Vol. 10, No. 1 (link).



A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by FUTURE OF PRIVACY FORUM FPF.org

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

Figure 10 - A visual guide to practical data de-identification (by the Future of Privacy Forum)

In 2016, the above practical guide was included in a valuable article on de-identification techniques[49]. No major changes occurred from that date as to the regulatory scenario on data anonymisation[50].

The already often-mentioned Recital 26 of the GDPR (see below) specifies that data are anonymous if it is 'reasonably likely' that they cannot – or can no longer – be linked to an identified or identifiable individual. By contrast, many national Supervisory Authorities and particularly the Article 29 Working Party ('**WP29**', now replaced by the EDPB) have, however, provided very rigid interpretations of the concept that conflict with this legislative text.

Recital 26 – GDPR

«The principles of data protection should apply to any information concerning an identified or identifiable natural person (…) 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. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable (…)»*.

Regarding the risk of re-identification, emphasis is placed by the EU legislator (in the GDPR) on the following criteria:

- ✓ reasonability, meaning that the relevant risks must not be completely excluded (in respect to anybody and any time);

[49] Polonetsky J., Tene O., Finch K. *Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification*. 56 Santa Clara L. Rev. 593, 2016 (link).

[50] In April 2021, the EDPS and the AEPD (Agencia Española de Protección de Datos) issued their 'Joint paper on 10 misunderstandings related to anonymisation' (link), whereby they stressed the following: 1) Pseudonymisation is not the same as anonymisation; 2) encryption is not anonymisation, but it can be a powerful pseudonymisation tool; 3) it is not always possible to lower the re-identification risk below a previously defined threshold, whilst retaining a useful dataset for a specific processing; 4) risks exist that some anonymisation processes could be reverted in the future. Circumstances might change over time and new technical developments and the availability of additional information might compromise previous anonymisation processes; 5) anonymisation process and the way it is implemented have a direct influence on the likelihood of re-identification risks; 6) it is possible to analyse and measure the degree of anonymisation (e.g. through Differential Privacy); 7) automated tools can be used during the anonymisation process, however, given the importance of the context in the overall process assessment, human intervention is needed on a case-by-case basis; 8) a proper anonymisation process keeps the data functional for a given purpose; 9) anonymisation processes need to be tailored to the nature, scope, context and purposes of processing, as well as the risks of varying likelihood and severity for the rights and freedoms of the data subjects; 10) personal data has a value in itself. Re-identification of an individual could have a serious impact for his rights and freedoms.

- ✓ current technology state-of-the-art, meaning that no data controller or processor is required to foresee what the future of technological progress will bring in terms of reverting the de-identification measures to single-out again the individuals to whom the data relate;
- ✓ means available to each operator/stakeholder involved in the processing to spot (or infer characteristics regarding) a specific individual, implying that the risks must be measured and differentiated based on each individual's position, without anonymisation having to be such 'for everyone and not just for some'.

Notwithstanding the above risk-based – and so flexible – approach established by the applicable law, in its 2014 guidelines on anonymisation and pseudonymisation[51], the WP29 adopted a very different and strict stance, prescribing a zero-risk test. Indeed, the WP29 specifies that:

- ✓ "*anonymisation results from processing personal data in order to irreversibly prevent identification*";
- ✓ "*the outcome of anonymisation as a technique applied to personal data should be, in the current state of technology, as permanent as erasure, i.e. making it impossible to process personal data*";

In addition, the WP29 crucially indicates 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*"[52].

Apart from being untrue in almost everyday practice, this interpretation met with considerable academic criticism as there are many scenarios where a controller can decide – or sometimes be required – to share anonymous data, while needing to keep the original dataset, as in the case where a hospital makes available anonymised data for research purposes while retaining the original data in clear for patient care.

In short, this approach by the WP29 – which has then been embraced also by some national data Protection Authorities – implies a rejection of the 'reasonable re-identifiability' approach set out by Recital 26 GDPR, as it considers the risk stemming from keeping the initial data to be intolerable in any event. Indeed, "*the concepts of irreversibility, permanence, and impossibility stand for a much stricter approach than that formulated by the legislative text itself*", with the effect that "*These diverging interpretations have prevented legal certainty as to what test ought to be applied in practice*"[53]. Confusion is also increased – and so reliance on anonymisation is made riskier – due to:

[51] Article 29 Working Party, *Opinion 05/2014 on Anonymisation Techniques* (WP216), adopted on 10 April 2014 (link).
[52] *Ibid*.
[53] Finck, M., Pallas, F. *They who must not be identified - distinguishing personal from non-personal data under the GDPR* (link).

i. the formalisation of the concept of pseudonymous data in the GDPR, because in some jurisdictions (e.g., in UK and Ireland)[54] and sectors (such as clinical trial), this type of data can, under certain circumstances, be considered as anonymous information;

ii. discordant interpretations by national competent Supervisory Authorities as to the degree of irreversibility that individual de-identification must achieve so that data can be deemed anonymous and not pseudonymous[55].

Concretely, this fragmentation results in situations where a company willing to implement a project (e.g., a clinical study, or a scientific research), or launch a new service or technology (e.g., based on generative AI) in more than one member State, which may entail or be based on the anonymisation of data and/or the use of anonymised data, must deal with many diverging applicable rules, deriving from the decisions taken by national Data Protection Authorities, which are often very different from a country to another.

As it has been highlighted, the solution to overcome this detrimental and still-locally-centered scenario depends on “*whether an ‘absolute’ or ‘relative’ approach is adopted; i.e. whether identifiability is judged according to the abilities of anyone and everyone (including the data controller) to re-identify data or whether the relevant question is whether the data are ‘identifiable’ in the ‘hands’ of a specific actor*”[56].

As obvious, the evaluation as to when re-identification may be considered reasonable (Recital 26 GDOR) heavily varies based on a number of factors which may affect the processing. The characterisation of data is context-dependent, so that personalisation of risk “*should not be seen as a property of the data but as a property of the environment of the data*”[57]. It is easy to figure out how risk prove to be very different subject to whether the entity which may attempt to reidentify the individuals is a private person, or a law enforcement agency, or rather a major online platform.

