DATA VALLEY

Data Valley White Paper

e-Health Data Sharing

Best practices and solutions for data sharing, anonymization, and data lake creation with health data.

ENG Version of the October 2021 DRAFT

Participants in the working table and drafting of the white paper:

Carlo Rossi Chauvenet, Data Valley

Silvia Martinelli, Data Valley - Università degli Studi di Torino

Paola Aurucci, Università degli Studi di Torino

Alessandra Salluce, Università degli Studi di Milano

Piergiorgio Chiara, University of Luxemburg - LAST-JD-RIoE

Vanessa Cocca, CRCLEX

Giorgio Presepio, San Raffaele (DPO)

Alexandru Raileanu – AWS

Elena Cappellaro, Astrazeneca

Federica Rizzo, EPEE / Green Innovation Network

Paolo Bartoli, Cloud-R - Soluzioni software per registri di malattie rare

Ruggero Di Maulo, Cloud-R - Soluzioni software per registri di malattie rare

Daniele Panfilo, Aindo - Tecnologia all'avanguardia per la produzione di dati sintetici

Jovan Stevovic, Chino.io - Soluzione software per la gestione del dato sanitario

Absctract

The White Paper of Data Valley "E-health Data Sharing" arises from the need, shared by operators in the sector, to reconstruct the applicable legal framework and define best practices for sharing data and creating data spaces in the healthcare sector.

The first section describes the impact of data analysis in the healthcare sector, the European perspectives for creating dataspaces, and the emergence of the need to identify, define and consolidate models and best practices for anonymization and sharing.

The second section, dedicated to the use of health data for treatment and research purposes, reconstructs the applicable legal framework, differentiating between personal data, anonymous data, and pseudonym data according to the GDPR; analyzing the specificities relating to the processing of health data for treatment and research purposes, also distinguishing between experimental research and observational research; deepening the problems relating to the sharing of "group data".

The third section focuses on the evolutionary profiles to enhance health data among the actors in the ecosystem by creating data lakes in health care to improve research/care through the concentration of data.

Finally, the fourth section describes the technologies for anonymization and synthetic data, the choice of technology, and its implementation.

TABLE OF CONTENTS

Abstract **3**

Data Sharing in Health Care and the Goals of the White Paper. **5**

Introduction 5

Description of the scenario: the impact of data analysis in the healthcare 7

European perspectives: the creation of data spaces in the health care sector 9

How the white paper has been developed: the need to identify, define and consolidate models and best practices for anonymization and data sharing in the health care sector 10

The use of health data for treatment and research purposes **12**

Personal data, anonymous data and pseudonymous data under the GDPR 12

Health-related data: treatment purposes and research purposes 14

Granular data, clusters, and groups 19

Evolving profiles of health data valorization among actors of the ecosystem **22**

Data access and security in sharing: the European perspective 22

Data lake in healthcare: improving research/care through data concentration 25

Organizational Tools: partnerships, consortia, contracts 26

RWE data in the case of rare diseases: how to extract new value from it 30

Technologies for anonymization **32**

The choice of technologies for anonymization 32

Aindo and synthetic data 35

The choice of dataset: primary scientific stage, feasibility,
stage of definition for secondary use. 35

Example of reference software architecture 37

Conclusions and future prospects 38

Data Sharing in Health Care and the Goals of the White Paper.

Introduction

Carlo Rossi Chauvenet, Silvia Martinelli

The use of data and algorithms for organizing production and bringing supply and demand together has brought about a paradigm shift that has affected both forms of production and exchange, as well as the product itself.

The paradigm shift described is enabled by the creation and management of data flow. These can be personal ones entered by the user or generated in the interaction with the product, or those collected by sensors and related to the surrounding environment, or even those collected by thousands of other applications with which the product and its user necessarily interact.

Thus, data-driven business models are multiplying in recent years and at this moment in history, all based on new forms of using the information collected in them. At the same time, the interest in having access to additional databases necessarily increases, in order to be able to generate new correlations and new services to be proposed to end users, consumer or business or even public entities.

The sharing of data and its reuse in innovative ways for creating new smart products and services, however, face some obstacles.

First, the use of the data, where personal or even where then anonymized to the point of its anonymization, requires as well known and dutiful the application of all the principles, cautions and procedures provided by our legal system for the processing of personal data.

Second, the sharing of data among different entities, private or public, requires agreements, partnerships or the construction of new legal structures for the management of data governance and for the regulation of all potential issues that may arise from the sharing itself. In particular, agreements will need to be made with regard to the possibilities and modalities for future decision-making, the sharing of risks and the predetermination of responsibilities, as well as the protection of the investment made.

Precondition for the sharing itself is, moreover, the encounter that generates it, becoming critical the identification of the partner who is in possession of or who is able to acquire the desired data asset or, on the other hand, manages the interface or product or sensor that dialogues with the user or the environment that one wishes to reach.

Third, but again a fundamental precondition, is technical dialogue and software integration. The latter is, in fact, critical for real-time communication between systems, for data quality, as well as for reaching the end customer itself, accessing the desired interface or product.

"Data Valley" - www.datavalley.it - is a project that was created to address this need for sharing and integration, carefully evaluating contractual and compliance elements as well. It specialized in the analysis of these issues, first by organizing meetings in the form of a symposium between Triveneto businesses and Big Tech, and later by creating a path of systematic analysis of economic, technical and legal aspects for the creation of new partnerships and synergies for data sharing and technological integration.

The initial experience was then continued online with the creation of restricted-participation working tables focused on specific, shared issues and needs, and the first one initiated led to the development of this White Paper.

The working tables were born from a need to connect multiple stakeholders belonging to the same sector but representative of different stakeholder categories, for the identification of sharing needs as well as obstacles in order to work together in overcoming them.

Starting in November 2020, on an almost monthly basis, Working Group members have been meeting to share experiences and needs, identifying and dissecting common needs and issues. Their experiences, discussions, concerns and aspirations led to the development of the White Paper.

The first section of the White Paper is devoted to describing the impact of data analysis in the health sector, European perspectives on the creation of data spaces, and the emerging need to identify, define, and consolidate models and best practices for anonymization and sharing.

The second section, focused on the use of health data for treatment and research purposes, reconstructs the applicable legal framework, differentiating between personal data, anonymous data, and pseudonymous data under the GDPR; analyzing the specificities related to the processing of health data for treatment and research purposes, including distinguishing between experimental and observational research; and delving into issues related to the sharing of "group data."

The third section focuses on the evolving profiles of the leveraging of health data among ecosystem actors for the creation of data lakes in healthcare for the improvement of research/care through data concentration.

The fourth section is, finally, devoted to technologies for anonymization and synthetic data, technology choice and implementation.

Description of the scenario: the impact of data analysis in the healthcare

Paola Aurucci

Over the past 25 years, the invention, development and diffusion of ICT (information and communication technologies) has greatly expanded the scope of data production, collection, storage and sharing[1]. Increasingly large digital databases and increasingly sophisticated systems of analysis have led to the rise of so-called data centrism[2], which has enormous implications for how scientific research is conducted, organized, governed and evaluated[3].

Going into the specifics, what really changes from the past due to the proliferation of devices suitable for digital data recording in heterogeneous environments is that it allows for continuous real-time digital imaging of different social and technical systems, on a global scale and with high resolution of individual behaviors. This new ability to measure humans is accompanied by new ambitions to understand these systems and control them. Not surprisingly, an immediate consequence of the change in the scale of the numerosity of the contactable and measurable population within the medical sciences has been an extraordinary development of observational clinical and epidemiological research on real word data, inclusive of primary and secondary prevention and care in the narrow sense. Such observational studies on large data sets, thanks to the availability of innovative computational technologies, have also made it possible to explore the best combination of available variables in a controlled setting to predict a given outcome (e.g., studies aimed at identifying patients who have a higher probability of benefiting from a specific treatment).

The biomedical sphere, then, has been particularly affected by the digital revolution. The growth rate of electronic data in this context is, in fact, above average. This is occurring by virtue of four major phenomena: (i) digitization of imaging; (ii) digitization of medical records and health files (iii) explosion of the Internet of Things (hereafter "IoT" and (iv) the development of Next Generation Sequencing (hereafter "NGS") sequencing techniques-also called Second Generation Sequencing or High-throughput Sequencing. The latter are employed in the field of so-called "omics" sciences and allow with reduced time and high analytical sensitivity, to acquire a huge amount of data related to different hierarchical levels of biological complexity (DNA, mRNA, proteins, metabolites, etc.). The use of such techniques has made it possible to provide a comprehensive view of the cellular and molecular processes that characterize individuals, contributing to revolutionize the study of complex systems (systems biology), which through integrative modalities and advanced computational models aims to answer complex biological questions such as pathogenesis, natural history and evolution of diseases.

[1] Pagallo, U., Il diritto nell'età dell'informazione, G. Giappichelli Editore, 2014, p. 174.
[2] Floridi, L., The 4th revolution: how the infosphere is reshaping human reality, Oxford, 2014, p. 96.
[3] Leonelli, S., La ricerca scientifica nell'era dei Big Data, Meltemi, 2018, p. 31.

This enormous amount of data, coming from heterogeneous sources that collect and update data for reasons largely unrelated to clinical and epidemiological research, if not properly analyzed and integrated, risks becoming a handicap when one wishes to translate them into new scientific discoveries. Fortunately, the availability of such data also represents a unique opportunity to train machine learning algorithms, a key factor in the development of artificial intelligence, to be used for disease prevention, diagnostics and new drug development. In fact, the analysis of this novel data stream by means of artificial intelligence techniques, and machine learning in particular, makes it possible to automatically identify correlations from which it will be possible to make "predictions" using inductive reasoning and formulating hypotheses. In particular, the use of machine learning to integrate the huge amount of data produced by second-generation sequencing techniques used in molecular biology makes it possible to objectify and quantify the heterogeneous nature of most diseases and the phenotypic variability of individuals at the level of genomics, epigenomics, transcriptomics proteomics and metabolomics, the so-called "panomics." It will then be possible to proceed with increasingly precise patient profiling and sew treatment according to the individual's genetic profile. A further conceptual progression can be identified in precision medicine, defined by the National Institute of Health (hereafter "NIH") as "an emerging approach to the treatment and prevention of disease that takes into account individual variability in genes, environment, and lifestyle,"[4] i.e., that takes into account not only genetic variability, but also environment and microbiota composition. The affirmation of this approach will depend on the integration of huge amounts of data produced by the use of high-processing methods for molecular characterization of patients, together with an equally huge amount of physiological, clinical and environmental data obtained from multichannel technologies such as smartphones and wearable sensors (as well as from information obtained through frequenting compulsive social media) and their analysis by machine learning tools.

