Big Health and Genetic Data:

Opportunities, Risks and Legal Conundrums regarding the application of GDPR

Ioannis Iglezakis/Theodoros Trokanas**/Panagiota Kiortsi***

Abstract: This article explores the major legal issues arising from the collection and processing of Big Health Data in light of the General Data Protection Regulation (EU Regulation 679/2016). It defines the concepts of big data, in general and big health and genetic data, in more particular, and their regulation by the GDPR. Then, it deals with the issue whether big health data are personal data and fall within the field of application of the GDPR. Subsequently, it applies the principles relating to data processing as regards Big Health and Genetic Data and the legal grounds which justify their processing by both public and private entities. Last but not least, it focuses on the mitigation of risks concerning data subject's rights while embracing the opportunities which Big Health Data has to offer by providing safeguards. In this respect, this article highlights the significance of DPIAs and privacy by design in the context of Big Health and Genetic Data processing.

Keywords: Big data, health data, genetic data, sources, stakeholders, GDPR, processing principles, legal bases, DPIA.

I. Introduction

The processing of Big Health and Genetic Data can bring significant benefits to medical research: it can improve decision-making in health care provision, e.g., development of personalized medicine, identification of new biomarkers to predict the emerging of complex or rare diseases, which would be impossible with an analysis of individual or small scale datasets, and help shape or reform healthcare policies; it can ameliorate public health monitoring, including disease outbreaks or disease spread prevention (epidemiology); it can curb the cost of healthcare (e.g. reduced patient readmission rates to hospitals), record and prevent side effects, and even acknowledge and extenuate medical errors.[1]

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

* Professor, Law Faculty, Aristotle University, email: iingleza@law.auth.gr.
**Researcher, Law Faculty, Aristotle University, email: trokanas@mycosmos.gr.
*** Researcher, Law Faculty, Aristotle University, email: kiortsip@gmail.com.

However, thanks to advanced processing techniques and analytics the ever-growing big data sets provide unprecedented insight into human behaviour, private life and our societies.[2] The endless and unpredictable uses and results of datasets[3] may compromise fundamental rights, such as the right to privacy, the right to the protection of personal data, the right to non-discrimination, which are all enshrined in the European Charter of Fundamental Rights.[4] Algorithmic systems could generate unlawful and harmful discrimination towards vulnerable groups. For instance, algorithms can associate or classify sexual transmitted diseases with social characteristics, ending up in the stigmatization of entire groups or the exploitation of their vulnerabilities[5]. Finally, the exponential growth of big datasets may upset the balance of power between citizens, governments, and private actors, resulting, inter alia, in the consolidation of monopolies and abusive practices on the market.[6]

Consequently, this paper will analyse whether the GDPR provides interpretative solutions which guarantee that risks deriving from Big Health and Genetic Data will be addressed, while data subjects' rights are enforced; at the same time, it should be ensured that the opportunities that Big Health Data have to offer will not be missed.

II. Big Data in general

Big Data has been hailed as a "socio-technical phenomenon", However, surprisingly, even today, the concept constitutes unchartered territory.[7] At a European level, a first attempt to describe the term was made in 2013. Article 29 Data Protection Working Group in its Opinion on purpose limitation[8] referred to gigantic digital datasets held by corporations, governments, and other large organisations, which are then extensively analysed with

[1] Hellenic Republic National Bioethics Commission, Report, Big Data in Health, p 7, available at: http://www.bioethics.gr/images/pdf/GNOMES/REPORT_Big_Data_FINAL_.pdf.

[2] European Parliament, Committee on Civil Liberties, Justice and Home Affairs, Report on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI)), 20 February 2017, Under C. It is argued that this transformation will signal the beginning of a new era named "surveillance capitalism" (See Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, PublicAffairs, 2019).

[3] Effie Vayena, Protecting Health Privacy in the World of Big Data, in: I. G. Cohen/H. F. Lynch/E. Vayena/U. Gasser (ed.), *Big Data, Health Law and Bioethics*, Cambridge University Press, 2018, p. 159.