Regulators and Supervisory authorities did not establish – nor have reachd an agreement on this at EU level – regarding (i) what standard of reasonableness should be applied to weigh up the risk by a

[54] In its draft ‘Anonymisation, pseudonymisation and privacy enhancing technologies guidance’ (Chapter I), the UK Information Commissioner’s Office points out that *“(…) you will not always be able to state that a specific technique or set of controls will achieve these aims, particularly as technology changes over time. This means that even where you use anonymisation techniques, a level of inherent identification risk may still exist. However, this residual risk does not mean that particular technique is ineffective. Nor does it mean that the resulting data is not effectively anonymised for the purposes of data protection law when you consider the context. Also, data protection law does not require anonymisation to be completely risk-free. You must be able to mitigate the risk of re-identification until it is sufficiently remote that the information is ‘effectively anonymised*” (link). Also, the Irish Data Protection Authority deems that it is not “*necessary to prove that it is impossible for the data subject to be identified in order for an anonymisation technique to be successful. Rather, if it can be shown that it is unlikely that a data subject will be identified given the circumstances of the individual case and the state of technology, the data can be considered anonymous*” (link).

[55] TEHDAS (the Joint Action Towards the European Health Data Space) specified in a number of reports that “*there is a lack of common European interpretation of what constitutes ‘sufficient anonymisation’ to transform personal data to non-personal data*” and “*what constitutes ‘pseudonymisation’*” (amongst others, link).

[56] Mitchell C., Redrup Hill, E., *Are synthetic health data ‘personal data’?;* PHG Foundation, Cambridge University; 2023.

[57] Stalla-Bourdillon S., Knight A. *Anonymous Data v. Personal Data - A False Debate: An EU Perspective on Anonymisation, Pseudonymisation and Personal Data* (2017) 34 Wisconsin International Law Journal 284, 301 (link).

uniform method across all member States, and (ii) whether an objective or subjective approach should be adopted[58].

3.1 Judgement in case SRB v. EDPS (case T-557/20): a needed turning point

On 26 April 2023, the General Court of European Union (‘**General Court**’ or ‘**ECG**’)[59] adopted a crucial decision regarding the issue of data pseudonymisation, in the Case T-557/20 (‘**Judgement**’ or ‘**Decision**’).

The dispute arose from the Single Resolution Board’s (‘**SRB**’, the central resolution authority within the EU Banking Union) request to Deloitte to carry out certain assessments, aimed at determining whether the shareholders and creditors of Banco Popular Español would have received a better treatment if the bank had been subject to normal insolvency proceedings. In this context, the affected shareholders and creditors submitted five complaints under Regulation 2018/1725[60] to the European Data Protection Supervisor, alleging that the SRB failed to mention the transmission of data collected to third parties in its privacy statement, thereby violating its transparency obligations relating to the processing of personal data under said Regulation.

In this proceeding, the SRB argued that the disclosure did not concern personal data, since Deloitte only received the data in pseudonymised form (alphanumeric codes) that would not have allowed the firm to re-identify the shareholders. However, the EDPS still concluded that the transmission involved pseudonymised and therefore personal data, due to the existence of additional information that could have allowed the re-identification of complainants, although such information were held and accessible only by the SRB. Therefore, the SRB brought an action before the EGC seeking, *inter alia*, annulment of the decision of the EDPS.

Given that the concept of personal data within the meaning of Article 3(1) of Regulation 2018/1725, coincides with the definition provided by Article 4(1) of the GDPR, the general Court was essentially called upon to decide whether the information that has been transmitted to Deloitte had to be qualified as personal data, *i.e.*, information relating to an identified or identifiable natural person.

Applying the principles already laid down by the Court of Justice of the European Union in the previous well-known Breyer judgment (C-582/14), the EGC stated that “*in order to determine whether the information transmitted to Deloitte constituted personal data, <u>it is necessary to put oneself in Deloitte’s position in order to determine whether the information transmitted to it relates to ‘identifiable persons</u>’*” (Para. 97 of the Judgement). In this light, the General Court held that the sole alphanumeric codes received by Deloitte would not have allowed it to identify the complainants, as it had no access to

[58] Finck, M., Pallas, F. *They who must not be identified - distinguishing personal from non-personal data under the GDPR.*

[59] Along with the Court of Justice, the General Court is one of the EU’s courts making up the Court of Justice of the European Union. The purpose of these courts is to ensure a uniform interpretation and application of EU law. Decisions of the General Court can be appealed to the Court of Justice, but only on a point of law. Before the Lisbon Treaty came into force on 1 December 2009, it was known as the Court of First Instance.

[60] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC.

additional information necessary to re-identify the data subjects, which was under the exclusive control of the SRB.

Therefore, according to the Decision, the EDPS should have verified whether the data subjects were re-identifiable by Deloitte and not by the SRB, since the mere fact that the latter held further information which allowed (only the SRB) singling-out the data subjects was not sufficient to conclude that the data transmitted to Deloitte were of personal nature. For these reasons, the EGC upheld the SRB's plea and annulled the decision of the EDPS.

This Judgment has been warmly welcomed by the European privacy professionals' community, primarily due to its innovative take concerning the processing of pseudonymised data.

Indeed, the principles expressed by the EGC represent a considerable step forward compared to the hard-line stated by the WP29 in the abovementioned Opinion 05/2014 on anonymisation techniques, since the Judgment specifies that the risk assessment on re-identifiability must be carried out in concrete, i.e., on the account of the specific position of the recipient of the data, and not of all the parties involved. Which is to say that the objective and absolute approach stemming out of the WP29's traditional interpretation of anonymisation – which affected tons of decision by some Supervisory Authorities after 2014 – is not compatible with the more flexible criterion of 'reasonable re-identification' established by Recital 26 of the GDPR.

This change of course could lead to significant practical implications for the management of pseudonymised data sharing operations: if re-identification is practically unfeasible for an entity receiving this kind of data, because it is not provided with the additional information or other means necessary for – or in any event allowing – reidentification, then specifically and solely for such entity the pseudonymised information could not be considered to be personal data. Therefore, data protection legislation would not apply, and data would be *de facto* shared as anonymous data.