We are thus only at the beginning of a process that could result in an epochal revolution in clinical practice and health care, which finds in the data intensive research the crucial element that proposes, raises and creates a number of critical issues and unprecedented questions. In addition to epistemological issues, related for example to the real reliability of the evidence produced by the analysis of heterogeneous data, and technical issues, related to the development of systems capable of safely processing and analyzing a huge amount of data, the most recalcitrant critical issues are ethical and legal. And this applies both to classical retrospective observational research on data from the real world, and to investigations in the field of precision medicine involving the systematic use - often by different research centers - of personal data of a sensitive and ultra-sensitive nature (genetic data) for purposes beyond those for which they were initially collected (secondary use). Indeed, considering the proliferation of digital recording systems, mobile devices, and wearables in the health care environment, as well as the inherent research value of health and genetic data, the rise of data-driven research implies, as pointed out by Mittlestadt and Floridi, "the impossibility of predicting at the time of collection all future uses

[4] National Institute of Health, The promise of precision medicine, reperibile sul sito internet: www.nih.gov/about-nih/what-we-do/nih-turning-discovery-into-health/promise-precision-medicine.

of the data"[5]. This undermines the use of consent as a legal basis for the processing of these data, since to be valid it must be informed and specific, i.e., referring to a specific purpose or purposes, and this is not possible since these data will likely need to be reused, shared, and aggregated to others for research purposes. In such cases, re-contacting each individual patient to inform them about the new research purpose is excessively costly, is organizationally impossible, or could jeopardize the achievement of the research purposes. Such purposes are even unknown at the time of processing in observational research based on the use of machine learning, which is capable of making transformative use of information, that is, identifying correlations invisible to the naked eye of the researcher, not even abstractly predictable prior to data analysis. Such research is, in fact, aimed at identifying the study hypothesis - not test it - and severely challenge the re-consensus approach. Alternative to the latter is anonymization, which, however, is difficult to achieve in a big data context.

European perspectives: the creation of data spaces in the health care sector

Vanessa Cocca

The General Data Protection Regulation (GDPR) has created a level playing field for the use of personal data, including health data. However, the landscape of digital health services, within and between European member states, remains fragmented due to different national regulatory transpositions.

Regulatory fragmentation in Europe regarding the processing of health data is a major obstacle for players in the health sector. As a result, the European Commission considers it essential to strengthen and expand the sharing, use and reuse of data health to boost innovation in the biomedical sector.

The Commission itself will promote, as discussed by Member States on the occasion of the "Recovery and Resilience Facility"[6], the establishment of European common data spaces (data spaces[7]) in strategic economic sectors and areas of public interest, in order to make large amounts of data available to actors in a sector.

Specifically, each common data space will feature a distinctive legislation and governance model based on the relevant sector to ensure full use and interoperability of data[8]. The data space is thus intended to be a tool regulated at the European level and developed in full compliance with

[5] Mittelsadt, B.D. e Floridi, L., The ethics of big data: Current and foreseeable issues in biomedical context, in Science and Engineering Ethics, vol. XXII, n. 2 (2016), p. 303-341.

[6] p. 17, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - 2030 Digital Compass: the European way for the Digital Decade, Bruxelles, 9 marzo 2021.

[7] A data space is an infrastructure that connects several virtual storage facilities containing only data (not, for example, common areas, system data or programs) and with which it interacts through an API or software.

[8] p. 29 , Comunicazione Della Commissione Al Parlamento Europeo, Al Consiglio, Al Comitato Economico E Sociale Europeo E Al Comitato Delle Regioni - Una strategia europea per i dati, Bruxelles, 19 febbraio 2020.

EU legislation on data protection data and compliant with the highest available cybersecurity standards.

The biomedical-health sphere, by virtue of the peculiarity and implications of the data being processed, represents an area in which data use can have a systemic impact on the entire ecosystem. Accordingly, the Commission proposes the establishment of a European health data space, aimed at: helping health authorities in making data-driven decisions to improve the accessibility, effectiveness, and sustainability of health systems; contributing to the competitiveness of the EU health industry; supporting the work of health system regulators in the evaluation of drugs or biomedical products and the demonstration of their safety and effectiveness; and finally, ensuring citizens' access, control, and portability of personal health data by implementing an electronic health record (EHR) safeguarding privacy.

How the white paper has been developed: the need to identify, define and consolidate models and best practices for anonymization and data sharing in the health care sector

Carlo Rossi Chauvenet

The goal of the white paper is to define the regulatory and technical framework that represents the new level playing field in which healthcare stakeholders are increasingly called upon to operate.

In the traditional relationship between the patient and the physician, important spaces of interaction have opened up, governed by technology, which require substantial investments and a systemic vision. The reference is obviously related to biomedical engineering companies, medical devices and all devices that measure people's lifestyles, telemedicine services to biomedical research platforms.

This domain is expanding greatly, but it is very fragile because it is left to the regulatory choices of individual nation states, which are always particularly constraining in this area. In turn, these choices depend on the assessment of available technologies that is made by individual regulatory bodies who are often influenced by news stories of incidents related to the use of certain technologies in their early stages and the resulting public concerns.

For this reason, it is increasingly appropriate for operators in the entire health sector to define in unified documents the framework of the needs and solutions envisaged in the interest of the patient, self-defining a consensus on regulatory and technological elements to encourage investment and shelter them from regulatory onslaughts in the early stages of development of a highly innovative sector. It is the first step in creating a testbed space for sharing solutions and integrating services along the lines of the "sandbox" model used in the United Kingdom with regard to financial regulation.

On the Data Protection front in health care, the issue is increasingly topical given the need to share and integrate large volumes of personal and non-personal data made possible by the use of innovative data anonymization techniques such as the use of synthetic data.

In the remainder of the paper, the needs of the industry, the current regulatory framework, and the available technological solutions will be analyzed, formulating proposals for advancing the regulatory framework to protect investment in the industry.

The use of health data for treatment and research purposes

Personal data, anonymous data and pseudonymous data under the GDPR

Piergiorgio Chiara

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie ITN EJD grant agreement No 814177

The GDPR applies only to personal data. Non-personal data therefore fall outside its scope. The legal classification of data is therefore a topic of central importance as it determines whether the entity processing the data is subject to the various obligations that the regulation imposes on data controllers. Yet the binary construction of the European data protection regime, five years after the Regulation came into force, still does not provide the legal certainty desired by market actors.

The Regulation defines personal data in Article 4(1) as any information relating to an identified or identifiable natural person. Furthermore, an identifiable person is any natural person who can be identified, directly or indirectly, with particular reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more features of his or her physical, physiological, genetic, mental, economic, cultural, or social identity.

Before examining in more detail, the test adopted by the Regulations to determine the personhood of data, a second relevant dichotomy should be highlighted in the context of non-personal data. Indeed, some data are always non-personal because they have never concerned an identified or identifiable natural person. Others, on the other hand, are originally personal data within the meaning of Article 4(1) but, as a result of an operation aimed at removing the link with the natural person, become non-personal because the natural person is no longer identified or identifiable. It is especially the latter category of data that gives rise to the aforementioned technical-legal issues that closely affect research, especially in biomedical field.

In this context, Recital 26 of the Regulation sets out the test to be performed to shed light on the different processing techniques that invest the binary distinction between personal and non-personal data. In particular, the case of pseudonymization and anonymization should be analyzed in more detail.

Pseudonymization is conceived by the GDPR as a means of reducing risks to data subjects by "hiding" the identity of individuals in a dataset, for example, by replacing one or more personal identifiers with so-called pseudonyms. Obviously, the technical-logical link between pseudonyms and initial identifiers must be appropriately protected by the data controller.

The risk of re-identification is reduced, and it is certainly true that such processing prevents direct identification of the data subject. Yet, according to Recital 26, personal data subjected to pseudonymization techniques should be considered information about an identifiable natural person and therefore fall under the scope of the Regulation, as it could still be attributed to a natural person through the use of additional information.

Conversely, the same Recital states that the Regulation should not apply to anonymous information, that is, information that does not relate to an identified or identifiable natural person or to personal data rendered sufficiently anonymous that it prevents or that the data subject can no longer be identified.

A closer reading of the text reveals the heart of the Recital 26 test. Indeed, to establish the identifiability of a person it is appropriate to consider all means, such as identification, that the data controller or a third party may reasonably use to identify that natural person directly or indirectly. To ascertain the reasonable likelihood of using the means to identify the natural person, all objective factors should be considered, including the cost and time required for identification, taking into account both the technologies available at the time of processing and technological developments.

The test elaborated in recital 26 of the GDPR essentially embraces a risk-based approach to determine the personhood or otherwise of the data. Where there is a reasonable risk of identification, the data should be treated as personal data. Where, on the other hand, that risk is negligible, the data can be treated as non-personal data, and this is so even if identification cannot be ruled out with absolute certainty[9].

This reading based on the risk approach of the Regulation has found resistance especially in the so-called "absolutist" reading of the Working Party art. 29, followed by some supervisory authorities, such as the French[10] and Irish[11] Data Protection Authorities. This interpretation takes into account all the possibilities and occasions in which anyone would be able to identify the data subject: while the GDPR explicitly refers only to the possibility of identifying the individual, the Working Group goes further, adding to the de-identification test the criteria of (i) linkability of information about the individual in different datasets; and, (ii) inference, i.e., the possibility of inferring, with significant probability, the value of an attribute from the values of a set of other attributes[12].

Thus, the Working Group art. 29 sets a high threshold to be met, establishing its "zero risk test" according to which no risk of re-identification can be tolerated. This would imply a perfect equation between anonymization and deletion: the result of such a technique should be permanent, making it impossible for any technical operation to re-identify the individual to whom the personal data originally referred.

The absolute approach, however, can hardly be sustained: there is a flourishing literature on the non-absolute nature of anonymization[13]. Therefore, if we could never rely on the non-personality

[9] Finck, M. e Pallas, F., “They who must not be identified—distinguishing personal from non-personal data under the GDPR” (2020) International Data Privacy Law, 10(1), 11-36.

[10] Commission Nationale de l'Informatique et des Libertés, “Comment prévenir les risques et organiser la sécurité de vos données ?” (2019).

[11] Data Protection Commission, “Guidance on Anonymisation and Pseudonymisation” (2019).

[12] Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques (WP 216) 0829/14/EN, 3.

[13] Ohm, P., “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization” (2010) UCLA Law Review, 57(2); Sweeney, L., “Simple Demographics Often Identify People Uniquely” (2000) Health, 671; Narayanan, A., e Shmatikov, V., “Robust De-anonymization of Large Sparse Dataset”(2008) IEEE Symposium on Security and Privacy.

of the data, then any information would always remain within the scope application of the GDPR.