[4] European Parliament, Committee on Civil Liberties, Justice and Home Affairs, Report on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI)), 20 February 2017, Under I.

[5] Favaretto, M., De Clercq, E. & Elger, B.S. Big Data and discrimination: perils, promises and solutions. A systematic review. J Big Data 6, 12 (2019). https://doi.org/10.1186/s40537-019-0177-4

[6] European Parliament, Committee on Civil Liberties, Justice and Home Affairs, Report on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI)), 20 February 2017, Under K.

[7] M. Oostveen, "Identifiability and the applicability of data protection to big data", International Data Privacy Law 6(4) (2016), p 2.

[8] Article 29 Data Protection Working Group, Opinion 03/2013 on purpose limitation, adopted on 2 April 2013.

computer algorithms.[9] The WP 29 also focused on the fact that Big Data can be used to identify general trends and correlations, but it can also directly affect individuals[10]. In 2017 a European Parliament Report[11] defined Big Data as "the collection, analysis and the recurring accumulation of large amounts of data, including personal data, from a variety of sources, which are subject to automatic processing by computer algorithms and advanced data-processing techniques using both stored and streamed data, in order to generate certain correlations, trends and patterns".[12]

Due to a lack of consensus on a definition Big Data are usually described by their main characteristics.[13] Big Data feature the so-called "4Vs +1": *Volume, Velocity, Variety, Veracity and Value.*[14] In other words, they are characterized by the large volume of information (whether structured or not), the high velocity at which the data are collected and analysed, the increased complexity and variety of data arriving at different formats, their reliability and the worth produced by their analysis. It is interesting to note that the term 'Big' data is generally understood both in quantitative and procedural terms: it denotes the electronic size of datasets and simultaneously the big computational or human effort it takes to analyse them.[15] Moreover, the term 'big' is dynamic in character, as it is conditioned by the advancement level of computing technologies.[16] To put it differently, what is characterised 'big' today might not to be so in one year or in a decade.[17] Finally, as regards the subcategory of genetic data it is correctly underlined that Genomics is an inherently Big Data science.[18] A

[9] Article 29 Data Protection Working Group, Opinion 03/2013 on purpose limitation, adopted on 2 April 2013, Under III.2.5. and Annex 2.

[10] Ibid.

[11] European Parliament, Committee on Civil Liberties, Justice and Home Affairs, Report on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI)), 20 February 2017.

[12] Ibid, Under A.

[13] Effy Vayena/Urs Gasser, "Strictly Biomedical? Sketching the Ethics of the Big Data Ecosystem in Biomedicine", In B. D. Mittelstadt/L. Floridi (ed.), The Ethics of Biomedical Data, Law Governance and Technology Series, Volume 29, Springer, 2016, p 18.

[14] Hellenic Republic National Bioethics Commission, Report, Big Data in Health, p 4, available at http://www.bioethics.gr/images/pdf/GNOMES/REPORT_Big_Data_FINAL_.pdf.See also Maria Tzanou, The GDPR and (Big) Health Data: Assessing the EU Legislator's Choice, in: Tzanou (ed.), *Health Data Privacy under the GDPR*, 2021, p 4. For other authors just "3Vs": Volume-Variety-Velocity (I. G. Cohen/H. F. Lynch/E. Vayena/U. Gasser, Introduction, In I. G. Cohen/H. F. Lynch/E. Vayena/U. Gasser (ed.), *Big Data, Health Law, and Bioethics*, Cambridge University Press 2018, p. 1).

[15] B. D. Mittelstadt/L. Floridi, Introduction, In B. D. Mittelstadt/L. Floridi (ed.), *The Ethics of Biomedical Data*, *Law Governance and Technology Series*, Volume 29, Springer, 2016, p 2.

[16] Ibid.

[17] Ibid.

