HEALTHYCLOUD
Health Research & Innovation Cloud

Glossary of commonly used terms in the field of health data research

Change Log

Version	Author	Date	Description of Change
V0.1	Irene Kesisoglou	07/02/2022	First published version
V1.0	Named authors	01/07/2022	Second published version

Notice: *The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement №965345*

1. Introduction

This glossary was created by the consortium partners of the European project HealthyCloud with the aim of establishing agreed definitions of commonly used terms to harmonise the work of the project. The aim is that the glossary can also be used by other projects working in similar areas. The glossary was created through a series of glossary working group calls, during which each definition was discussed and agreed upon. This is a living document that will be updated regularly with the addition of new terms and modifications, if needed, of the terms and definitions currently there.

This is the second version of the HealthyCloud glossary, in which new terms have been added based on a need perceived in the glossary working group. Some existing terms have been modified to align with the recently published Proposal for a Regulation on the European Health Data Space (EHDS). In addition, sections have been created for terms relating to specific topics (e.g., artificial intelligence, profiles as defined in HealthyCloud).

2. Glossary

Aggregated data:

Aggregated data is pooled data. Statistical data about several individuals that have been combined to show general trends or values within the data.[1] Aggregated data are not necessarily anonymised data.

Anonymisation:

The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject. Removing personally identifiable information, so as to definitively not allow the identification of the data subjects.[2] The methods used to anonymise the data are context dependent.

Biometric data:

Personal data resulting from specific technical processing relating to the physical, physiological or behavioral characteristics of a natural person, which allow or

[1] EMA, External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use, October 2018. Available at: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-3.pdf

[2] Par. 11 Proposal for a Regulation on European data governance (Data Governance Act), November 2020. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

confirm the unique identification of that natural person, such as facial images or fingerprint data.[3] [4]

Cloud:

Network of computing facilities providing remote data storage and processing services through the internet.[5]

Cloud computing:

Paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with administration on-demand.[6]

Consent:

An individual's agreement e.g. to participate in research, undergo a healthcare procedure, to personal data processing.

Within the context of personal data, the GDPR defines consent as: Any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.[7]

Data:

1. Data can be defined as the recorded factual material that is commonly accepted in the scientific community as information that is required to support research findings.[8]

[3] Art. 4(14) GDPR, Regulation (EU) 2016/679 of the european Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

[4] Proposal for a Regulation laying down harmonised rules on artificial intelligence. https://ec.europa.eu/newsroom/dae/items/709090

[5] Lexico.com / Oxford University Press

[6] ISO. Information technology, Cloud computing, Overview and vocabulary. Available at: https://www.iso.org/obp/ui/#iso:std:iso-iec:17788:ed-1:v1:en

[7] Art. 4(11) GDPR, Regulation (EU) 2016/679 of the european Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

[8] Based on https://nnlm.gov/data/thesaurus/data

2. Refers to any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording.[9]
3. There are four major categorical types of data for where the data comes from: observational; experimental; simulated and derived.[10]
4. Data is information available for processing.

Data discoverability:

The ability or a mechanism to browse and locate available data relevant to a specific user's purpose (e.g., research project) in a non-targeted search. Data is more discoverable if the data collection has metadata and the metadata is publicly accessible. Discoverability is related to findability from the FAIR principles.

Data access:

The processing of data by a data user, which was provided by a data holder, in accordance with specific technical, legal, or organisational requirements, without necessarily implying the transmission or downloading of such data[11].

Data access right: the ability, right or permission to act on data in a defined location.[12]

Data altruism:

Consent by data subjects to process personal data pertaining to them, or permissions of other data holders to allow the use of their non-personal data without seeking a reward, for purposes of general interest, such as scientific research purposes or improving public services.[13]

Data centric health research computational infrastructure:

This infrastructure provides data as a service. This infrastructure includes services, such as data visualisation, hosting and processing of data. In particular, it can

[9] Proposal for a regulation of the European Parliament and of the council on European data governance. 2020/0340. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

[10] Based on https://nnlm.gov/data/thesaurus/data

[11] Proposal for a regulation of the European Parliament and of the council on European data governance. 2020/0340. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

[12] From the Beyond 1 Million Genomes project.