Another reading of Recital 26, so-called relativist, considers only the efforts required to identify an individual, without delving into the murky field of mere theoretical possibilities. Several authors, and to some extent the UK Data Protection Authority[14], have argued that data resulting from anonymization operations by cryptographic techniques should not be considered personal data if two requirements are met: the cryptographic method must be effective, robust, and up-to-date, and the data controller (or any third party) is not in possession of the decryption key, nor is there a reasonable chance that he or she will obtain the key. This reasoning has been particularly successful in the field of cloud computing[15].

However, the most convincing position is the one based on the risk approach, which inspired the Regulation and has been confirmed by the Court of Justice[16]. If there is a reasonable likelihood that certain data, even if subjected to irreversible encryption operations (e.g., salted/peppered hash function) in order to achieve anonymization, can be (re)linked to the natural person to whom they were originally referenced, they must be qualified as personal data. In contrast, if de-identification has been sufficiently robust so that identification is no longer reasonably likely, that data should be considered non-personal[17].

Health-related data: treatment purposes and research purposes

Paola Aurucci, Giorgio Presepio

"Health-related data," along with biometric and genetic data, are included by the GDPR in the "Special Categories of Personal Data" (what are referred to in common parlance as "sensitive" or "super-sensitive" data). The identification of these Special Categories of data, made specifically by Article 9 of the GDPR, is functional to the provision of a more restrictive regulation of their processing. A higher protection for this type of data is guaranteed by virtue of their inherent dangerousness, since they not only identify the individual (as it is for common data) but they indefectibly contribute to the construction of his or her identity, and for this reason they are susceptible to be a source of abuse and discrimination if improperly processed. In confirmation that the processing of health-related data (and biometric and genetic data) is even more dangerous than the processing of other special categories of data, paragraph 4 of Article 9 authorizes the Member States to maintain or introduce additional conditions and possibly limitations to the discipline provided by the aforementioned Article 9, which turns out to be only a minimum discipline with respect to this type of data. Article 4 of the GDPR defines health-related data as that "pertaining to the physical or mental health of a natural person,

[14] Information Commissioner's Office, "Anonymisation: Managing Data Protection Risk Code of Practice" (2012)

[15] Hon, K.W., Millard, C. e Walden, I., "The problem of 'personal data' in cloud computing: what information is regulated? -the cloud of unknowing" (2011) International Data Privacy Law, 1(4).

[16] Case C-582/14 Patrick Breyer [2016] EU:C:2016:779.

[17] Cfr. con AEPD ed EDPS, "Introduction to the Hash Function as a Personal Data Pseudonymisation Technique" (2019)

including the provision of health care services, revealing information relating to his or her state of health." Recital 35 specifies a further important aspect, namely the temporal aspect, by specifying that health-related data relate to an individual's physical or mental condition whether past, present, or future. The recital goes on to specify that these are data typically collected (although not stated as an exclusive situation) in the course of registering for access to a benefit and goes on to provide a (nonexhaustive) list of examples of such data: a specific symbol or element attributed to a natural person to uniquely identify him or her for health purposes; information resulting from examinations and checks carried out on a body part or organic substance, including genetic data and biological samples; and any information regarding, for example, a disease, disability, disease risk, medical history, clinical treatments, or physiological or biomedical status of the data subject, regardless of the source, such as, for example, a physician or other health care provider, a hospital, a medical device, or an in vitro diagnostic test. The European Data Protection Board went on to add that information that when cross-referenced with other data is likely to reveal health status or health risks (e.g., the presumption that a particular person is at a higher risk of heart attacks based on repeated blood pressure measurements over a certain period of time), in addition to information collected by a health care provider in a medical record, should also be considered sources of health-related data, self-assessment tests, in which individuals answer questions related to their health (e.g., describing symptomatology) and information that as a result of its use in a specific context reveals the individual's health status (e.g., information related to a recent trip or stay in a COVID-19-affected region processed by a health professional to make a diagnosis). On the basis of these assessments, on the other hand, a mere representation of the subject's physical reality (e.g., his or her picture, audio of his or her voice or heartbeat) should not be considered health-related data if that data is not then processed in such a way as to reveal elements related to the subject's health status. The definition of state of health should then include-in addition to a pathological condition-both the condition of good health, both physical and psychological[18] -and that of recovery from a disease.
In light of these assumptions and the relevant case law of the Court of Justice of the European Union ("CJEU")[19], it can be inferred that the term "data relating to health" should be interpreted broadly.

The regulations under the GDPR seek to reconcile the protection of the individual, with regard to sensitive information about him or her, with the need for economically and socially relevant activities, such as scientific research, to be carried out. The latter, in particular, enjoys particular favor within the Regulation, which although it does not provide an explicit definition of "data processing carried out for the purpose of scientific research" at Recital 159 states that this assumption "should be interpreted broadly" and "taking into account the Community objective of creating a European research area - as provided for in Article 179 of the TFEU - so that "researchers, scientific knowledge and technologies circulate freely." The same recital goes on to give a wide range of examples of what should be meant by scientific activities in which "privately funded research" as well as "studies carried out in the public interest in the field of public health"

[18] See case C-101/01, Lindqvist, point 50.
[19] Ibidem.

are included. There is no doubt that this broad definition is intended to ensure that it includes both clinical trials, funded in most cases by pharmaceutical companies, and observational clinical studies. The " Article 29 " Working Group went on to point out that the interpretation of the term "scientific research" should not go beyond the meaning commonly given to it, i.e., "a research project established in accordance with relevant ethical and methodological standards sectors, in accordance with good practice."

In addition, as it will be seen below, for the first time the GDPR provides a specific exception to the prohibition on the processing of health-related data if it is necessary for such scientific research purposes.

With respect to the "processing of health-related data for scientific research purposes," it is necessary to distinguish between two different uses that can be made of these data. We speak of "primary use" when such data are collected directly for scientific study purposes. Examples of studies in the biomedical field that assume the primary use of health-related data are clinical trials and prospective observational studies. In such studies, the patient's health-related data are in fact collected ab origine for the specific purposes for which the study itself is being conducted and must be fully described to the subject before participating in the research. On the other hand, we speak of "secondary use" when the health data that are used research purposes were initially collected for other purposes (e.g., for treatment purposes within normal clinical practice, previous clinical trials, or previous and different observational studies). This is also referred to as "further processing for research purposes." A typical example of secondary use of health-related data for research purposes is found in retrospective observational studies in which personal data were previously collected for health care purposes or for the execution of previous research projects or were derived from biological samples taken previously for health care purposes or for the execution of previous research projects.

The distinction between scientific research based on the primary or secondary use of health-related data assumes particular importance in determining the legal basis for processing, information requirements, and the application of the principle of purpose limitation.

The proliferation of digital recording systems, mobile devices, and wearables in health care settings, considering the inherent research value of biometric and genetic health-related data, has triggered an unprecedented proliferation of data-intensive observational studies based on secondary use of health-related data. Health data routinely collected in normal clinical practice are thus continually being reused, shared, and aggregated with others for purposes other than those for which they were collected. Such additional research purposes are even unknown at the time of data access by the researcher in studies involving data analysis by artificial intelligence systems capable of identifying correlations and links invisible to the researcher's naked eye, not even abstractly predictable prior to data analysis, on which to then base predictive models that allow understanding that certain combinations of values of certain parameters are often associated with specific clinical conditions.

Experimental and observational research

Paola Aurucci

Biomedical research in very general terms can be defined as research of a multidisciplinary nature that increasingly employs integrated approaches that make use of complementary notions and methodological inputs typical of different scientific disciplines to understand physiological, pathological, and pharmacological mechanisms. It is preliminarily divided between preclinical research (research that is not conducted on humans) and clinical research (that is conducted on humans). The latter is conducted directly on humans (both healthy and sick) and is aimed at the direct study of disease for the development of new effective treatments for prevention, diagnosis, rehabilitation/assistance, and cure. Clinical research is based on various types of studies using both experimental and observational methodology. For this reason, at the regulatory level, Regulation (EU) No. 536/2014 to define clinical trials first establishes what should be meant by a clinical trial, i.e., " any investigation carried out in relation to human subjects aimed at: a) discovering or verifying the clinical, pharmacological or other pharmacodynamic effects of one or more medicinal products; b) identifying any adverse reactions of one or more medicinal products; or c) studying the absorption, distribution, metabolism and elimination of one or more medicinal products, in order to ascertain the safety and/or efficacy of those medicinal products. Only in the subsequent Article 2(2) does it specify that a clinical trial represents a subcategory "that meets one of the following conditions: (i) the assignment of the subject to a particular therapeutic strategy is decided in advance and is not part of the normal clinical practice of the member state concerned; (ii) the decision to prescribe the investigational medicinal products and the decision to include the subject in the clinical trial are made at the same time; (iii) additional subjects diagnostic or monitoring procedures in addition to normal clinical practice.

The Regulation then, following an approach that takes due account of international guidelines, particularly that of the United States whose regulations on experimental studies have for several years provided for a classification of these according to level of risk-introduces the concept of low-intervention level trials, in which the investigational drugs have already received marketing authorization, and are used according to the terms of the marketing authorization (comparative studies of authorized drugs) or on the basis of published/documented scientific evidence (e.g., off-label trials). Low-intervention trials are distinguished from "standard" clinical trials in that they involve minimal additional risk to subject safety compared to normal clinical practice. To assess the risk that exists for the subject under study, Recital 11 reminds us that this originates from two areas the investigational drug and the intervention, i.e., the clinical trial procedures. In qualifying which types of clinical trials can be qualified as "low-intervention," committees should therefore focus on whether there is real scientific evidence to support the use of the drug in the trial according to an indication other than that established by the AIC and on the possible risks "additional to normal clinical practice" posed by by the procedures involved in the trial (e.g., diagnostic and monitoring). This type of clinical trial should be subject to less stringent rules regarding monitoring, requirements applicable to the content of the permanent file, and traceability of investigational drugs. Studies classified as observational involve--like experimental studies--establishing a comparison between groups, only the phenomenon under study is not the effect of an experimental intervention, but of an exposure to a risk or protective factor. The latter is spontaneous in nature and is, therefore, not conditioned by the researcher who merely observes what occurs in nature (in clinical practice), not acting on the condition being studied, neither randomly assigning it nor modifying it. Research subjects are placed in comparison groups on the basis of personal characteristics or their experiences not conditioned by the study.