[18] Stephens ZD, Lee SY, Faghri F, CampbellRH, Zhai C, Efron MJ, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195, p 1.

quintessential example is that even the genome sequence of one single person could be considered big data.[19]

Big Data analysis is often illustrated as a three-phase model, which in simplified terms comprises a collection, an analysis and an application phase.[20] The first phase (collection) involves amassing personal or non-personal data, and often a combination of both types, that may be further analysed /processed.[21] In this phase data are obtained in various ways: direct collection from the data subject, purchase from third parties (i.e. data brokers), harvest from publicly available data sources. The second phase (analysis) encompasses both storage and further processing of the data collected. Apart from pre-processing techniques this phase features data mining techniques, i.e. the discovery of useful patterns in large data sets, building on statistics, machine learning and artificial intelligence. Based on the analysis phase outcomes, the third phase (application) focuses on reaching decisions, either general or individual ones, and it is either automated or performed by individuals.

III. From Health and Genetic Data to Big Health and Genetic Data

Personal data concerning health[22] are all data pertaining to the health status of a data subject, including the provision of health care services, which reveal information relating to the past, current or future physical or mental health status of the data subject.[23] Health data are pieces of information, originating from testing or examining a body part or bodily substance, including genetic data and biological samples;[24] for example, any information on a disease, disability, disease risk, physical or mental medical history, clinical treatment or the physiological or biomedical state of the data subject, irrespective of its source (physician or other health professional, hospital, medical device or an in vitro diagnostic test)[25]. Personal data concerning health fall under the exhaustive list of special categories of personal data, for which Article 9 GDPR establishes a blanket principle of prohibition of processing.

Before the regulation of genetic data in the GDPR, these did not constitute a standalone subcategory of sensitive data. To fill this legislative gap, the Article 29 Data

[19] K. Pormeister, The GDPR and Big Data: Leading the Way for Big Genetic Data?, In: Schweighofer E., Leitold H., Mitrakas A., Rannenberg K. (eds) *Privacy Technologies and Policy. APF 2017. Lecture Notes in Computer Science*, vol 10518, Springer, 2017, p 13.

[20] M. Oostveen, op. cit., p 2.

[21] Ibid., p 3.

[22] The term "health data" or "data pertaining to the health status of a data subject" is preferable to "medical data", since the former has a broader scope than the latter (see in more detail W29 Working Party, Letter, Annex, Health Data in Apps and Devices, 5 February 2015).

[23] GDPR Art. 4 (15) and Recital 35. Directive 2016/680, Art. 3 (14) and Recital 24. Regulation 2018/1725, Art. 3 (19).

[24] GDPR Recital 35. Directive 2016/680, Recital 24.

[25] GDPR Recital 35. Directive 2016/680, Recital 24.

Protection Working Party[26] classified genetic data into the subcategory of "health data" of Article 8 (1) of Directive 95/46.[27] By contrast, due to the genetic data explosion[28] witnessed in about 20 years that elapsed between the implementation of Directive 95/46 and the adoption of the GDPR, the EU legislator explicitly defined genetic data as a subcategory of "special categories of personal data" in Article 9 GDPR.

Article 4 (13) GDPR defines genetic data as personal data relating to the inherited or acquired genetic characteristics of a natural person and data that result from the processing of biological samples. In Recital 34 it is stated, more precisely, that genetic data in particular result from chromosomal, deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) analysis or from the analysis of another element enabling equivalent information to be obtained.

The EU data protection reform, which led to the adoption of the GDPR, has been hailed as "an enabler for Big Data services in Europe"[29]. Regrettably, GDPR lacks a straightforward legal definition of Big Health Data or Big Genetic Data. Only GDPR Recital 157 mentions them. Specifically, it stipulates that researchers can obtain new knowledge of great value regarding widespread medical conditions (such as cardiovascular disease, cancer, and depression). The above-mentioned information, which is obtained through registries "provides solid, high-quality knowledge, which can form the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for several people and enhance the efficiency of social services".

IV. Big Health Data Sources

Big Health Data can be derived from various sources. A primary source is clinical and administrative health records, especially electronic health records (EHRs).[30] They incorporate medical histories, regular doctor visits and emergency department visits, therapeutic schemes, electronic prescription records, social security institutional records and insurance claims. Another major source is scientific/research databases and registers, which comprise clinical and laboratory data either open-access or closed (biobanks).