As a consequence, the data controller would not be required to fulfill several burdensome obligations, such as designating the third-party as data processor and, particularly in relation to the purposes which are relevant for DataTools4Heart, ensure transparency towards the patients and find an adequate legal ground for any possible re-use (i.e., secondary processing) of the data.

Moreover, in the case of transfer of pseudonymised data outside the European Economic Area (such as in Turkey), should the EU-based controller undertake a contractual obligation not to provide the receiving entity (unless some competent public competent or judicial authority require or order otherwise) with the means to re-identify the data (e.g., the Secret private pseudonymisation Key mentioned in §1.2.2), it will not be necessary for the controller to comply with the challenging rules on international data transfers set out in Chapter V of the GDPR.

Hopefully, too strict applications of data anonymisation by Supervisory Authorities – stemming from the interpretation offered by the WP29 almost 10 years ago – will be toned down in the light of this clear, up-to-date and more than commendable Decision by the General Court. This would greatly simplify the sharing of unidentifiable data, while still protecting both patients and their sensitive data by means of advanced pseudonymisation techniques, also in combination with novel PETs, triggering highly positive effects in the field of scientific and medical research.

Moreover – but not less importantly – this 'subjective' view of the concept of anonymous data[61] constitutes an first response, by the European institutions, to the increasingly frequent requests coming from the research environment all across member to set more standardised and uniform criteria based on which to determine whether data can considered anonymous or not.

However, some additional aspects need to be taken into account, as the Decision is neither a silver bullet for all circumstances, nor it is definitive from a strictly legal perspective.

Regarding the core principles affirmed by the EGC, it should be noted that the Judgement does not introduce broad presumptions or general rules for verifying the re-identifiability of a given pseudonymised dataset. Instead, the General Court ruled that it is necessary to carry out a case-by-case analysis, aimed to evaluate in detail the concrete position of the parties involved and, particularly, the means at their disposal to re-identify the data subjects. This means that the very same data which have undergone pseudonymisation techniques may prove to be anonymous for a party, while still reidentifiable for another, with totally different implications in terms of compliance.

On a purely judicial side, and so considering the legal effects of the case at stake, the Judgement is not yet final: the EDPS recently brought an appeal before the CJEU, alleging that the General Court misinterpreted the provisions regarding privacy obligations laid down by Regulation 2018/1725. The CJEU final ruling is going to have a huge impact for the regulation of all the situations in which the identifiability of data subjects is at issue in relation to pseudonymous data.

Furthermore, as already specified above, new EDPB Guidelines on pseudonymisation and – separately – anonymisation are expected for 2023/2024, and these will certainly take into account the case law related to the Decision T-557/20.

3.2 Practical legal implications in line with the vision of DataTools4Heart

A clear practical distinction exists in law between anonymisation and pseudonymisation, which goes beyond their conceptual differences in the GDPR.

Indeed, pseudonymisation is a security and data minimisation measure, thus not qualifying as a data processing itself.

On the contrary, anonymisation constitutes and implies a 'standalone' processing.

That being said, when it comes to compliance, practical and significant implications stem from the legal qualification of the de-identification process implemented by the data controller, or by a processor on the latter's behalf, in that:

i. transparency must be ensured only in regards of data processing activities as such, and not relating to the specific minimisation and security techniques which are applied, for instance, to prevent data breaches or enhance the confidentiality of the data;

[61] The concept of 'qualified anonimity', based on the idea that the outcomes of the assessment on re-identifiability risks must vary depending on the person or entity processing or receiving the data, was legally put forward for the first time and then detailed in MyHealthMyData, and EU-funded project in which many DataTools4Heart partners participated (PANETTA; Lynkeus; Athena and Siemens).

ii. lawfulness of the processing – which depends on whether (i) an appropriate legal basis exists among those established by art. 6 of the GDPR, and (ii) an additional condition is satisfied according to Art. 9.2 GDPR when special categories of data are involved – must be guaranteed only to anonymise and not to pseudonymise the data.

In concrete, point i. means that no processing carried out in order to make the data pseudonymous can be considered in breach of the applicable law, solely because the controller did not specifically inform the data subjects about the intention to pseudonymise the data (i.e., to secure them and enforce minimisation).

At the same time, point ii. entails the possibility for controller (or any of its processors) to apply pseudonymisation techniques at any time to both personal and more sensitive data, without the need to have a specific legal ground in place for this operation.

As already pointed out, the scenario is complicated by reason of the lack of coincidence between the legal definitions and the more technological (and mathematical) conceptions of these phenomena. While in the EU data protection legislation, the reversibility (or not) of the de-identification measures applied to the data is the key to classify the output as anonymous or pseudonymous data, the dividing line is much more blurred from a technical (e.g., IT, mathematical, machine learning) point of view.

Notwithstanding this, a fundamental indication is provided by the WP29 – in its already mentioned WP216[62] (§ 2.2.1) – which applies to data anonymisation and, even more so (given that "*Plus semper in se continet quod est minus*"), to pseudonymisation:

> «*[t]he Working Party considers that **anonymisation** as an instance of further processing of personal data **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 brief, the Opinion 05/2014 of the WP29 – which to date still represents the main legal guidance in the EU on data anonymisation (at least until the EDPB will adopt new guidelines)[63] – establishes a presumption of compatibility between the initial purposes for which the data were collected and the operations necessary to render them anonymous. To put it even more simply, controllers can rely on a pre-validation by the WP29 (namely now the EDPB) of their legitimate interest to undertake a secondary processing consisting in the anonymisation of the data they have collected over time. And this is true both for anonymising and pseudonymising the data.

On the account of this principle and with specific reference to DataTools4Heart, the combination of the pseudonymisation measures and advanced PETs described above in detail (as summarised in §2.3.1.1), allows stating that:

- ✓ **the patients' health data that clinical partners will process – or rather, that technical partners will process on behalf of the hospitals – for the purposes of DataTools4Heart project, will reach a degree of unintelligibility and security such as to reasonably**

[62] See Note no. 50.
[63] Refer to Note no. 20.

prevent any further reidentification of the data subjects, except by the hospitals themselves, in case a legitimate need arises or a request by a competent authority is issued;

- ✓ **provided that sufficient transparency was ensured at the time of data collection (or later, by any admitted informative means as per each hospital's own privacy practices and policies) vìs-a-vìs the patients regarding the possibility to make their data no longer identifiable for any permitted purpose (e.g., for reusing them for scientific research), the clinical partners can rely on their pre-verified legitimate interest in order to implement data anonymisation**.