The observational method has developed mainly in epidemiological research, which has been defined as "the study of the distribution and determinants of health-related situations or events in a specific population, and the application of this study to the control of health problems." Observational studies can be prospective and retrospective. In the former at the time the study is planned both the exposure and the outcomes of interest have not yet occurred, consequently the data are collected prospectively and directly for the specific purposes of the study. In the latter, they have already occurred and the relevant data are collected retrospectively-as they are recorded in different datasets-and thus involve further processing of data initially collected for other purposes ("secondary use"). We find the normative definition of an observational study in subparagraph (p) of Article 1 of Legislative Decree No. 200/2007, according to which in such research "medicines are prescribed in accordance with the indications of the marketing authorization where the assignment of the patient to a particular therapeutic strategy is not decided in advance by an experimental protocol, is part of normal clinical practice and the decision to prescribe the medicine is completely independent of the decision to include the patient in the study, and in which no additional diagnostic or monitoring procedures are applied to patients." This definition reveals multiple critical issues. First, it includes only observational drug studies and does not pertain to observational studies in general. There are, in fact, numerous types of studies that methodologically can be classified as observational but do not fall under the definition provided in the regulations because they do not involve prescribing drugs (e.g., epidemiological studies, observational studies of medical devices, studies of biological samples and genetic data; studies of behavior or quality of life) contributing to the patchy regulation of observational studies. Moreover, according to epidemiological theory, the observational study may involve diagnostic and evaluative procedures that are not routine in the clinical practice of the participating subject, and therefore do not qualify as merely "additional" to them.

A partial remedy was AIFA's Guidelines for the Classification and Conduct of Observational Drug Studies. In that document, the following are included in the definition of "current clinical practice": "questionnaires, interviews, diaries, health economics and drug economy surveys, subjective assessments by the subject of his or her own health status, rating scales and blood chemistry tests, the use of which is justified by the study protocol." Ultimately, the national framework proves to be wholly deficient in regulating and classifying the various types of observational studies that can be conducted in health care and epidemiology. In Law No. 3/2018, "Delegation to the Government on the subject of clinical trials of medicines as well as provisions for the reorganization of the health professions and for the health management of the

Ministry of Health," the need for a new regulatory instrument - of a binding type - is highlighted. relating to observational studies in the biomedical and health fields.

Granular data, clusters, and groups

Alessandra Salluce

The increasingly pervasive use of IT tools and resources has led, as also mentioned in the previous Paragraphs, to the production and circulation of a truly impressive amount of data. However, one should not be surprised: the post-modern era, characterized by so-called "data-driven" business models, represents nothing but the natural evolution of the operational paradigms adopted in all productive sectors, and is undoubtedly doomed, in the near future, to further expansion.

The healthcare and research sectors are also involved: as an example, one only needs to think of the ever-increasing number of apps on the market capable of collecting in real time numerous personal information of a particular nature, as well as the use - increasingly frequent - of technological tools in the performance of more traditional medical activities or, again, the role of research, which, since its beginnings, has been nourished by datasets, but now has at its disposal increasingly cutting-edge tools that enable it to process once unimaginable masses of data.

That being said, on the one hand, there is no denying the absolute usefulness of such a development on the technological front - which, in the health field, has enabled the achievement of once unimaginable milestones - on the other hand, it is important to recognize the existence and relevance of the right to privacy, which is incumbent on each individual, especially in relation to certain types of information.

It is necessary, therefore, to find a "balance point," which allows for the need for knowledge required for certain purposes to be met, especially where related to collective interests worthy of satisfaction, such as health and progress in the medical field, but, at the same time, to protect the privacy of the individuals involved, while also taking into account the critical issues arising from the use of algorithms and the possibility of inferring personal information through data correlation. This crucial node, moreover, is linked to several aspects, including ethical ones, which primarily concern the possibility of giving rise to discrimination. In addition, along with the more strictly legal aspects, it is necessary to include in the analysis also the more technical aspects, especially related to the security of the information and the methods chosen for storage, access and transmission. In the context of health research, moreover, access to data does not concern so much a problem from the authorization point of view, since the transmission of the same is certainly authorized and necessary, but rather from the technical-organizational point of view, since the most critical aspect is related to the choice of the modalities of release of the information: from this choice, in fact, violations could result to the most confidential sphere of the individuals belonging to the sample to be analyzed, where from the data released, whether in granular or aggregate form, more intimate aspects could be deduced or, in the most serious, completely reconstruct the identity of the subject.

Within this complex framework, the application of data release techniques, combined with others of a more strictly IT nature aimed at preserving the security of information, can make a significant contribution[20].

On this point, as a preliminary matter, it is necessary to clarify the differences - mentioned just above - with regard to the type of data released for research purposes: in this regard, one speaks of "microdata" when the information contained in the statistical database is "pure," single; one speaks instead of "microdata" when the information is released in aggregate, statistical form.

In turn, the data, whether released in the form of microdata or macrodata, can be aggregated, going to compose groups, or clusters, on the basis of certain parameters that they have in common, depending on the purpose of the analysis to be conducted. In this case, however, the choice of grouping criterion must be well thought out, since a suboptimal choice may make the path to achieving one's research purpose more difficult, as well as, in some cases, lead to untrue results. Choosing the type of data best suited to our purpose, in any case, presents several critical issues from the point of view of privacy and personal data protection, since not only from such a choice derives the very application of the GDPR-which applies, as is well known, only to personal data, thus excluding anonymous data, assuming they really are-but also the application of the de-identification or pseudonymization measures deemed most appropriate. Relative to this aspect, it is possible to discern some substantial differences in the release of data in "pure" or aggregated format. First of all, pure data is the one that, by definition, is more delicate and deserving of stronger protections: if adequate data protection techniques are not applied, in fact, it is possible to trace back much more easily to the subject to whom they refer and, consequently, also to deduce additional information pertaining to his or her person. This can generally occur in two specific cases:

- when there is a particularly "conspicuous" piece of data within the dataset (such as may be, in a socioeconomic analysis, a salary much larger than others in a database confined to a small geographic area);
- when the data within the database are easily correlated with external information (this occurs when there are numerous matching attributes in the two related databases and the information in them is very accurate and detailed).

Obviously, individuals with peculiar or, even, unique characteristics are more exposed to the so-called "disclosure risk," which involves the identification of the individual or the inference of certain data, in some cases of a particularly confidential nature. Moreover, the two possibilities just outlined become more substantial where particularly accurate data are published and there is more than one external database with which to make connections.

Even the release of data in aggregate form, in any case, is not without risks and critical issues of this sort. First of all, it is worth specifying the existence of two possible forms of macrodata release: so-called "frequency tables" report the exact number or percentage of subjects sharing that particular attribute; "magnitude tables," on the other hand, report aggregate values (generically in the form of a mathematical mean) related to a particular attribute under analysis.

[20] *On this point, it should be specified that while cybersecurity is concerned with the security of information systems and the information flows exchanged through them, providing tools aimed at countering possible infiltration or, more generally, damage to software, hardware, or compromise of data security, data protection techniques are aimed at preventing the correlation and inference of information and identification of individuals.*

Of the two, the latter represent the more problematic ones, since the protection techniques applicable to frequency tables-such as, for example, sampling - may prove insufficient.

It was seen just above how the data released in its "pure" form presents more critical issues from the point of view of data protection, where even the application of specific techniques devoted to this (such as, among others, sampling, generalization, suppression or the addition of "noise") may in many cases allow for the re-identification of the data subject or the inference of other peculiar personal characteristics of the data subject. To assess more realistically the protection of the anonymity of the subjects represented in the dataset, however, it can be very useful to apply the parameters of k-anonymity, l-diversity and t-closeness[21] which allow, through different methods, to add in different and gradually increasing degrees of difficulties in the actual identification of the individual depicted in a given analysis group. These have been joined in recent years by more innovative techniques involving the addition of so-called "noise," such as differential privacy.

What has just been observed with reference to microdata is also valid, albeit with some exclusions, in the case where one has opted for the release of macrodata: the risks one may run into are the same, although the possible data protection techniques applicable are different. These include, but are not limited to, value sampling or the application of "threshold rules" or other special rules.

The main, and underlying, problem with both types of datasets discussed above can ultimately be channeled into the problem of anonymity: when can it be said with reasonable certainty that a piece of data has been rendered anonymous? And when, after applying one of the techniques known to date to protect the privacy of the individual, can one have less fear of possible re-identification?

In addition to such questions of a more purely technical-legal nature, those of an ethical derivation are also emerging more and more powerfully: for example, what are the most correct criteria for the homologation of individuals and their grouping into clusters? What are the possible consequences of a "privacy leak"? What are the discriminations that might result in the reuse of such data-even in aggregate form-for further purposes, especially where processing is done in an automated form?

These are crucial questions, to which too little attention is still paid in many cases, but which, especially in the health field, can lead to truly worrisome implications.

[21] *The property of "k-anonymity" allows for the identification, within a group, of at least k individuals who have a characteristic in common; the property of "l-diversity" indicates the amount of different sensitive attributes that each individual represented in a dataset must have to ensure a certain value of anonymity; the property of "t-closeness," on the other hand, aims to redistribute the data so that among the entire distribution of records and a selected part of it are very similar.*

Evolving profiles of health data valorization among actors of the ecosystem

Data access and security in sharing: the European perspective

Vanessa Cocca

Moving our analysis to the European level, it is worth noting that there have been numerous European impulses over the past few years to digitize the health sector and to sharing of health data[22]. Key action points include interoperability of information systems, data security and privacy-enhancing technologies, improved digital service infrastructure for eHealth, cross-border exchange of health data, common disease registries and platforms, tools for rare disease research, prevention and control of cross-border health threats, better use of European funding, and sharing of best practices.

Health and care systems need deep reforms and innovative solutions to become accessible and effective in providing care to European citizens. Data sharing is an essential step in achieving these goals: however, data are often available in formats that do not guarantee their interoperability and are often managed in ways that are disparate both across member states and within national health systems[23].

The emergency context related to the deployment of Covid-19 showed the potential and paved the way for the widespread use of innovative medical solutions, the use of telemedicine and remote assistance. Digital technologies can enable citizens to monitor their health status, prevent the onset of new diseases, and streamline the operation of healthcare systems. However, the health crisis has also exposed the vulnerabilities of the digital space, its dependence on critical infrastructures, often not based in European territories; highlighted dependence on a few large tech companies; seen an increase in the influx of counterfeit products and cyber theft; and amplified the impact of misinformation on our democratic societies[24]. In this regard, the European Commission estimates that the introduction of greater integration of online services, improved infrastructure for electronic transmission, and access to data could lead to benefits of up to €120 billion per year[25].

Building a common, multi-purpose pan-European interconnected infrastructure for data processing to be used in full compliance with fundamental rights, developing real-time peripheral

[22] See Strategy for a Digital Single Market in Europe, COM(2015) 192 final, 2015.