---

[26] The Working Party, "ARTICLE 29 Date Protection Working Party", Working Document on Genetic Data, 12178/03/EN WP 91, adopted on 17 March 2004, part III. Cf Section 1180(a)(1) of the United States Genetic Information Non-Discrimination Act (GINA) (enacted by the 110th United States Congress, 21 May 2008, 122 Stat 881): "Genetic information shall be treated as health information". See also Taylor M., Genetic Data and the Law A Critical Perspective on Privacy Protection, Cambridge University Press, 2012, p. 70.

[27] Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

[28] Or according to some authors "a coming genomic data flood" (I. G. Cohen/H. F. Lynch/E. Vayena/U. Gasser, Introduction, In I. G. Cohen/H. F. Lynch/E. Vayena/U. Gasser (ed.), Big Data, Health Law, and Bioethics, Cambridge University Press 2018, p. 2)

[29] European Commission, The EU Data Protection Reform and Big Data, Factsheet, January 2016.

[30] S. E. Malanga/J. D. Loe/C. T. Robertson/K. S. Ramos, Who's Left Out of Big Data? How Big Data Collection, Analysis, and Use Neglect Populations Most in Need of Medical and Public Health Research and Interventions, In I. G. Cohen/H. F. Lynch/E. Vayena/U. Gasser (ed.), *Big Data, Health Law, and Bioethics*, Cambridge University Press 2018, p. 102 et seq.

Increasingly, Big Health Data are being derived from non-biomedical sources. General Internet use (Internet search terms, webpage visits and clicks, movie ratings, online purchases etc.), social network use (whole profile structure, 'likes', posts, emotions, videos in Facebook, Twitter or Instagram) and mobile device use can disclose individual preferences, opinions and health status.[31] Surprisingly, non-medical Big Data have high biomedical value.[32] For instance, it was reported[33] that Target[34] identified and analysed together a group of purchased products to assign each shopper a 'pregnancy prediction' score, so that the company could send coupons timed to different stages of pregnancy. In other cases, researchers could predict, track and map obesity rates in a neighbourhood by means of Facebook 'likes',[35] and also identify the severity of depression symptoms based on daily locations of mobile phone users and their total time spent on their mobile devices.[36]

Admittedly, the most conspicuous example of Big Health Data in action was Google Flu Trends.[37] This model was launched in 2008 to help detect early seasonal influenza outbreaks among the general population worldwide by gathering and analyzing Google search engine queries about flu-related topics.[38] Similarly, during the outbreak of Ebola crisis in 2014-2015 mobile phone data collected in affected regions were exploited for contact tracing and for public health surveillance detecting human mobility.[39] In parallel, there exist even more specialized social network platforms, such as PatientsLikeMe[40], where users exchange sensitive information about medical conditions they experience.

In addition, broadband-enabled digital tools, generically called ingestibles, wearables and implantables (or alternatively embeddables), are a treasure trove of Big Health Data.[41] Ingestibles are edible digital tools (e.g., "smart pills"), which help track blood levels of medications in a patient's body or monitor his or her internal reactions to them, or pill-shaped

[31] Ibid, p. 101.

[32] Effy Vayena/Urs Gasser, "Strictly Biomedical? Sketching the Ethics of the Big Data Ecosystem in Biomedicine", In B. D. Mittelstadt/L. Floridi (ed.), The Ethics of Biomedical Data, Law Governance and Technology Series, Volume 29, Springer, 2016, p 19.

[33] See Forbes, Kashmir Hill, How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did, 16 February 2012. See also Charles Duhigg, The New York Times Magazine, How Companies Learn Your Secrets, 16 February 2012.

[34] www.target.com

[35] Carisa Véliz, Medical Privacy and Big Data, In Andelka M Phillips/Thana C de Campos/Jonathan Herring (edited by), Philosophical Foundations of Medical Law, Oxford University Press, 2019, p 309.

[36] Effy Vayena/Urs Gasser, "Strictly Biomedical? Sketching the Ethics of the Big Data Ecosystem in Biomedicine", In B. D. Mittelstadt/L. Floridi (ed.), The Ethics of Biomedical Data, Law Governance and Technology Series, Volume 29, Springer, 2016, p 26.