4 Primary and secondary processing of health data

On the account of the higher risks and more harmful impacts that may be determined in any case of misuse, and even more of a breach, of special categories of data – including (i) data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership; (ii) genetic data; (iii) biometric data for the purpose of uniquely identifying a natural person; (iv) data concerning health; (v) data concerning a natural person's sex life or sexual orientation – the GDPR explicitly prohibits the processing of this type of data, except when one of the exemptions provided for by Art. 9.2 can apply[64].

Many of the conditions which permit the collection and processing of sensitive data relate to the health sector – both for primary (i.e., to provide the patients with healthcare services) and secondary use (i.e., to support the safe and efficient functioning of national healthcare systems, as well as to drive health research and innovation) – such as when the processing is necessary for:

i. Patient care: for the purposes of the provision of health or social care or treatment or the management of health or social care systems and services on the basis of Union or Member State law or by or under the responsibility of a health professional subject to the obligation of professional secrecy;

ii. Public health: 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, on the basis of Union or Member State law which provides for suitable and specific measures to safeguard the rights and freedoms of the data subject;

[64] The following definitions are set out in Art. 4 GDPR: 'data concerning health' means "*personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status*"; 'genetic data' means "*personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question*"; 'biometric data' means "*personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data*".

iii. Scientific research: for medical and clinical research purposes, provided that appropriate technical and organisational measures are adopted to ensure data minimisation, such as pseudonymisation, "*based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject*" (Art. 9.2, j)).

The term 'secondary use' is not found in the GDPR, but it is to be understood as being broadly in line with the term 'further processing' of data as described in the purpose limitation principle set out in Art. 5(1)(b) of such Regulation. This crucial provision states that processing data for a purpose different to that specified at the time when the data were collected (namely in the privacy notice given to the patients) must not be allowed when this is incompatible with the original purpose, unless such further processing is for (*inter alia*) research purposes and is undertaken in accordance with safeguards described in Art. 89(1) GDPR, i.e., ensuring that data minimisation is in place.

In short, a genuine presumption of compatibility is laid down in the GDPR between the primary and secondary processing of the data, as long as the reuse takes place for scientific research objectives and under appropriate security and minimisation measures, including pseudonymisation. Then, should the research project also imply the (re)processing of data falling into special categories, such as health ones, the existence of an EU or national law permitting the reuse must be ascertained, according to Art. 9.2(j) GDPR.

Unfortunately, the intention of the EU legislator to foster scientific research in the GDPR, by establishing simplified procedures and lighter requirements in connection with the re-use of data, has been in many cases frustrated at national level, since member States are empowered to "*introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health*" (Art. 9.4 GDPR).

This inevitably led – notwithstanding the Recital 53 GDPR states that the possibility for member States to establish stricter requirements for processing health data should not hamper the free flow of personal data within the Union, when those conditions apply to cross-border processing of such data – to a loss of homogeneity in the way the EU legislation on the circulation and protection of health data was integrated at local level, "*resulting in a complex and fragmented landscape for researchers to navigate. Consequently, differences between Member States in the way the GDPR is implemented and interpreted in the area of scientific research has made data exchange between Member State and EU bodies for research purposes difficult and in some cases highly technical*" (so-called '**Nivel Study**')[65].

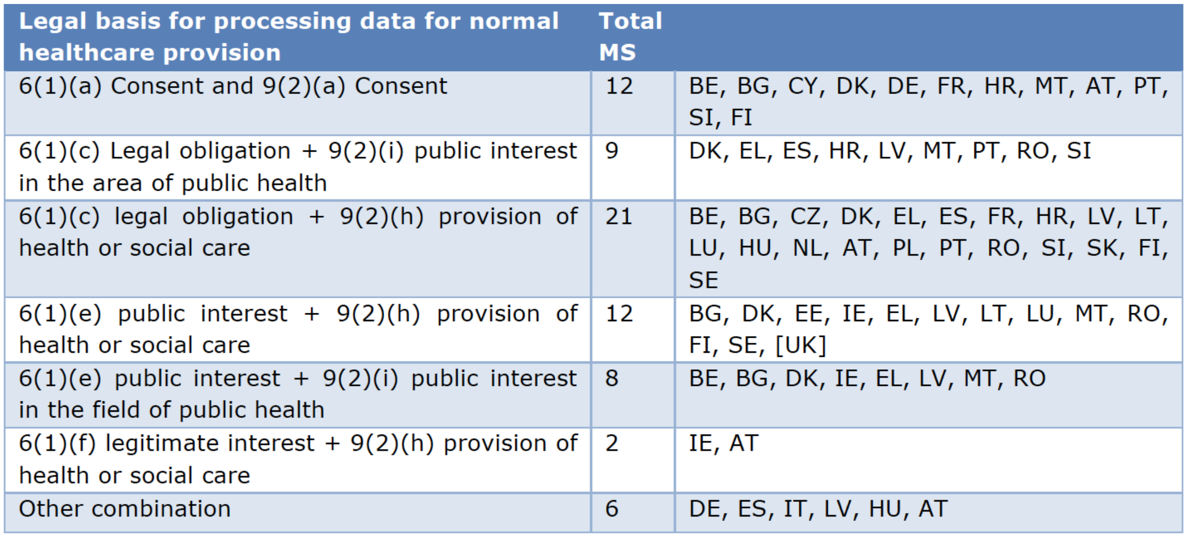
In practice, this means that sharing and re-using health data for cross-border scientific or medical research project poses to date serious risks of non-compliance, as the obligations and regulatory

[65] *Assessment of the EU Member States' rules on health data in the light of GDPR*, by the EU Commission's Health and Food Safety Directorate-General, dated 12 February 2021.

procedures to be abided (e.g., vìs-a-vìs the Ethics Committees and/or competent Supervisory Authorities) greatly differ from a member State to another.

This stokes fear of violations and sanctions by controllers and thus concretely prevents scientific progresses, as well as the implementation of the European Health Data Space and, in the end, the creation of the coveted European Research Area[66].