[23] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS on the. digital transformation of health and care in the digital single market, empowering citizens and creating a healthier society," COM(2018) 233 final.

[24] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS "2030 Digital Compass: European way for Digital Decade," COM(2021) 118 final. Compass: European way for Digital Decade," COM(2021) 118 final.

[25] Ibidem 8.

capabilities to serve the needs of end users close to where the data is generated, designing secure, low-power and interoperable middleware platforms for sectoral uses, and enabling easy exchange and sharing of data are among the priorities the European Union has identified in the Digital Compass 2030[26].

This vision has been described and incorporated within the EU4Health[27] Program for 2021-2027, which aims to digitally transform health services, promote interoperability, and develop a European health data space. EU4Health represents the European Union's response to the deployment of Covid-19. With an investment of €9.4 billion, EU4Health becomes the largest ever health program in terms of financial resources, and will provide funding to member states, health organizations and NGOs.

The program will also fund actions related to the creation of the European Health Data Space, among others. The creation of a European Data Space is one of the Commission's priorities for 2019-2025, including in the health sector. A common European Health Data Space will promote better exchange and access to different types of health data (electronic health records, genomic data, data from patient registries, etc.), not only to support health care delivery (the so-called primary use of data), but also for health research and health policy making (the so-called secondary use of data).

The system will revolve around adherence to the principles of transparency and protection of personal data of patients, on strengthening data portability, based on the provisions of Article 20 of the GDPR. The Commission will work together with member states to develop the European Health Data Space, the construction of which will revolve around three pillars:

1. A strong data governance system and a framework of data exchange rules;
2. Data quality;
3. Establishing a structure that can enable interoperability.

The Commission had already announced, in the European Data Strategy28 and in the more recent Data Governance Act[28], its intention to achieve concrete results in the area of health data and to exploit the potential created by developments in digital technologies to introduce innovation in health. The collection, access, storage, use and reuse of health data poses challenges that need to be addressed in a regulatory framework that best serves the interests and rights of citizens, particularly with regard to the processing of health status data.

Although the Cross-Border Healthcare Directive[29] created a collaborative framework among national authorities responsible for eHealth (the "eHealth Network"), existing agreements and tools provide and only partially address the challenges.

[26] Ibidem 8

[27] Regulation (EU) 2021/522 of the European Parliament and of the Council of March 24, 2021 establishing a programmen of Union action in the field of health for the period 2021-2027 ("EU Health Programme") (EU4Health) and repealing Regulation (EU) No. 282/2014.

[28] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European data governance, COM(2020) 767 final.

[29] Directive 2011/24/EU of the European Parliament and of the Council of 9 March 2011 , on the application of the rights of patients concerning cross-border health care.

Failure to exchange health data has a negative impact on the delivery of health services (and thus on the primary use of health data). The level of digitization varies widely within each Member State, and interoperability among health service providers remains limited. The eHealth network-and the related IT infrastructure-has improved the cross-border exchange of health data for health care, especially with regard to patient records and e-prescriptions. However, its voluntary and non-binding nature has negatively affected its adoption and limited its impact.

Exercising access and control over one's health records is often extremely difficult for patients. Electronic health records (EHRs) are not yet a reality throughout the Union, and many patients cannot easily access and use the information they contain, or transfer it between different providers, especially when the transfer is cross-border. This leads to duplication of effort, inefficiencies, delays in care, and higher costs for health systems and patients. The sharing of EHRs is limited, which means that this information cannot be easily shared in the treatment of patients.

As for the secondary use of health data, access to and exchange of health data for scientific research and innovation, new policy-making and regulatory activities remains very limited within the Union.

The collection, access, storage, use, and reuse of health data in health care poses specific challenges, mainly legal and technological in nature. In fact, from a regulatory perspective, the GDPR establishes a common framework of rules to which member states have added additional specifications and restrictions in relation to the processing and sharing of healthcare data. Thus, the processing of personal health data in member states appears to be fragmented, leading to obstacles and limited access by researchers and public institutions, which in turn reduces the EU's competitiveness and innovation potential globally.

Member states have different approaches to accessing and sharing health data. Some member states have established national bodies that facilitate access to health data; however, such bodies do not exist in all member states. Limited cooperation, governance, and IT infrastructure at the EU level hinder access to health data by researchers, public institutions and regulatory bodies.

A growing number of digital health tools then integrate artificial intelligence (AI) systems. The Commission is already working on a horizontal framework for AI that covers security and fundamental rights aspects, which is intended to be applied in various sectors, including health products. However, specific health-related aspects that rely on the future AI framework, including training, testing, and validation of AI systems, as well as aspects not covered by this horizontal framework may require further consideration.

The use of AI tools, and in particular the opacity of some applications, may make it difficult to assign responsibility or ensure compliance. It is therefore important to ensure adequate safeguards on fundamental rights and damages.

All these issues should be analyzed and solved in the European Heath Data Space; in particular, the program aims to:

a. Ensure the access, sharing, and optimal use of health data for health care delivery, as well as its reuse for research and innovation, policy development, and regulatory activities, in a secure, timely, transparent, and reliable and with appropriate institutional governance;
b. Promote a true single digital health market, covering health services and products, including telemedicine, telemonitoring and mobile health;
c. Improve the development, deployment and application of digital health products and services that are reliable, including those incorporating artificial intelligence in the health sector;
d. Establish an appropriate legal and governance framework to cover access and exchange of health data for health care delivery, research, policymaking, and regulatory activities.

The European Health Data Space, integrated with aspects of the Data Governance Act, will provide for the designation of national digital health bodies and sectoral bodies to deal with the secondary use of health data. It will also include: support for public authorities (e.g., medical agencies, epidemiological institutions, National Institutes of Health, HTA bodies, EMA, ECDC) to access health data in full compliance with data protection rules; access to genetic data and linkage with health data; reuse of data held by private entities; and support for training and testing of AI health applications. The interaction with the GDPR, particularly Articles 9 and 89, regarding the regulation of health data will be the subject of detailed study and analysis.
Efforts will also be made toward eliminating technical barriers to the use and reuse of data, particularly those related to infrastructure, interoperability, data quality, and standards in healthcare. Options regarding infrastructure for the use of data for health care will be examined, building on the eHealth digital service infrastructure (MyHealth@EU) for cross-border exchange of patient data when traveling abroad. Options regarding enhanced interoperability of electronic health records, in line with the European exchange format, as well as semantic and technical interoperability of different types of data will be explored. Regarding data access for research, policymaking, and regulatory purposes, options will address different models of interoperable data access infrastructure and related services to facilitate secure and cross-border storage, processing and analysis of health data.

Data lake in healthcare: improving research/care through data concentration

Daniele Panfilo

The ability to aggregate information into systems that allow for quick and easy access and reuse, such as data lakes or data warehouses, is one of the key drivers for research and development of data analytics-based solutions. This is even clearer in the case of health data. This has been further underscored by the current pandemic that has made the need for an infrastructure capable of facilitating the sharing, access and safe reuse of health data imperative.
While the exponential growth of available information follows a very high rate of growth due to the significant development and subsequent deployment of data acquisition devices, the access and reuse of the information asset shows very different growth factors quite different.

The causes hindering a broader democratization of health information, aimed at encouraging the rapid implementation of research and development projects, are varied and have different origins.
On the one hand, the lack of a standard platform for secure sharing of health data, and on the other hand, the sensitive nature of the data processed constitute some of the main factors behind the poor reuse of the information asset.
While The European Monitoring Tool[30] predicts that by 2025 the European data market will reach the value of more than 140 billion euros, it is clear that a real paradigm shift in technology will be necessary for the valorization of the information asset to take place as expected.
To this end, several European initiatives have emerged and are emerging, with the goal of enabling access to and reuse of health information assets through aggregation in data lakes or data warehouses. One example of such initiatives is provided by the eHealth platform Belgium, a Belgian government service, which offers the chance for actors in the health care landscape to securely exchange even sensitive information. Another case is that represented by the UK's The Health Data Research Hubs, which facilitates access to national health system data for the public, academic, and industrial research sectors in the United Kingdom.
The emergence of such hubs in several member states shows how central the issue of data sharing is and how this need is felt throughout the member states of the union.
The aggregation of such information through dedicated platforms not only stimulates research by facilitating access to data, but also enables the elimination of geographic barriers, encouraging the emergence of international collaborations and synergies-the lifeblood of scientific progress.
For the European data market forecasts to be realized, and for the European Union to fully benefit from the strategies presented in the report "Impact Assessment on enhancing the use of data in Europe," it is necessary that the technologies available to the IT world, and the search for innovative data privacy solutions, develop and be adopted in a synergistic manner.

Organizational Tools: partnerships, consortia, contracts

Paolo Bartoli, Ruggero Di Maulo

Registries supporting observational studies of rare diseases are important tools for a strategy aimed at accelerating medical research and the development of new therapies and solutions to improve patients' quality of life.

The involvement of patients through their Associations therefore plays an increasingly central, in fact it is now clear that these are at the heart of the entire implementation process and have briefly some key prerogatives, such as:

- They are the stakeholders with respect to the ultimate goal (new therapies or solutions);

[30] First Report on Facts and Figures Updating the European Data Market Study Monitoring Tool By International Data Corporation (IDC) and the Lisbon Council, European Data Market Study Updated SMART 2016/0063.

- They guard the instrument and the data collection;
- They are guarantors of the interests of the community for which they work;
- They are vehicles for engagement and sharing with participants, both physicians/researchers and patients.

The particular nature of registries in the head of an Association requires the ability to structure and maintain operational capacity over time, to engage patients and physicians, and to collect data while ensuring its security and availability. Such data are of increasing value to the ecosystem as they are the basis for the generation of real-world evidence (RWE), which is increasingly required by regulators and companies in both the early stages of drug research and subsequent registration, pricing, and post-marketing surveillance activities.

These elements rely on "industrial" and administrative management capacity, the greater the underlying technological and regulatory complexity of a registry that wants to survive over time and be in compliance with legal regulations. Therefore, economic and human resources are required that are not easy to obtain and maintain over time.

Such management capacity cannot be provided outside of an enterprise approach. The structured organization of human and material resources for a defined purpose offers these guarantees, provided it can sustain itself financially. Preparation and experience, safety, quality are actors that have a high cost, and must come together in a model where roles and responsibilities are clear.

Another key aspect is an approach aimed at the medium to long term, as it is not uncommon in this context for research results to lead to concrete solutions for subsequent generations of patients who have participated in the early-stage research. A "project-based" approach, with teams organized on an extemporaneous basis and timelines of 1 to 3 years is effective for a clinical product trial, for example, but not so for a long-term observational study, where a permanent organization for data quality collection and management on the one hand, and for patient and clinician engagement on the other, must be implemented.
Also from the point of view of economic sustainability, for a 1-3 year project a grant may be a suitable tool, whereas for a long-term observational registry, funds must be secured funds in a structural way (industry model).