[37] Google Flu Trends, as a website open to the general public, has shut down.

[38] See indicatively J. Ginsberg/M. H. Mohebbi/R. S. Patel/L. Brammer/M. S. Smolinski/L. Brilliant, Detecting influenza epidemics using search engine query data, Nature, Vol 457, 19 February 2009, p. 1012-1014. See also D. Butler, When Google got flu wrong, Nature, Vol 494, 14 February 2013, p. 155.

[39] Effy Vayena/Urs Gasser, "Strictly Biomedical? Sketching the Ethics of the Big Data Ecosystem in Biomedicine", In B. D. Mittelstadt/L. Floridi (ed.), The Ethics of Biomedical Data, Law Governance and Technology Series, Volume 29, Springer, 2016, p 26.

[40] www.patientslikeme.com.

[41] S. E. Malanga/J. D. Loe/C. T. Robertson/K. S. Ramos, op. cit., p. 101.

video cameras, which are meant to replace conventional diagnostic procedures. Wearables are devices (such as smartwatches, smart wristbands, patches, contact lenses, advanced textiles, etc.) intended to monitor vital signs (such as body temperature, blood pressure, heart rate, glucose levels, brain activity, muscle motion, etc.) through skin contact and to transmit such data wirelessly to smartphones, with or without the wearer's knowledge. Wearables can serve either as activity (or alternatively fitness) trackers (e.g., for measuring calorie consumption, sweat rate, distance waked or steps made, etc.) or for medical purposes (e.g. for long-term monitoring of patients). One real example of the former is Fitbit[42], whereas of the latter are ResearchKit and Carekit.[43] Both ResearchKit and Carekit were developed by Apple as open-source software development framework, on which other medical apps for the iPhone and Applewatch will be built. To be clear, ResearchKit was specifically designed to facilitate medical research and clinical trials, whereas CareKit was built for patients to self-manage their ongoing medical conditions. And last but not least, implantables (or embeddables) are miniature devices which are inserted under the skin or deeper into the body (e.g. a heart pacemaker). It is noteworthy that all these gadgets collect health data, even though some of them have not been labelled as medical devices.[44]

V. Traditional and novel stakeholders

Information technology revolution has had a far-reaching impact on the collection and processing of health-related data. Traditionally, health data were collected and processed by health practitioners, hospitals, pharmaceutical companies, pharmacies, academic researchers, insurance firms, all of which could be aptly described as trustworthy entities, of a conservative nature, tightly regulated or at least committed to codes of ethics.[45] States could exert regulatory control over them[46] to ensure that the analysis and use of health data will not pose serious threats for data subjects. Nevertheless, the rapid expansion of information technologies has been exciting the interest of various stakeholders in individuals' health data. The so-called data brokers[47] rush to fulfil this insatiable desire of interested parties for all sorts of data.

[42] www.fitbit.com.

[43] www.researchandcare.org.

[44] Art. 29 Working Party, Letter, Annex-health data in apps and devices, 5 February 2015, Under Defining health data.

[45] T. Z. Zarsky, Correlation versus Causation in Health-Related Big Data Analysis, The role of Reason and Regulation, In: I. G. Cohen/H. F. Lynch/E. Vayena/U. Gasser (ed.), *Big Data, Health Law and Bioethics*, Cambridge University Press, 2018, p 44.

[46] Ibid., p 45.

[47] Companies like Acxiom, Experian, Nielsen, Oracle, Salesforce and others reportedly provide one-stop shopping for hundreds of different data (Financial Times, Aliya Ram/Madhumita Murgia, Data brokers: regulators try to rein in the 'privacy deathstars', 8 January 2019, available at: https://www.ft.com/content/f1590694-fe68-11e8-aebf-99e208d3e521.