A quick look at the *Figure* below (no. 11) is more than sufficient to give an idea of the degree of fragmentation which today affects the health sector when it comes to data protection requirements.



Legal basis for processing data for normal healthcare provision	Total MS	
6(1)(a) Consent and 9(2)(a) Consent	12	BE, BG, CY, DK, DE, FR, HR, MT, AT, PT, SI, FI
6(1)(c) Legal obligation + 9(2)(i) public interest in the area of public health	9	DK, EL, ES, HR, LV, MT, PT, RO, SI
6(1)(c) legal obligation + 9(2)(h) provision of health or social care	21	BE, BG, CZ, DK, EL, ES, FR, HR, LV, LT, LU, HU, NL, AT, PL, PT, RO, SI, SK, FI, SE
6(1)(e) public interest + 9(2)(h) provision of health or social care	12	BG, DK, EE, IE, EL, LV, LT, LU, MT, RO, FI, SE, [UK]
6(1)(e) public interest + 9(2)(i) public interest in the field of public health	8	BE, BG, DK, IE, EL, LV, MT, RO
6(1)(f) legitimate interest + 9(2)(h) provision of health or social care	2	IE, AT
Other combination	6	DE, ES, IT, LV, HU, AT

Figure 11 - Legal basis for normal healthcare provision

This hurdle is even more critical in relation to the re-use of health data for research purposes, because articulated evaluation procedures and, often, obligations of prior authorisation by Supervisory Authorities are added to the patchwork of legal grounds and specific processing conditions applicable at national level.

With the aim of getting an up-to-date picture of the national regulatory scenarios in this field, a detailed questionnaire was prepared and circulated, within WP1, to a project's Consortium sub-group focused on ELSI (Ethical, Legal and Social Issues) requirements. The scope of this action was obtaining, from clinical partners' Data Protection or Privacy Officers (or Legal departments), indications of the main local obligations they must comply with and hinders to be overcome in order to lawfully reuse health data for research purposes.

Some of the main findings of this survey are summarised in the table below:

[66] BBMRI-ERIC pointed out that "*The existing regulatory framework seems insufficient to deliver on the promises of the EHDS. Health data governance remains fragmented at national and regional level, hindering any effort to scale up research and healthcare solutions. Most importantly, it is necessary to protect and promote the use of health data, defining clear pan-European rules to overcome the existing gaps in practice*" (link).

Table 1 - Main findings of the DataTools4Heart ELSI questionnaire with clinical partners

Country	Clinical partner	Further conditions or limitations established by national law for reusing health data for scientific research (Q4)	Description of provided limitations/conditions and legislative reference (Q7)	Data sharing with third parties operating in the research field (Q6)	Prior authorisation from national competent supervisory Authority (Q8)	Transparency and lawfulness of health data reuse for research purposes towards the patients at the time of data collection (Q9)
Romania	**BUCH**	No, only GDPR requirements apply	None	Yes, the reuse also includes sharing the data with third parties.	No	Yes. The privacy notice provided to data subjects (patients) at the time when the data were collected includes information regarding the use of their data for scientific and medical research/studies. Article 6.1. (e) or (f), combined with the derogations adopted under Article 9.2. (j) or (i) of the GDPR, provides a legal basis for the processing of personal medical data for the purpose of scientific research.
UK	**UCL**	Yes	Research studies involving patients require ethics approval by a Health Research Authority (HRA) Ethics committee.	Yes, subject to HRA ethics approval, as long as appropriate data sharing agreements are	No. The HRA ethics approval is sufficient	Yes. Privacy notice for UCLH patients on hospital website. Privacy notice for participants

			In the case of research databases or re-use of existing de-identified clinical data, the HRA approval can be at a database-wide level, where the organisation curating the research database needs to put in place a governance process to approve individual studies. This devolved process is in place at UCLH, which has overarching research database approval. Identifiable data used without consent for research requires approval according to Regulation 5 (section 251) of the Health and Social Care Act 2012. This requires approval from the HRA Confidentiality Advisory Group (England and Wales), the Patient and Public Benefit Panel (Scotland) or the Privacy Advisory Committee (Northern Ireland).	in place with the third party.		in UCL-sponsored studies on UCL website.
Netherlands	**AMC** **UMCU**	Yes	Data can be accessed when required for the proposed research while complying to minimisation measures (and pseudonymisation is the most important one). Specifically, the Dutch Medical treatment	For data sharing purposes, all data can be shared, if the data is definitely required for the project. This should be substantiated by study protocols and proposed	No, only ethical boards' approval	**AMC:** Yes. Data reuse for medical scientific research purposes is mentioned in the informative notice given to all patients when they first visit the outpatient clinic or get

			Contracts Act (WGBO) regulates the relation between patients and care providers and stipulates that those involved in medical scientific research have to be adequately informed about the research and must give their legally valid permission for their data to be accessed. Upon collaboration with commercial partners, additional rules/requirements may apply. Research is carried out based on the no-objection rule for large amounts of data, according to the WGBO (article 7:458). Patients can object at any moment to the re-use of their data and, in that case, they will no longer be included in the research projects	research plans. Additionally, necessary agreements on data sharing and transfer should be established. In some cases, data can be shared with third parties, but this will have to be assessed in a project-to-project manner together with both the legal and privacy department. In the case of sharing with a third party, appropriate agreements should be established between the participating parties (either data processing agreements or data transfer agreements)		admitted for the first time. Additionally, the same information notice is also published in AMC website. In short, clinical data can be reused for research purposes, but if patients do not want this, they can indicate this to their attending physician and data will not be re-used for research purposes. The no-objection check is carried out when data is collected. Specifically: data is retrospective and encoded, no informed consent will be sought for participation in the study, according to the no-objection rule. On a project-to-project basis, it will be assessed which data is required for the proposed research and based on this description, data will be delivered. The ‘no-objection’-check will be performed by RDM at time of data extraction, to fulfill patients’ wishes when not wanting to participate in the research. The researcher receives the dataset on the same date as