Cloud-R aims to structure processes and make them organizationally and technologically operational by applying a proven industrialization model, and most importantly, it can finance from the beginning the implementation and maintenance of the registry. However, this capacity is possible if, and only if - having achieved the goal of quality data collection - these data, through their proper anonymization, can then be shared for secondary research purposes with other researchers and industries, for a fee, to remunerate the cloud-R business activity, a necessary condition to sustain the system in the medium to long term.

In the practice, the ability to bear the costs of the registry for the benefit of the Association and researchers is based on Cloud-R's ability to bear the business risk of registry implementation due to the exclusive availability of the anonymized registry data and the consequent ability to create, organize, and share such data for secondary research purposes, thus remunerating its investment.

The sharing of secondary data for Cloud-R has business value, based on an ethical purpose that is summed up in making knowledge and information available to other stakeholders even outside the single context in which the data are collected (see EMA recommendations) and at the same time to raise new financial resources, which are not available today for such independent research.

For its part, the Association is free to use the registry data for the primary purposes of research, typically on a non-profit basis, and in accordance with physicians' plans as outlined in the observational study protocol served by the registry. These are activities within the powers of the Association, for which it must nonetheless equip itself with the governing bodies and the minimum essential legal, administrative and regulatory expertise.

It is therefore important to mutually recognize the clear distinction between the nonprofit activities for-profit activities that gravitate to the Association and business activities characterized by the necessity inescapable financial balance and risk remuneration in the hands of Cloud-R.

- Primary research purposes: not-for-profit:

    - Association
    - Medical Doctors

The Association works in collaboration with physicians and referral centers by regulating relationships through agreements signed with the health facilities where the physicians work, and which implement what is stated in the protocol for aspects related to data collection and its use for publications and in general for primary uses.

- Secondary research purposes: for profit and social impact

    - Cloud-R

Extended knowledge sharing based on anonymized data, financial balance and risk reward.

The Association uses Cloud-R for all IT and technical activities as well as compliance aspects (from privacy to both process and facility security) for the purposes of data collection and registry management.

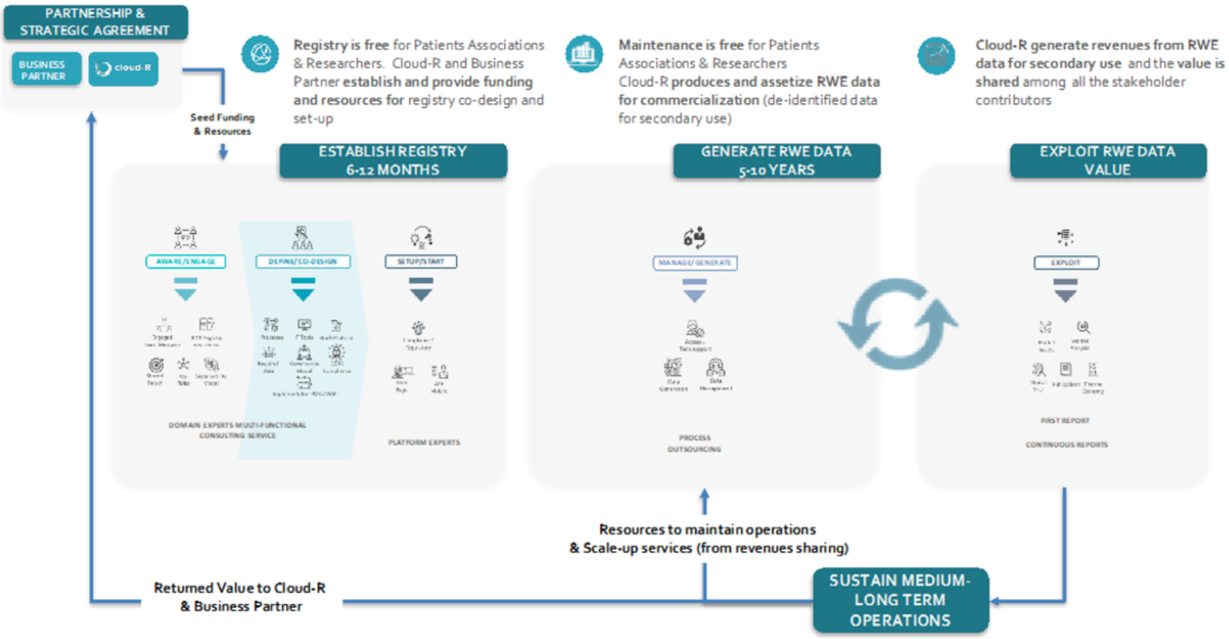
Cloud-R's ability to have the data anonymized allows the supply chain to be covered for both structure and process IT costs, as well as potentially covering all data entry and data monitor costs, which are usually covered by the participating centers and the promoter and allows the registry to be maintained in the medium to long term.

This distinction responds well to the different nature of the Association and Cloud-R, and corresponds to their respective corporate purposes.

The above premises form the basis of the fundamental clauses of the service contract underlying the implementation of the registry, and are signed by Cloud-R and the Association, which , as a rule, is the Promoter of the observational study.

The model can be viewed in its entirety through the following infographic:

Cloud-R CIRCULAR MODEL FOR LONG-TERM REGISTRY SUSTAINABILITY



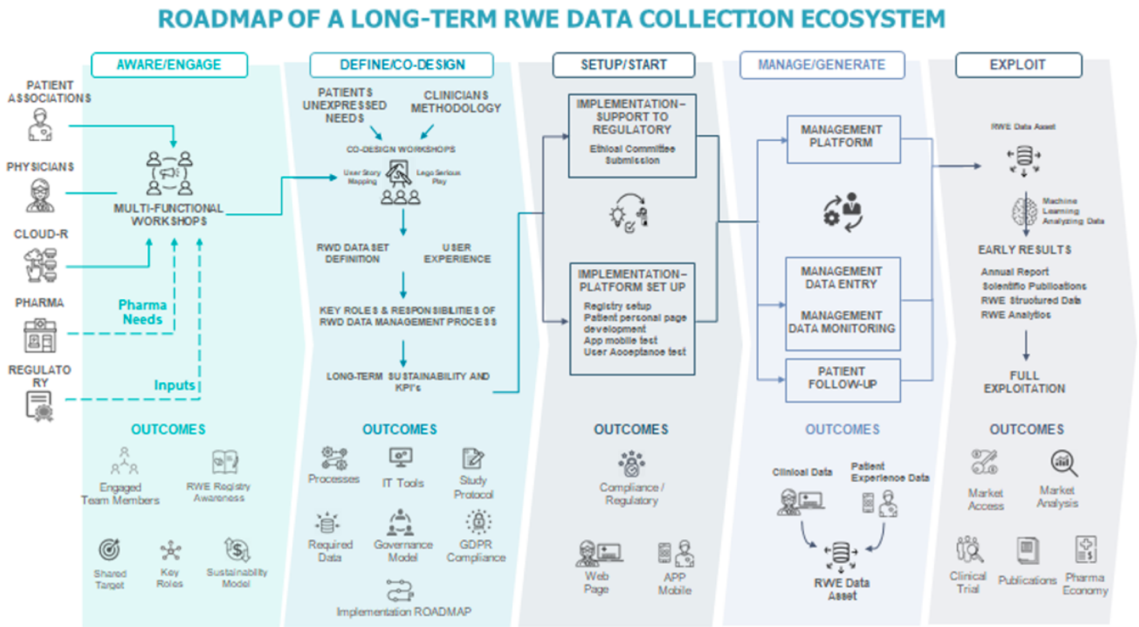
Registry is free for Patients Associations & Researchers. Cloud-R and Business Partner establish and provide funding and resources for registry co-design and set-up

Maintenance is free for Patients Associations & Researchers
Cloud-R produces and assetize RWE data for commercialization (de-identified data for secondary use)

Cloud-R generate revenues from RWE data for secondary use and the value is shared among all the stakeholder contributors

PARTNERSHIP & STRATEGIC AGREEMENT
BUSINESS PARTNER
cloud-R
Seed Funding & Resources
ESTABLISH REGISTRY 6-12 MONTHS
AWARE/ENGAGE
DEFINE/CO-DESIGN
SETUP/START
DOMAIN EXPERTS MULTI-FUNCTIONAL CONSULTING SERVICE
PLATFORM EXPERTS
GENERATE RWE DATA 5-10 YEARS
MANAGE/GENERATE
PROCESS OUTSOURCING
EXPLOIT RWE DATA VALUE
EXPLOIT
FIRST REPORT
CONTINUOUS REPORTS
Resources to maintain operations & Scale-up services (from revenues sharing)
Returned Value to Cloud-R & Business Partner
SUSTAIN MEDIUM-LONG TERM OPERATIONS

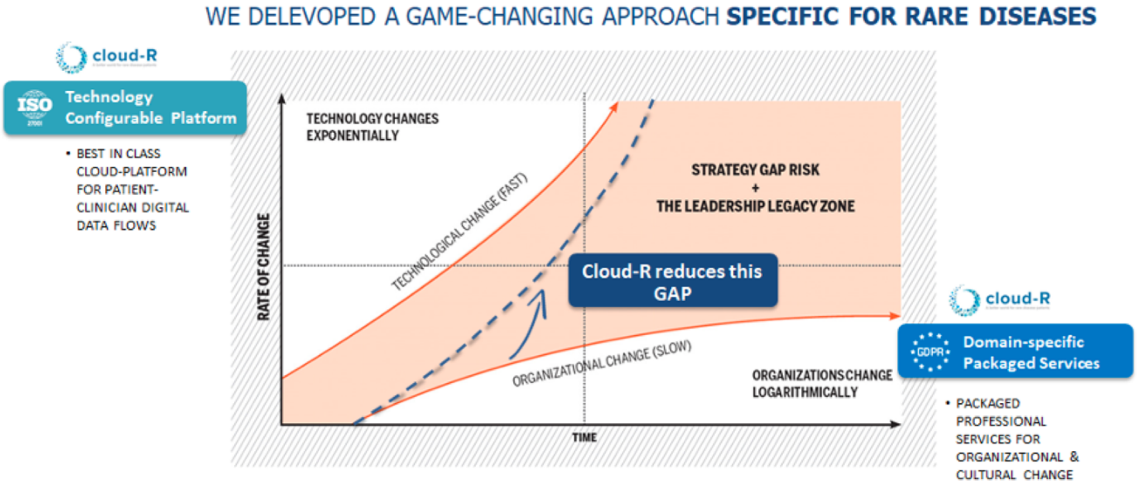
The generation of usable data for the market can be seen in more detail in this infographic:



RWE data in the case of rare diseases: how to extract new value from it

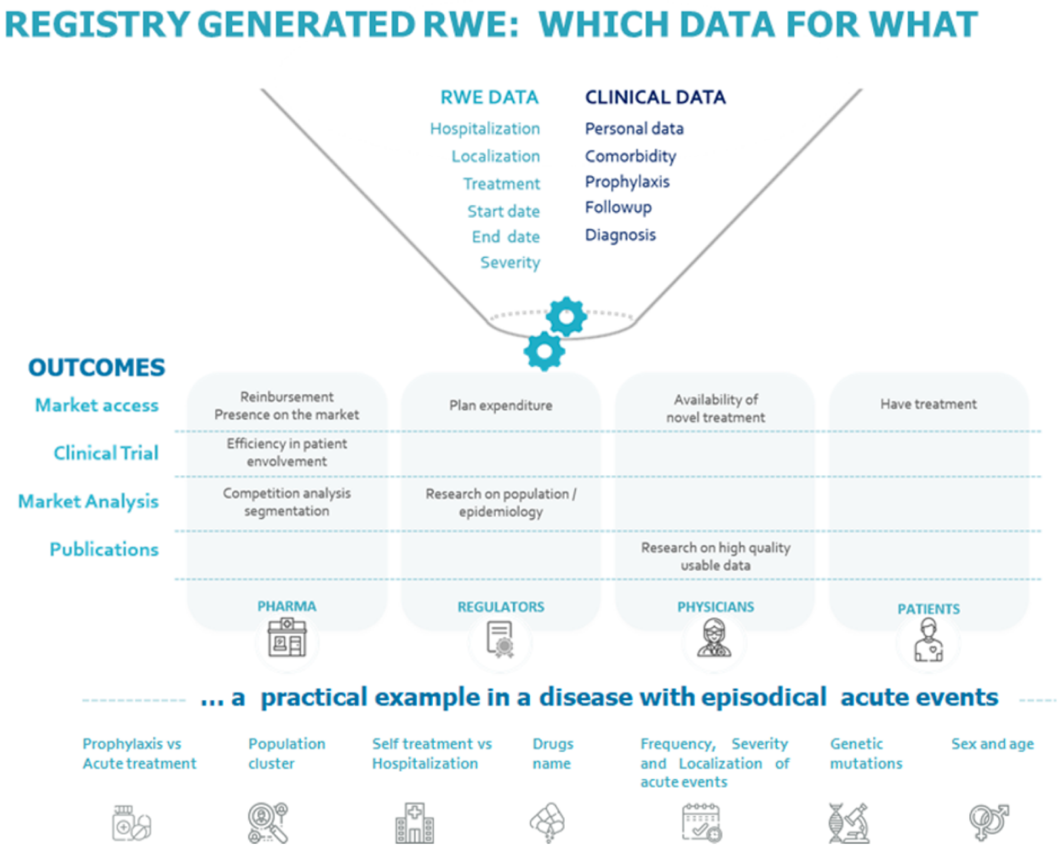
Paolo Bartoli, Ruggero Di Maulo

The model that is being carried out by Cloud-R has the capacity to revolutionize data collection in rare diseases. In fact, it intervenes on the governance phase of processes that are traditionally managed in an unstructured way, not integrated with the technological and IT phases, thus creating the breaking points that often then manifest themselves in the defaults under the compliance profile, in the unavailability of data and in the dissolution of the study organization itself. The real weakness of these projects lies in the different speed between the technological and organizational evolution - the human factor - processes, between the hard and soft parts of the necessary skills/culture and attitudes, which instead must be harmonized to lead to lasting results.


WE DELEVOPED A GAME-CHANGING APPROACH SPECIFIC FOR RARE DISEASES
cloud-R
ISO
Technology Configurable Platform
• BEST IN CLASS CLOUD-PLATFORM FOR PATIENT-CLINICIAN DIGITAL DATA FLOWS
RATE OF CHANGE
TECHNOLOGY CHANGES EXPONENTIALLY
TECHNOLOGICAL CHANGE (FAST)
STRATEGY GAP RISK
+
THE LEADERSHIP LEGACY ZONE
Cloud-R reduces this GAP
ORGANIZATIONAL CHANGE (SLOW)
ORGANIZATIONS CHANGE LOGARITHMICALLY
TIME
cloud-R
GDPR
Domain-specific Packaged Services
• PACKAGED PROFESSIONAL SERVICES FOR ORGANIZATIONAL & CULTURAL CHANGE

The focus of Cloud-R's action is therefore in cultural change, in the development of a collective self-awareness - mindfulness with a term in vogue today - of the stakeholders of the entire ecosystem, who never more than now are called upon to seize the opportunities given by the power of digital while at the same time managing the growing complexities that come with digitization.

The use cases can be many, such as:


REGISTRY GENERATED RWE: WHICH DATA FOR WHAT
RWE DATA
Hospitalization
Localization
Treatment
Start date
End date
Severity
CLINICAL DATA
Personal data
Comorbidity
Prophylaxis
Followup
Diagnosis
OUTCOMES
Market access
Clinical Trial
Market Analysis
Publications
Reinbursement Presence on the market
Efficiency in patient envolvement
Competition analysis segmentation
Plan expenditure
Research on population / epidemiology
Availability of novel treatment
Research on high quality usable data
Have treatment
PHARMA
REGULATORS
PHYSICIANS
PATIENTS
... a practical example in a disease with episodical acute events
Prophylaxis vs Acute treatment
Population cluster
Self treatment vs Hospitalization
Drugs name
Frequency, Severity and Localization of acute events
Genetic mutations
Sex and age

Technologies for anonymization

The choice of technologies for anonymization

Daniele Panfilo

Access to and reuse of health data are often hindered by various factors. A major cause of difficulty is the fact that many of the health data are classified as sensitive data. The personal information contained within this data, if improperly disclosed, could cause a serious breach of the right to privacy. The risks associated with the use of health information also impact various dimensions from security to reputational. This requires that the person in charge of managing the health information asset adhere to the highest standards aimed at protecting the privacy of individuals. This will go a long way toward encouraging research and innovation while ensuring the security and rights of individuals.

From a technical perspective, there are several approaches to protecting the privacy of individuals ranging from standard anonymization and pseudo-anonymization solutions to modern data synthesization. The choice of one or the other technique depends on the degree of security one desired to achieve and the type of intended use for the data.

- **Anonymization:**
  Anonymization is that process of data obfuscation that irreversibly involves the total removal of the identifying element.
  There are multiple anonymization techniques, and a first macro distinction is between randomization and generalization techniques.

  - **Randomization:** operates on the degree of truth of the data to undermine the correlation that exists between the same and the person. The main randomization techniques fall under:

    - Noise addition: you go to add noise on certain columns with the goal of decreasing the accuracy of the information while still trying to keep the distribution unchanged.
    - Permutation: this is done by mixing the values of some attributes so that they turn out to be related to different people or entities. However, if there are strong logical relationships between some attributes the effect could be easily reversed (example treating physician and hospital department). This operation aims to break certain correlations between attributes that would facilitate identification of subjects. Nevertheless, this technique ensures that the marginal distributions of the attributes remain unchanged.

  - o <u>Differential privacy:</u> this technique is in principle similar to noise addition. The main difference lies in the fact that the latter involves "a priori" noise insertion. In contrast in differential privacy, the addition of noise occurs "on the fly" at the time of the execution of the database query. Thus, the query result has an appropriate amount of noise and can be shared with third parties since, if properly implemented, this technique does not allow easy re-identification of subjects. However, the true data remains available to the data controller.

- **Generalization/Aggregation:** aims to group individual records into classes containing multiple subjects to eliminate the possibility of point identification. This can be achieved by changing the scale of an attribute. For example, if we had a city column we could replace it with region by enlarging/diluting the information by enlarging the search set.

  - o <u>K-anonymity:</u> Generalization techniques ensure the anonymity of people through their grouping into sets with other k-people. The idea behind k-anonymity techniques is to replace the point value of identifiers with ranges of values that include at least other k-subjects. For example, one replaces precise dosage values of a drug with ranges of ranges of values.
  - o <u>L-diversity/T-vicinity:</u> extends and improves k-anonymity by requiring that within each value range with k subjects there exist at least L-different values. Strengthens the concept of k-anonymity against attacks by inference.

- **Pseudoanonymization:**

Pseudonymization aims to replace the identifying attributes of a piece of data with other values that do not allow for subject identification.
Thus, the goal of pseudonymization is to reduce the possibility of correlation of a data set to the original identity of the data subjects. Such data transformation, unlike the case of anonymization, is often a reversible process.
In essence, it is done by separating the data into direct identifiers (e.g., Social Security number) that allow easy identification of the subject, and which must be encrypted or masked, and indirect identifiers (e.g., place of birth) that can be shared instead without pre-processing steps.
The main pseudonymization techniques include:

- **Hashing functions:** this is a noninvertible function that takes as input an attribute of arbitrary length and returns a string of predefined length. Although it is not invertible, if the nature of the input attribute is known and if it has a finite size(e.g., acronyms of Italian provinces), the function allows the hashing result to be reproduced by simply leaving the input again.

- **Hashing functions and salt:** is an improved version of the classic hashing function that to limit the possible reidentification of subjects adds to the original input data a random value called salt that can be known.
- **Hash encrypted with stored key:** very similar to hash with salt. In this case the salt is a private key known only to the controller.
- **Secret key encryption:** is a reversible operation in which the original data is transformed using encryption techniques based on secret keys secret keys.

  The main distinction in pseudo-anonymization techniques is between. symmetrical and asymmetrical pseudo-anonymization:

  - Symmetric: in this case, the encryption and decryption key coincide;
  - Asymmetric: in this case, one key is used to encrypt and another separate key is used to decrypt the data, making it unnecessary to share the encryption key.

- **Tokenization:** involves assigning a randomly generated value to each instance of the attribute we intend to pseudonymize. Obviously, the mapping must take care that random numbers are not assigned same to different instances to avoid confusion.

- **Synthesization:**

Unlike previous anonymization and pseudonymization techniques, modern data synthesization solutions based on generative machine learning models represent an entirely new paradigm for personal data management.

These techniques assume that in most cases and applications, the individual record constitutes solely a liability while the real asset is the statistical content of the dataset.

Data synthesization by means of AI systems represents the new frontier in sensitive data management.

Modern synthetic data generation systems are tools capable of sampling new records from the input data distribution thus generating new data (synthetic).