Employers take a keen interest in learning health information about their prospective employees.[48] That is because, being able to determine and to reject candidate employees who are prone to certain diseases, they may enhance their productivity and reduce healthcare costs. Financial and banking institutions would eagerly collect and process health data of potential clients, in an attempt to screen out applicants with an increased risk of default due to future medical conditions.[49] Furthermore, in countries where private insurance models dominate or where private insurance models are subsidiary to the statutory health insurance models, Big Health and Genetic Data mining could be a common practice for insurers who make risk assessments for each insured person. Inevitably, such practice entails adverse effects for individuals, such as refusals to insure altogether, impositions of higher premiums, refusals to provide coverage for particular treatments or refusals to compensate.[50]

Advertisers and marketers are greatly interested to obtain and analyse health data about their potential customers, so as to better direct their products.[51] As a matter of fact, the location of a customer has been reported to play a decisive role in the determination of a price of a product. For example, individuals suffering from serious illnesses may receive less attractive promotional offers, as advertisers and marketers would be unwilling to invest in consumers with a short life span. And also, educational institutions could equally have a vested interest in examining health data of their applicants.[52] Private universities would be motivated to track applicants' social media behaviour to garner 'big data' and predict not only which students are likely to succeed academically, but also which ones have the best career prospects. Their enrolment can both enhance university prestige and ensure future donations to fund the university.

VI. Qualification of Big Health and Genetic Data

The key issue related to Big Health Data is their legal qualification. As already mentioned above, accumulating Big Health Data does not necessarily involve the collection of personal data and not every stage of Big Health Data analysis may require personal data processing.[53] In light of these considerations the question to be discussed is whether Big

[48] S. Hoffman, Big Data's New Discrimination Threats, Amending the Americans with Disabilities Act to Cover Discrimination Based on Data-Driven Predictions of Future Disease, In: I. G. Cohen/H. F. Lynch/E. Vayena/U. Gasser (ed.), *Big Data, Health Law and Bioethics,* Cambridge University Press, 2018, p 86.

[49] Ibid.

[50] T. Trokanas (2011) The use of genetic data in private insurance. Problems and global perspectives. In: M. Bottis M (ed.) An information law for the 21st century. Third international seminar on Information law, Athens, 2010, p 559.

[51] S. Hoffman, op. cit, p 86.

[52] S. Hoffman, op. cit., p 87.

[53] K. Pormeister, The GDPR and Big Data: Leading the Way for Big Genetic Data?, In: Schweighofer E., Leitold H., Mitrakas A., Rannenberg K. (eds) Privacy Technologies and Policy. APF 2017. Lecture Notes in Computer Science, vol 10518, Springer, 2017, p 12.

Health Data constitute personal data[54] or whether they should be recognised as a new category of data.[55] The answer to the said questions is crucial, insofar as it entails the application or not of GDPR.

An argument against characterizing Big Health Data as personal data might be derived from Recital 9 of the Regulation 2018/1807 on a framework for the free flow of non-personal data in the European Union[56], which cites any aggregate and anonymized datasets used for big data analytics as a typical example of non-personal data. Nonetheless, this argument alone is not convincing for several reasons.

The application of GDPR is dependent on the data type under processing. In principle, Big Health Data may blend different categories of data: *identifiable data*, *de-identified data*, *non-personal data*. Recall that personal data definition in Article 4 (1) GDPR is mainly based on two alternative constructs: identification or identifiability of the data subject.[57] In more detail, GDPR Recital 26 clarifies that the identifiability of a natural person is determined through all the means reasonably likely to be used (such as singling out) by the controller or by another person to identify the natural person, directly or indirectly. Indirectly identifiable data are the result of pseudonymization process. Identifiability is the minimum threshold for the application of data protection rules.[58] On the other hand, de-identified data refers to data that used to be personal data but are now de-identified to an extent that identification of individuals has been rendered unreasonably difficult.[59] Lastly, non-personal data are data which have never been personal; they are equated with anonymous data. How shall each Big Health Data category be legally treated under GDPR?

As a starting point, if Big Health data processing is based on pseudonymized (i.e., indirectly identifiable) data, it is uncontested that GDPR is applicable. According to Article 4 (5) GDPR pseudonymization means processing person data in such a way that they can no longer be attributed to a specific data subject without the use of some additional information. And as it is further stated in GDPR Recital 26, personal data which have undergone pseudonymization *should* be considered to be information on an identifiable natural person.