						when the objection check is performed by RDM. **UMCU:** Both transparency and lawfulness requirements are properly complied with by (i) providing the patients with a privacy notice which specifies, inter alia, that their data may be reused for scientific and medical research; and (ii) identifying an appropriate legal basis (Art. 6.1 GDPR) and a valid condition which legitimises the secondary processing (Art. 9.2 GDPR)
Sweden	**KUH**	Yes	The Patient Data Act (*Patientdatalag* (2008:355)) governs processing of patients' personal data in the healthcare sector. With regards to the secondary use, the purpose limitation principle applies as described in the GDPR, meaning that patients' health data can undergo a secondary processing provided that the new purpose is compatible with the one for which the data were originally collected. As to the	Consent of the patient must be obtain to share the data, unless prior anonymisation is applied	No. Only ethics approval is necessary	Yes (i.e., specific information was given to the patients and an appropriate legal ground was indicated)

			legal ground permitting the reuse, the opt-out shall apply. Furthermore, the Act concerning Ethical Review of Research Involving Humans (*2003:460*) applies to research that includes the processing of sensitive personal data, so requiring an authorisation from the Ethical Review Authority.			
Italy	**GEM**	Yes	Art. 110 (titled 'Medical, Biomedical and Epidemiological Research') of the Italian Privacy Code (Legislative Decree no. 196/2003) reads that: "*The data subject's consent shall not be required to process data relating to health for scientific research purposes in the medical, bio-medical or epidemiological sectors if the said research is carried out in accordance with laws or regulations or EU law pursuant to Article 9(2), letter j), of the Regulation (…) and if a data protection impact assessment is carried out and published in accordance with Articles 35 and 36 of the Regulation. Additionally, consent shall not be necessary if informing the data subjects*	Art. 110-*bis* (titled 'Further processing of personal data by third parties for scientific research or statistical purposes') states that: "*The Garante may authorise further processing of personal data, including the special categories of personal data referred to in Article 9 of the Regulation, for scientific research purposes or statistical purposes by third parties that carry out such activities to a prevailing extent if informing the data subjects proves impossible or entails a*	Yes, in the cases set out by Articles 110 and 110-bis of the Italian Privacy Code described in the two left columns. In addition, the Ethics Board approval is needed.	Yes, both the obligations were fulfilled when data were collected.

			proves impossible or entails a disproportionate effort on specific grounds, or if it is likely to render impossible or seriously impair the achievement of the research purposes. In such cases, the controller shall take appropriate measures to protect the rights, freedoms and legitimate interests of the data subjects and the research programme shall be the subject of a reasoned, favourable opinion by the geographically competent ethics committee as well as being submitted to the Garante for prior consultation in accordance with Article 36 of the Regulation”. Evidently, a prior green light by the Italian Supervisory Authority would be needed in this latter case.	*disproportionate effort on specific grounds, or if it is likely to render impossible or seriously impair the achievement of the research purposes. In such cases, the controller shall take appropriate measures to protect the rights, freedoms and legitimate interests of the data subjects in accordance with Article 89 of the Regulation including arrangements for the prior minimisation and anonymisation of the data. (…) Processing for scientific purposes of the personal data collected in the course of clinical activities by public and private* Istituti di ricovero e cura a carattere scientifico *shall not be an instance of further processing by third parties on account of the instrumental nature of the health care activities carried out by such Istituti vis-à-vis research activities, subject to*		

				compliance with Article 89 of the Regulation”.		
Czech Republic	**FNUSA**	No	Currently, there is no legal framework for the easy secondary use of medical data as stated. Art. 16 para 1 of Act No. 110/2019 Coll., specifies which safeguards referred to in Art. 89 of GDPR must be adopted when processing health data for scientific research. Act No. 372/2011 Coll., on health services and the conditions for their provision – which does not regulate the conditions for secondary use for research – affects the data kept in the patient's medical documentation. An amendment to the law is currently being discussed, which would make this possible. However, its adoption cannot be expected early. Currently, only data from patients who have consented to such use can be used for scientific and research purposes	Patient’s consent is needed in order to lawfully share his/her personal health data with third parties for research purposes	No	Yes. Informed patient consent is currently the only reliable way to use such patient data for research. However practice is not uniform in the situation of the absence of a specific legal framework on research and science in the Czech Republic and some data controllers rely on other legal titles (e. g. public interest) and general exemptions from GDPR.