Data generated through advanced artificial intelligence solutions are highly representative of the input statistical distribution, such that they can be used for training machine learning models or descriptive statistical analysis by exhibiting results that are statistically comparable to those obtainable with real data.

Given the artificial nature of the synthetic data obtained through generative models, the identification of real subjects or the possibility of Membership Inference Attacks (MIAs), in the absence of access to the parameters of the generating model or the real dataset and except for degenerate cases (datasets containing only a few units), is unlikely and much more complex when compared to the case of MIA on discriminative models[31].

[31] "An Overview of Privacy in Machine Learning." https://arxiv.org/abs/2005.08679. Accessed 17 Jun. 2021.

Aindo and synthetic data

Daniele Panfilo

Aindo Ltd. has developed a technology based generative machine learning models for producing synthetic data in a healthcare context. This technology starting from real patients enables the creation of artificial patients that exhibit the same statistical characteristics as the real population. The artificial patient, generated by means of AI models, exhibits the statistical characteristics of the real one, thus maintaining utility in analysis but preventing the reidentification of real subjects or the sharing of personal information.

The synthetic patients generated can be used for statistical analysis or training of machine learning models without ever exposing the actual data. Such technology is intended to facilitate medical research and development projects by greatly speeding up data access time by allowing sharing of the statistical asset without compromising patient privacy.

The choice of dataset: primary scientific stage, feasibility, stage of definition for secondary use.

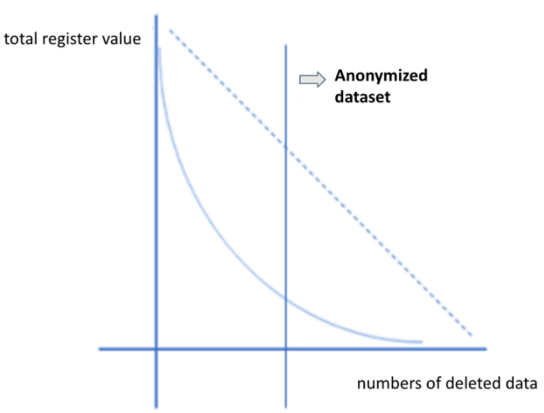
Paolo Bartoli, Ruggero Di Maulo

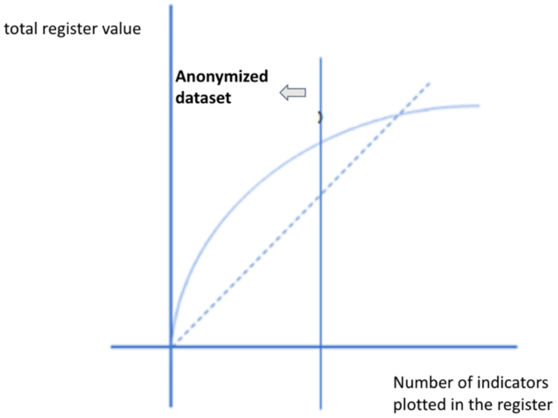
Datasets are defined by the study sponsor, with IT, cloud-R process, as well as regulatory/legal advice if needed.

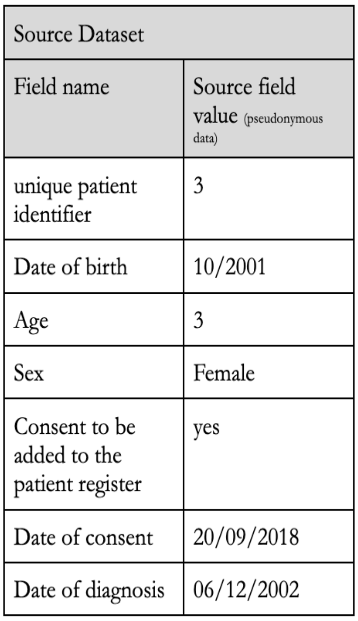
The anonymization techniques can be defined by the promoter as owner, or approved by it when proposed by cloud-R, which has the role of Data Processor and will have, by contract, the rights to use such anonymized data.

The parties' interests are consistent with each other and have the same functional goal: to make the data an autonomous object, devoid of reference to individuals.

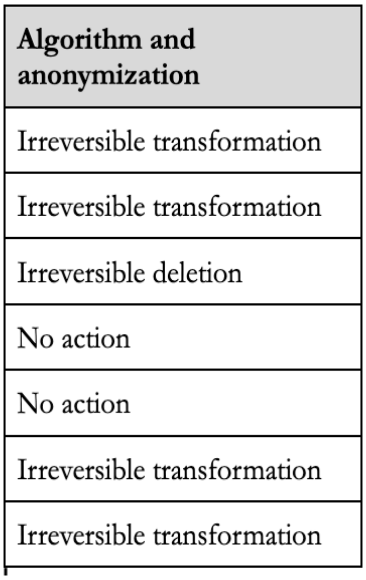
The limited amount of data on rare diseases (few patients) makes the choice of anonymization techniques difficult. One of these involves deletion of items that could lead to re-identification of the patient. The more data that are deleted or merged, the more difficult it is to to trace the patient's identity, the less information there will be.



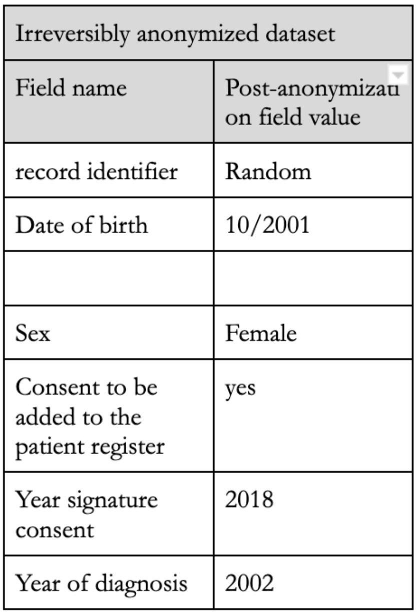




Source Dataset	
Field name	Source field value (pseudonymous data)
unique patient identifier	3
Date of birth	10/2001
Age	3
Sex	Female
Consent to be added to the patient register	yes
Date of consent	20/09/2018
Date of diagnosis	06/12/2002



Algorithm and anonymization
Irreversible transformation
Irreversible transformation
Irreversible deletion
No action
No action
Irreversible transformation
Irreversible transformation



Irreversibly anonymized dataset	
Field name	Post-anonymizau on field value
record identifier	Random
Date of birth	10/2001
Sex	Female
Consent to be added to the patient register	yes
Year signature consent	2018
Year of diagnosis	2002

There are several techniques that can be adopted, but they all pose the trade-off between data quality and anonymization, taking into account that the latter is not gradable. In addition, it may be necessary to consider the characters that define a piece of data as "anonymous." These could change due to digital technological evolution and AI, potential elements that could make recognizable tomorrow personal profiles underlying data now considered anonymous.

The possibility of bringing the dataset to a state of "essential" non relation to the concept of personal data can be experienced with the adoption of algorithms that generate a dataset - derived from the original - defined as synthetic, in which granular data are transformed into information entities different from the original, while retaining the ability to generate statistics superimposed on those generated by the source dataset containing personal references. This topic is the subject of discussion elsewhere.

Example of reference software architecture

Alexandru Raileanu

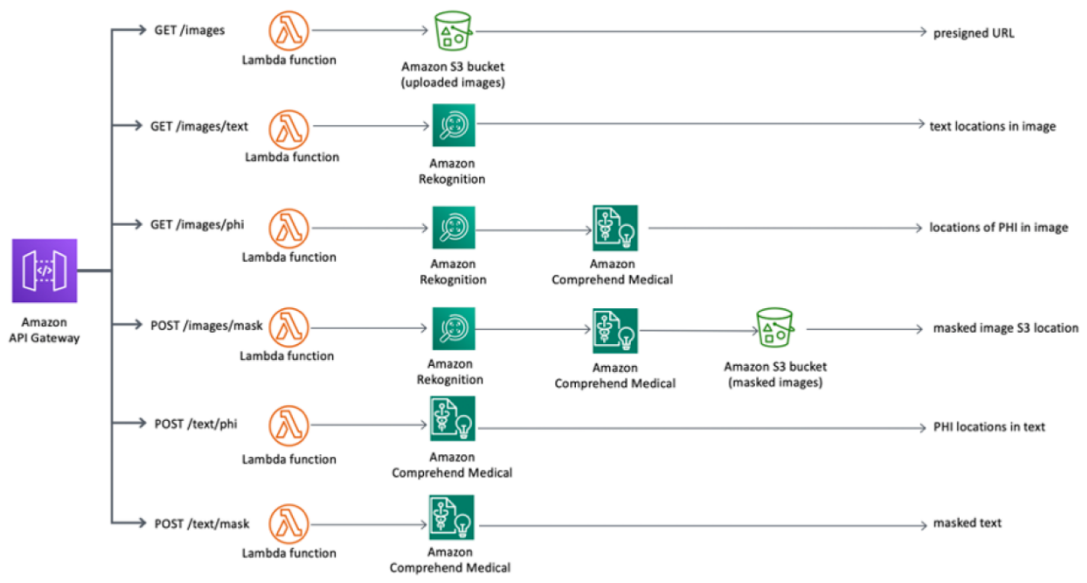
Healthcare and Life Science organizations generate large amounts of medical data such as RX or MRI images or patient information that are exchanged among various Software applications. A very common challenge for both medical personnel and developers designing systems to manage sensitive medical data is sharing data but at the same time complying with industry compliance rules (e.g., GDPR, PHI).

The solution presented below is an example of data masking architecture from which one can take a cue to build a more complex system customized to the needs of the case.

Microservices, which are nothing more than a paradigm for optimizing resources from a computational and cost perspective, provide for the creation of the functional logic for managing pre-processing, configuration, identification and finally for masking patient data.

The microservices interact with the managed Amazon Recognition service to identify text in an uploaded medical image and with Amazon Comprehend Medical to identify potentially sensitive information in the texts.

In addition, the template configures an Amazon Simple Storage Service bucket (flexible storage space) to store raw data and processed images, AWS CloudTrail to track data access (text or images), and Amazon CloudWatch logs for operational management. By default, the logs are encrypted, using the HTTPS protocol.



Conclusions and future prospects

The publication of the white paper "E-health Data Sharing" marks the conclusion of the first phase of a journey and the starting point for building a multi-stakeholder discussion, involving institutions, academia and the private sector. More specifically, as Data Valley we will continue to collect and integrate new contributions to enrich the first version of the white paper and create complementary tools to this document, such as checklists and toolkits, so to highlight the most relevant elements.

We invite those interested in sharing further experiences and models for data sharing in the health sector to write to us at info@datavalley.it