On the other end of the spectrum, it might be argued that GDPR is inapplicable when Big Health Data processing is based on anonymous, i.e., non-personal data. It is reminded here that according to GDPR Recital 26 data protection principles do not apply to anonymous

[54] European Parliament, Committee on Civil Liberties, Justice and Home Affairs, Report on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI)), 20 February 2017, Ibid, Under J.

[55] Ibid, Under I.

[56] Regulation 2018/1807 of 14 November 2018 on a framework for the free flow of non-personal data in the European Union.

[57] For a full analysis of the definition of personal data see W29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, Adopted on 20th June 2007, Under III.

[58] M. Oostveen, op. cit., p 7.

[59] Ibid.

information, insofar as they do not relate to an identified or identifiable natural person. Apart from the fact that, technically speaking, it is questionable whether anonymization of Big Health Data is irreversible[60], the subcategory of Big Genetic Data adds a new dimension to the anonymity issue. Some scholars subscribe to the view that Big Genetic Data can never truly be rendered anonymous; they advocate that Big Genetic Data are *per se* identifying[61], since they can identify an individual without any other links to the data subject.[62] For these reasons, Big Genetic datasets cannot be exempted from the application of personal data protection legal framework.[63]

Furthermore, Big Health Data analysis seems to blur the line between personal and non-personal data.[64] To rephrase it, the correlation of different types of datasets (e.g. de-identified and non-personal data with identifiable data, de-identified and non-personal data with each other) may create new data sets, ultimately leading individuals to be re-identified.[65] It is the risk of re-identification of individuals which poses a potential threat to the fundamental right of natural persons to the protection of personal data. Theoretically, it would be possible for an insurance company to link non-personal environmental data, such as high air pollution levels in a specific area, to personal data of the residents, in order to charge increased insurance premiums to people who run higher health risks.[66] Another real example is the case of the advertising technology company Amobee, which suggested buying more drinks on a certain day to Mr. D., based on a decision of The Weather Company, a business owned by technology group IBM, that hot weather conditions in his area were likely to cause him an 'overactive bladder'.[67] Consequently, the application of personal data protection rules cannot be excluded *a priori* for Big Health and Genetic Data.[68]

This position may be reinforced with an additional argument that derives from Regulation 2018/1807 on a framework for the free flow of non-personal data in the European Union. As already explained, Big Health and Genetic Data are typical examples of mixed datasets[69], which call for the application of Regulation 2018/1807. According to Article 2 (2) of this Regulation, in cases where personal and non-personal data are inextricably linked, the

[60] M. Oostveen, op. cit., p 8.
[61] K. Pormeister, op.cit., p 9.
[62] Ibid.
[63] Ibid., p 13.
[64] European Parliament, Committee on Civil Liberties, Justice and Home Affairs, Report on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI)), 20 February 2017, Under J.
[65] Ibid, Under 7. See also M. Oostveen, op. cit., p 4 & 9.
[66] Borgesius, F., Gray, J., & Van Eechoud, M. (2015). Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework. Berkeley Technology Law Journal, 30(3), pp 2121-2122.
[67] Financial Times, Aliya Ram/Madhumita Murgia, Data brokers: regulators try to rein in the 'privacy deathstars', 8 January 2019.
[68] Paul Voigt/Axel von dem Bussche, The EU General Data Protection Regulation (GDPR). A Practical Guide, Under 9.1.1.
[69] Communication from the Commission to the European Parliament and the Council, Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union, 29.05.2019, Ibid., p 8.

Regulation shall not hinder the implementation of GDPR. This provision gives rise to two questions: first, how to define the term "inextricably link", and second, if the provision presupposes a certain proportion of personal to non-personal data to the set.