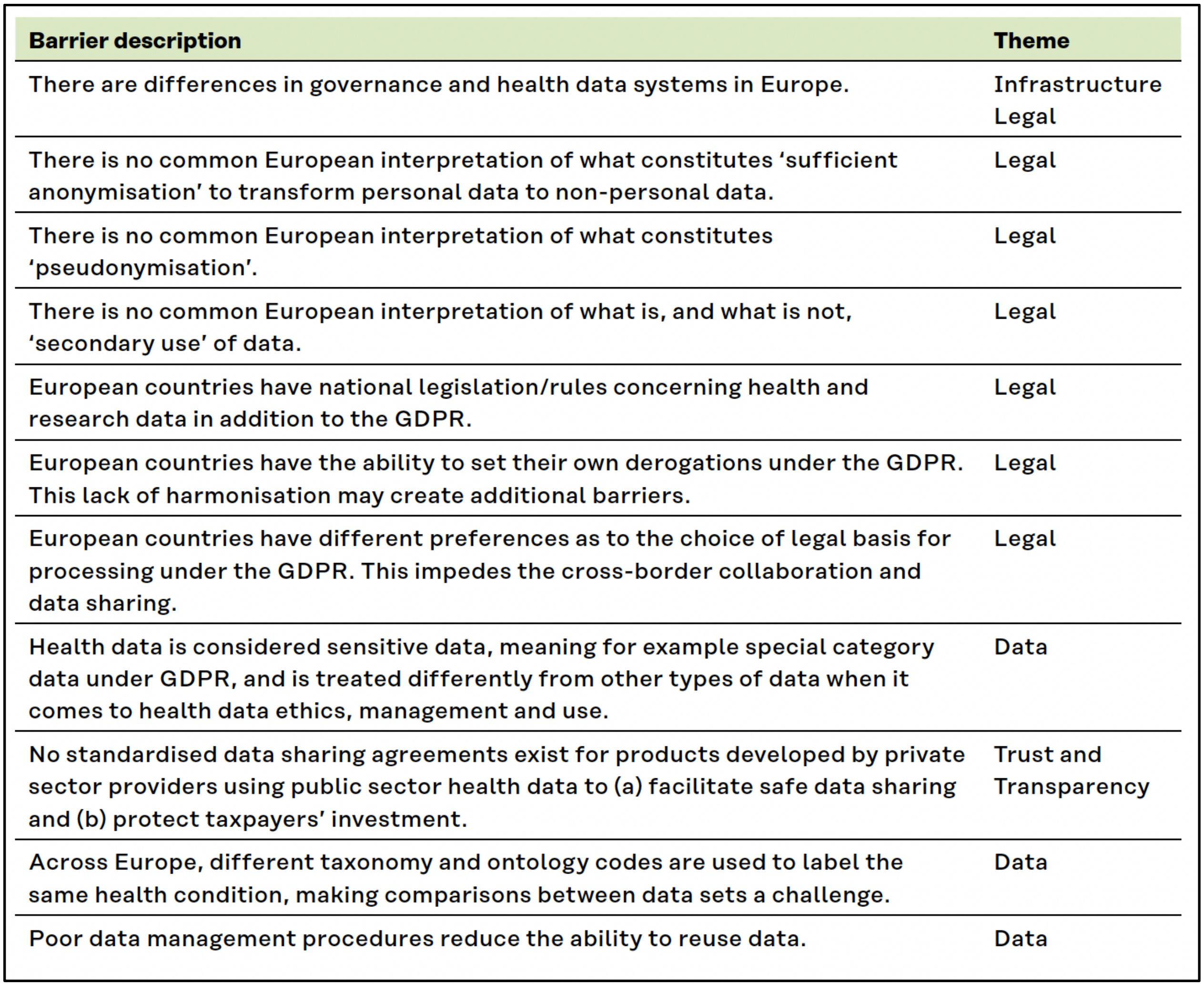
Spain	**VHIR**	No (more favourable conditions)	*Ley Organica* 3/2018 (*Disposiciòn adicional decimoséptima*: "*Tratamientos de datos de salud*") provides that: “The reuse of personal data for health and biomedical research purposes will be considered lawful and compatible when, having obtained the data subject’s consent for a specific purpose, the data are used for purposes or areas of research related to the area in which the initial study was scientifically integrated. In such cases, those who are responsible must publish the information notice (…) in an easily accessible place on the corporate website of the center where the research or clinical study is carried out, and, where appropriate, on that of the sponsor, and notify the existence of this information by electronic means to the data subjects. When the latter lack the means to access such information, they may request its submission in another format. For the processing provided for in this letter, a prior favorable report from the research ethics committee will be required. The	Yes	No	Yes VHIR has not adhered to date to the Code of Conduct Regulating the Processing of Personal Data in Clinical Trials and Other Clinical Research and Pharmacovigilance Activities (link), but still follows its rules

			use of pseudonymised personal data for health and, in particular, biomedical research purposes, is considered lawful. The use of pseudonymised personal data for public health and biomedical research purposes will require: 1. technical and functional separation between the research team and those who carry out the pseudonymisation and preserve the information that enables re-identification. 2. That the pseudonymised data are only accessible to the research team when: i) There is an express commitment to confidentiality and not to carry out any re-identification activity. ii) Specific security measures are adopted to prevent re-identification and access by unauthorised third parties. The re-identification of the data at its origin may be carried out when, due to an investigation that uses pseudonymised data, the existence of a real and specific danger to the safety or health of a person or group of people, or a serious threat for their rights, are proven, or when it is necessary to guarantee adequate health care.			

The findings above ruthlessly capture the current legal scenario on data reuse in the EU.

As repeatedly underlined by TEDHAS in its reports, including the last one, issued on 14 September 2023: "*Different national governance systems, lack of standardisation of data sets and variations in legal interpretations of EU data protection law*" are only some examples of "*the most common barriers that make transnational studies difficult and increase the costs of research and compliance*"[67].

More generally, most of the hurdles that to date make the implementation of the European Health Data Space much more complex to achieve are of legal nature.



Barrier description	Theme
There are differences in governance and health data systems in Europe.	Infrastructure Legal
There is no common European interpretation of what constitutes 'sufficient anonymisation' to transform personal data to non-personal data.	Legal
There is no common European interpretation of what constitutes 'pseudonymisation'.	Legal
There is no common European interpretation of what is, and what is not, 'secondary use' of data.	Legal
European countries have national legislation/rules concerning health and research data in addition to the GDPR.	Legal
European countries have the ability to set their own derogations under the GDPR. This lack of harmonisation may create additional barriers.	Legal
European countries have different preferences as to the choice of legal basis for processing under the GDPR. This impedes the cross-border collaboration and data sharing.	Legal
Health data is considered sensitive data, meaning for example special category data under GDPR, and is treated differently from other types of data when it comes to health data ethics, management and use.	Data
No standardised data sharing agreements exist for products developed by private sector providers using public sector health data to (a) facilitate safe data sharing and (b) protect taxpayers' investment.	Trust and Transparency
Across Europe, different taxonomy and ontology codes are used to label the same health condition, making comparisons between data sets a challenge.	Data
Poor data management procedures reduce the ability to reuse data.	Data

Figure 12 – Barriers to cross-border data sharing and reuse of health data identified by TEHDAS (see Note 67)

[67] '*EU-wide collaboration needed to optimise health data use for research and innovation*', available here (link).

4.1 Achieving uniform legal compliance in DataTools4Heart

The enormous potential of Privacy Enhancing Technologies – whatever the definition that one prefers to attribute to these hybrid legal/IT measures – is still almost completely unexplored by the applicable legislation.

The regulatory framework is almost stuck in April 2014[68] and the progress of scientific research area still trapped in the grip of the super-rigid interpretation of anonymisation first offered by the WP29 and then embraced by some national Supervisory Authorities.

DataTools4Heart Consortium designed a technical and legal architecture to reverse this paradigm.

A. Pseudonymous data

First of all, as the project aims to enable clinical partners to easily and lawfully reuse all of their existing EHR data, in any format, unstructured ones will be converted to structured format by means of Natural Language Processing tools, which will be designed for all data controllers' 7 languages (English, Spanish, Italian, Romanian, Czech, Swedish and Dutch), specifically within the cardiology domain. This will ensure accuracy and precision of the data used in the project.