[13] Proposal for a regulation of the European Parliament and of the council on European data governance. 2020/0340. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

process health-related sensitive data. Technological infrastructures for data analysis, exploitation and/or processing.

Data controller:

Under Regulation (EU) 2018/1725, as well as under the GDPR, the data controller is the party that, alone or jointly with others, determines the purposes and means of the processing of personal data. The actual processing may be delegated to another party, called the data processor. The controller is responsible for the lawfulness of the processing, for the protection of the data, and respecting the rights of the data subject. The controller is also the entity that receives requests from data subjects to exercise their rights.[14] [15]

Data governance:

Assembly of policies and processes, coordination aspects, data usage and accessibility principles and data management procedures for a certain health data infrastructure to ensure legal compliance, consistency and good data quality throughout the different stages of the data life cycle.

Data processor:

According to Article 3 (12) of Regulation (EU) 2018/1725, a processor shall mean "a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller." The essential element is therefore that the processor only acts "on behalf of the controller" and thus only subject to his instructions.[16]

In some cases, the processor may choose not to process the data himself, but may have recourse to a subcontractor who processes the data on his behalf. In practice, this will depend upon the processor agreement entered into with the controller.

Data provider/holder:

Any natural or legal person, which is an entity or a body in the health or care sector, or performing research in relation to these sectors, as well as European Union institutions, bodies, offices and agencies who has the right or obligation, or the

[14] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1725&from=EN

[15] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

[16] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1725&from=EN

ability to make available, including to register, provide, restrict access or exchange certain data.[17]

Data quality:

The degree to which a set of inherent characteristics of data fulfills requirements.[18]

Notes: The requirements are defined by the purpose of the processing and hence data quality can be viewed in other words also as a "fitness for purpose". The purpose can be any use of the data, including primary use or secondary use.

For the purpose of data protection, data quality refers to a set of principles laid down in Article 5 of the GDPR and Article 4 of Regulation (EU) 2018/1725, namely[19]:

- Lawfulness, fairness and transparency
- Purpose limitation
- Data minimisation
- Accuracy
- Storage limitation
- Integrity and confidentiality

Data sharing:

Provision of data by a data controller to a data user for the purpose of joint or individual use of the shared data, based on conditions of use, directly or through an intermediary.[20]

Data subject:

As defined in the GDPR, a data subject is a person who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.[21]

Data user:

[17] Proposal for the European Health Data Space: https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF

[18] https://www.iso.org/obp/ui/#iso:std:iso:8000:-2:ed-4:v1:en

[19] https://edps.europa.eu/data-protection/data-protection/glossary/d_en#data_quality

[20] Proposal for a regulation of the European Parliament and of the Council on European data governance. 2020/0340. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

[21] Art. 4(1) GDPR, Regulation (EU) 2016/679 of the european Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

A natural or legal person/organisation who has lawful access to certain personal or non-personal data and is authorised to use that data for commercial or noncommercial purposes.[22]

Dataset:

Collection of data that is represented in a particular form. Datasets will vary depending upon the type of intended use, and how the collecting organization has decided to organize their data upon collection. Dataset is essentially a heterogeneous term that could be made up of any type of collection for any type of data.[23]

Dataset catalogue:

A collection of datasets descriptions, which is arranged in a systematic manner and consists of a user-oriented public part, where information concerning individual dataset parameters is accessible by electronic means through an online portal.[24]

FAIR Principles:

Principles to define the Findability, Accessibility, Interoperability, and Reuse of resources for humans and computers at the source. For example, the principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.[25]

- Findable: Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.
- Accessible: Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository
- Interoperable: Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

[22] Proposal for a regulation of the European Parliament and of the council on European data governance. 2020/0340. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

[23] https://meshb.nlm.nih.gov/record/ui?ui=D064886

[24] Proposal for the European Health Data Space: https://eur-lex.europa.eu/resource.html?uri=cellar:dbfd8974-cb79-11ec-b6f4-01aa75ed71a1.0001.02/DOC_1&format=PDF

[25] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

- Re-usable: Data and collections have a clear usage licenses and provide accurate information on provenance.[26]

Filing system:

Any structured set of personal and non-personal data which are accessible according to specific criteria, whether centralised, decentralised or dispersed on a functional or geographical basis.[27]

Genetic Data:

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.[28]

Health data:

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.[29]

Health data collection:

A technical infrastructure that holds datasets, makes datasets available for use, and organises data in a logical manner. The datasets may come from different sources, hospitals and/or research institutes from the same country (national data repositories) or different countries (international data repositories). Data collections may also cover appropriate, subject-specific locations where researchers can submit their data. Data collections may have specific requirements concerning subject or research domain; data reuse and access; file format and data structure; and the types of metadata that can be used.[30]

Minimal inclusion criteria:

1. A digital platform that receives and stores data

[26] https://upload.wikimedia.org/wikipedia/commons/b/b7/Implementing_FAIR_Data_Principles_-_The_Role_of_Libraries.pdf

[27] Art. 4 (6) GDPR, Regulation (EU) 2016/679 of the european Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

[28] Art. 4(13) GDPR, Regulation (EU) 2016/679 of the european Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

[29] Art. 4(15) GDPR, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

[30] Based on https://nnlm.gov/data/thesaurus/data-repository#:~:text=A%20data%20repository%20can%20be,data%20in%20a%20logical%20manner.&text=Data%20repositories%20may%20have%20specific,metadata%20that%20can%20be%20used.

2. It receives data from a single source and/or multiple sources
3. Allows discovery of the stored health data
4. It must have control over the data stored

Other possible characteristics of a data collection:

5. It could have a specific thematic, data type that it collects (e.g. a particular disease, a particular data type: genomic data, clinical data, EHRs…)
6. It could be part of one or more overarching data hubs
7. It could generate data

Health data hub:

Minimal inclusion criteria:

1. A digital technical infrastructure with the core mission of enabling health data sharing
2. It provides health data from different sources
3. It allows discovery of health datasets
4. It has a metadata discovery service
5. It has a data accessibility mechanism in accordance with existing regulation
6. It has an authorization functionality, provided by the same Data Hub or by an external institution.

Health information:

All organised and contextualised data on population health and health service activities and performance, individual or aggregated, that improves health promotion, prevention, care, cure and policy-making.[31]

Health Information System (HIS):

A health information system is the total of resources, stakeholders, activities and outputs enabling evidence-informed health policy-making. The health information system manages all types of health data, from EHRs to imaging data and population health data. HIS activities include: data collection, interpretation (analysis and synthesis), health reporting, and knowledge translation, i.e. stimulating and enhancing the uptake of health information into policy and practice. Health information system governance relates to the mechanisms and processes to coordinate and steer all elements of a health information system.[32]

[31] PHIRI glossary. Available at: https://www.phiri.eu/glossary

[32] Marieke Verschuuren, Hans van Oers. Population Health Monitoring: Climbing the Information Pyramid. Available at: https://be1lib.org/book/3661052/728e4e

Metadata:

A set of data that defines and describes a resource (e.g., data, dataset, sample...) so that it can be understood, discovered and reused. There are different levels of metadata. Since metadata can be used to describe different aspects of data, we can group metadata properties in terms of quality, availability, provenance, processing, among others. Then there are metadata catalogues that can be developed to describe the available data collections in a repository or hub. Metadata is important to make data understandable, and can contribute to increase the findability, accessibility, interoperability and reusability of the data. Metadata can be collected or compiled in repositories to improve the FAIRness level of the data collections.