Then, the resulting structured data along with other structured information from the EHR will be captured in the Common Data Model[69], which also ensures proper implementation of FAIR (Finable, Accessible, Interoperable and Reusable) principles. During the ingestion/mapping by relevant Data Ingestion Suite, all patients' identifiable data are pseudonymised, while the association table is kept in a secured folder which will be only accessible to each clinical partner's divisions data manager and the physician with care relation to the patient. Furthermore, in case data relating to a single patient should come from different data sources within the same hospital, then such data will be linked using solely a pseudonymous code by the assigned data manager, so that no identifiable data will be stored within the database.

All the IT transmission modules between the hospitals and the technical partner developing the Data Ingestion Suite (associated with the Common Data Model and the data Feature Extraction Tool), are configured to operate with the highest level of cybersecurity, via the TLS-protected communication channel aligned with the IETF (Internet Engineering Task Force) Best Current Practice.

Moreover:

- access to FHIR-based Health Data Repository is enabled only within the DT4H virtual docker network, so preventing any unauthorised party from having visibility of the data;
- each data mapping activity is logged;

[68] When the *Opinion 05/2014 on Anonymisation Techniques* was adopted by the WP29.

[69] The structured and unstructured data curated the CDM include: clinical presentation (symptoms, co-morbidities; ECG data measurements; imaging (echocardiography, cardiac MRI) data measurements; cardiovascular risk factors; laboratory measurements; diagnoses; medications; clinical notes, discharge letters, referral letters, procedural reports (i.e., surgical, catheterisation, mapping); clinical outcomes (e.g., procedures, death, major adverse cardiac events).

- Health Data Repository implements the suggested best practice from HL7 FHIR standard for audit logging; for any FHIR interaction (search, CRUD, FHIR operation), a detailed AuditEvent record (when, who, which resource(s), query details, etc.) is generated and stored within the Health Data Repository;
- each dataset extraction activity will be logged;
- each interaction with the federated querying platform activity will be logged.

In light of the principles enshrined in the Judgement (case T-557/20) detailed in §3.1 – as contextualised to DataTools4Heart in §3.2 – and given the combination of health-domain specific state-of-the-art data security measures and legal solutions implementing data minimisation and privacy-by-design, with special reference to Federated Learning which enables machine learning application to run within the clinical partners' nodes, without any personal data having to leave the controller's local repository, the risks of patients reidentification can to date be reasonably excluded.

Furthermore, leveraging on the same subjective interpretation of data anonymisation, given that clinical partners will be responsible to share in no case with third parties, even when acting as data processors, the mapping table or the secret pseudonymisation key, it can be affirmed that the resulting data – which remains re-identifiable solely for the hospitals – will prove as reasonably anonymous for any other entity receiving them.

This also means that, legally, there is no reuse of these data, as they can no longer be used by other clinical partners or by data processors to single-out an individual, falling outside the scope of the GDPR.

Notwithstanding the soundness of the rationale underpinning this approach, all the partners of the project must be aware that, in some jurisdiction, there is an extremely remote risk that this could still be challenged by Data Protection Authority, in case the initial consent of the patients for the reuse of their data for scientific research should lack.

B. Synthetic data

The legal status of synthetic data in general and specifically in DataTools4Heart has already been detailed in §2.3.1.1 and 2.3.2.

The project is a testbed for an innovative integration of security layers aimed to protect both the data and the patients and PETs which help implementing robust forms of data minimisation, while ensuring accountability for the clinical partners (as well as for all parties acting on their behalf).

Synthetic data will be produced through federated and differentially private generative models (including *Bayesian Networks* and *Generative Adversarial Networks*) which allow learning the data distribution within the clinical partners' datasets, which are then sampled to produce AI-driven fully made-up unidentifiable data.

Federated Learning avoids any data sharing, because the ML models are trained locally at each clinical site and only model characteristics (e.g. parameters, gradients) are transferred to a centralised location for iterative aggregation. This allows attaining high performance models without the data ever leaving the hospital, so offering a concrete paradigm of minimisation under the GDPR.

In addition, to further guarantee the privacy of both the patients' data during computation and the output of the learned model, Secure Multi-Party Computation will be deployed to encrypt the shared model parameters, alongside Differential Privacy to hide the participation of an individual in a specific training task, based on different 'privacy budget' (*epsilon*) values. The combination of FL, SMPC and DP strengthens data protection guarantees during both the computation and use of the generated AI model, offering the highest level of accountability for all parties involved in the process.

In sum, synthetic data generation will be carried out:

- ✓ through Federated Learning, with a view to minimising both the data processed during computation taking place at local level and the risk arising from data breaches;
- ✓ enforcing privacy guarantees through appropriate Differential Privacy mechanisms (Laplace, Gaussian, or Exponential, sparse-vector technique, etc.);
- ✓ pondering the differences between probabilistic graphical models (e.g., Bayes/Markov nets) and deep learning models (e.g., GANs)
- ✓ through standard techniques (Markov Chain Monte Carlo, generative network forward computation);
- ✓ implementing other forms of algorithmic privacy during various steps of the training process (such as gradient or parameter estimation and structure learning).

Moreover, data synthesis will also be explored for improving real-data utility, e.g., by filling in missing data entries, balancing bias and discrimination and augmenting datasets with the aim of adding variety in training AI models and reducing model 'brittleness'.

Following on from the novel rules of the Artificial Intelligence Act, where synthetic data are explicitly considered as equivalent to anonymous (or other non-personal) data, DataTools4Heart aims to set an industry standard for this technique in the clinical sector, testing a combination of some of the most innovative Privacy-Enhancing Techniques to: (i) collaboratively train AI models across all the hospitals without the need to exchange any local data; (ii) secure the parameters computed by each data controller to ensure they do not reveal any of their initial inputs (personal data); (iii) to make the data no longer intelligible and so enhance security and individual privacy.

From a legal standpoint, this practically means that all clinical partners are entitled to rely on the presumption of compatibility laid down by Art. 5.1(b) GDPR (see §3.2), thus considering anonymisation as compatible-by-default with the original purpose for which they collected the data, at the sole condition that transparency was ensured at that moment towards the patients regarding the possible reuse of their data.