Non-personal data:

All data other than personal data.[33]

Open data:

Data that is freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

Open license is a license agreement which contains provisions that allow other individuals to reuse another creator's work, giving them four major freedoms. Without a special license, these uses are normally prohibited by copyright law or commercial license. Most free licenses are worldwide, royalty-free, non-exclusive, and perpetual (see copyright durations). Free licenses are often the basis of crowdsourcing and crowdfunding projects.[34]

Open science:

The movement to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of an inquiring society, amateur or professional. Open science is transparent and accessible knowledge that is shared and developed through collaborative networks. It encompasses practices such as publishing open research, campaigning for open access, encouraging scientists to practice open-notebook science, and generally making it easier to publish and communicate scientific knowledge.[35]

Personal data:

[33] Proposal for a regulation of the European Parliament and of the Council on European data governance. 2020/0340. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

[34] Wikipedia, Open Data. Available at: https://en.wikipedia.org/wiki/Open_data

[35] Wikipedia, Open Science. Available at: https://en.wikipedia.org/wiki/Open_science

According to Article 3 (1) of Regulation (EU) 2018/1725: "'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".[36]

Name and the social security number are two examples of personal data which relate directly to a person. However, the definition extends further and also encompasses for instance e-mail addresses and the office phone number of an employee. Other examples of personal data can be found in information on physical disabilities, in medical records and in an employee's evaluation.

Personal data which is processed in relation to the work of the data subject remain personal/individual in the sense that they continue to be protected by the relevant data protection legislation, which strives to protect the privacy and integrity of natural persons. As a consequence, data protection legislation does not address the situation of legal persons (apart from the exceptional cases where information on a legal person also relates to a physical person).

Personal data breach:

A breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed.[37]

Primary use of data:

The use of any data for the purpose for which it was originally collected.

Processing (personal and non-personal):

Any operation or set of operations which is performed on data or on datasets, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure

[36] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1725&from=EN

[37] Art. 4(12) GDPR, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.[38]

Profiling:

Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.[39]

Pseudonymisation:

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.[40]

Restriction of processing (personal and non-personal data):

As defined by the GDPR, methods by which to restrict the processing of data could include, inter alia, temporarily moving the selected data to another processing system, making the selected personal data unavailable to users, or temporarily removing published data from a website. In automated filing systems, the restriction of processing should in principle be ensured by technical means in such a manner that the personal data are not subject to further processing operations and cannot be changed. The fact that the processing of data is restricted should be clearly indicated in the system. [41]

Secondary use of data/data re-use:

[38] Art. 4(2) GDPR, Regulation (EU) 2016/679 of the european Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

[39] Art. 4(4) GDPR, Regulation (EU) 2016/679 of the european Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

[40] Art. 4(5) GDPR, Regulation (EU) 2016/679 of the european Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

[41] Par.67 GDPR, Regulation (EU) 2016/679 of the european Parliament and of the Council of 27 April 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=es

Secondary use refers to using data for a different purpose than the one it was originally collected for (i.e. than the primary use).

According to the European Data Governance Act 2020 're-use' means the use by natural or legal persons of data held by public sector bodies, for commercial or non-commercial purposes other than the initial purpose within the public task for which the data were produced, except for the exchange of data between public sector bodies purely in pursuit of their public tasks.[42]

Clinical definition: Secondary use of health data applies personal health information (PHI) for uses outside of direct health care delivery.[43]

Secure processing environment:

The physical or virtual environment and organisational means to provide the opportunity to re-use data in a manner that allows for the operator of the secure processing environment to determine and supervise all data processing actions, including to display, storage, download, export of the data and calculation of derivative data through computational algorithms.[44]

Sensitive data:

Information that is regulated by law due to possible risk for plants, animals, individuals and/or communities and for public and private organisations. Sensitive personal data include information related to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership and data concerning the health or sex life of an individual. These data that could be identifiable and potentially cause harm through their disclosure. [45]

[42] Proposal for a regulation of the European Parliament and of the council on European data governance. 2020/0340. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

[43] Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper J Am Med Inform Assoc. 2007 Jan–Feb; 14(1): 1–9. Charles Safran, MD, MS, Meryl Bloomrosen, MBA, W. Edward Hammond, PHD, Steven Labkoff, MD, Suzanne Markel-Fox, PHD, Paul C. Tang, MD, Don E. Detmer, MD, MA. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2329823/

[44] Proposal for a regulation of the European Parliament and of the council on European data governance. 2020/0340. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN)

[45] https://hal.archives-ouvertes.fr/hal-03226010

2.1. Artificial intelligence (AI) related terms

Federated data analysis:

Federated data analysis describes an analysis that is performed on multiple (often geographically) separated datasets. During this analysis, the data is not exchanged and can stay, for example, behind a given institution's firewall. Only the interim results of a local analysis are exchanged between the data-hosting sites)[46]. The aggregated non-identifiable results from each local analysis are pooled and returned to the data user.

Federated learning:

This is a specific case of federated data analysis, for machine learning purposes. It is a learning technique that allows users to collectively reap the benefits of shared models trained from rich data collections. The learning task is conducted across multiple separate sites coordinated centrally. Each site has a local training dataset which is never shared . Instead, each site computes an update to the current global model maintained centrally, and only this updated model is communicated[47].

Input Data:

Data provided to or directly acquired by an AI system on the basis of which the system produces an output.[48]

Machine learning:

A subset of AI techniques based on the use of statistical and mathematical modelling techniques to define and analyse data. Such learned patterns are then applied to perform or guide certain tasks and make predictions.[49]

Synthetic data:

[46] https://www.foldercase.com/blog-federated-data-analysis-how-to-get-started.php

[47] Adapted from Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54:1273-1282, 2017.

[48] Regulation of the European Parliament and Council laying down harmonised rules on Artificial Intelligence https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

[49] WHO. (2021). Ethics and Governance of Artificial Intelligence for Health. Available at: https://apps.who.int/iris/rest/bitstreams/1352854/retrieve

The concept of synthetic data generation is to take an original data source (dataset) and create new, artificial data, with similar statistical properties from it.

Keeping the statistical properties means that anyone analysing the synthetic data, a data analyst for example, should be able to draw the same statistical conclusions from the analysis of a given dataset of synthetic data as he/she would if given the real (original) data.

The use of synthetic data is growing in many fields: from training of artificial intelligence models within the health sector to computer vision, image recognition and robotics fields.[50]

Testing Data:

Data used for providing an independent evaluation of the trained and validated AI system in order to confirm the expected performance of that system before its placing on the market or putting into service.[51]

Training Data:

Data used for training machine learning algorithms (e.g., an artificial intelligence (AI) system) through fitting its learnable parameters.[52]

Validation Data:

Data used for providing an evaluation of the trained AI system and for tuning its non-learnable parameters and its learning process.[53]

2.2. Profiles as defined within the scope of HealthyCloud

Data curator:

[50] https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en

[51] Regulation of the European Parliament and Council laying down harmonised rules on Artificial Intelligence https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

[52] Regulation of the European Parliament and Council laying down harmonised rules on Artificial Intelligence https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

[53] Regulation of the European Parliament and Council laying down harmonised rules on Artificial Intelligence https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

A person who is responsible for the quality and FAIRness of the health-related data, and to make sure the value of the data is discovered and accessible. This role also considers the possibility of enriching data when increasing its quality. Importantly, data curators might play a role regarding being processors, e.g. responsible for the data at hand.

Data steward:

A person who has an administrative role; they do not really use the data. They create guidelines to make data FAIR and advice on how to do it. Stewards might have direct responsibility on the data at hand (processors) or not.

Infrastructure provider:

The responsible organisation to support the physical management of health-related data following existing regulations. Parent definition for data hub, data collection and secure processing environment.