The concept of "inextricably linked" is left undefined in Regulation 2018/1807, but it is interpreted in its Guidance. According to the Guidance, the meaning of "inextricably linked" is any situation whereby separating personal from non-personal data would be impossible, or technically unfeasible, or economically inefficient or likely to significantly decrease the value of the dataset.[70] Moreover, in the Guidance it is stressed that the data protection rights and obligations stemming from GDPR fully apply to the whole mixed dataset, even when personal data represent only a small part of the dataset.[71] Therefore, the assertion that GDPR is inapplicable to Big Health or Genetic datasets, because, for instance, personal data and non-personal data are not inextricably linked, or because their personal data representation is proportionately smaller would be erroneous. In sum, if Big Health or Genetic Data is composed of both non-personal and personal data, a presumption in favour of the application of GDPR rules will apply.

After all, even when Big Health Data processing does not entail the re-identification of individuals, one might question whether identification or identifiability should be the sole criteria under consideration. Nowadays, thanks to algorithmic analytics individuals can be classified according to behaviours, preferences and other characteristics, while still retaining their anonymity.[72] In fact, it is the creation of and classification into groups that challenges the fundamental right of individuals whose Big Health Data are processed.[73] On the contrary it is not necessary for group members to be identified, but it suffices for them to be *classified*. In that sense, a broad interpretation of the legal definition of personal data in Article 4 (1) GDPR could suggest that Big Health Data is deemed personal data not only because they relate to an identified or identifiable natural person, but because they relate to a classified or classifiable natural person.

VII. Principles related to Big Health and Genetic Data Processing

In line with Article 5 of the Convention 108+ of the Council of Europe[74] and Article 5 GDPR, a principled approach to personal data processing is adopted. In other words, the following set of fundamental principles is set forth to govern the processing of personal data: (a) lawfulness, fairness, and transparency, (b) purpose limitation, (c) data minimisation, (e) accuracy, (f) storage limitation, (g) integrity and confidentiality, (h) accountability. It is

[70] Ibid., p 10.

[71] Ibid., p 9.

[72] B. Mittelstadt, From Individual to Group Privacy in Biomedical Big Data, In: I. G. Cohen/H. F. Lynch/E. Vayena/U. Gasser (ed.), *Big Data, Health Law and Bioethics*, p 178.

[73] For instance, advertisers are not so much interested in targeting a specific person as in reaching a target market (See B. Mittelstadt, op. cit., p. 178).

[74] Convention of the Council of Europe for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS No. 108), as amended by its Protocol CETS No. [223].

arguable that the continuous accumulation and analysis of Big Health and Genetic Data undermines at least four of the abovementioned key principles of processing: purpose limitation, data minimisation, storage limitation and transparency.[75]

1. Purpose limitation principle

The purpose limitation principle is the cornerstone of personal data processing.[76] Article 5 (1) lit. b GDPR requires that personal data be collected for specified, explicit and legitimate purposes and not be further processed in a manner that is incompatible with those purposes. In more detail, this means that the purpose of processing shall be determined at the outset of data collection and must be clearly and unambiguously expressed by the controller.[77] Conversely, this also implies that a controller may perform further processing, provided its purpose is considered compatible with the initial purposes. Obviously, processing personal data for undefined or unlimited purposes does not satisfy the purpose limitation requirement.[78]

As highlighted above, Big Data, and specifically Big Health and Genetic Data, can be endlessly used and have unpredictable impact on the data subjects. Their collection and processing go beyond the initial purpose of collection and processing, actually constituting further processing for a different/secondary purpose.[79] Thus, the collection and processing of Big Health and Genetic Data test the limits of purpose limitation principle[80].

A first question that arises is whether Big Health or Genetic Data processing could be considered a compatible further processing within the meaning of Article 5 (1) lit. b GDPR. The compatibility of purpose of any further processing with the original purpose is assessed according to certain criteria, which are indicatively listed in Articles 6 (4) GDPR (e.g., link between purposes, collection context, personal data nature, possible consequences, safeguards). A second question that arises is whether these criteria justify the further processing of Big Health and Genetic Data. In our opinion, the fact that they constitute sensitive data (criterion c) and that their intended further processing can have dire consequences for the data subject (criterion d) are compelling arguments against contemplating Big Health and Genetic Data processing as acceptable further processing.

This view is supported by the fact that in the GDPR Proposal this article foresaw hat even if the secondary purpose were incompatible with the original one a data controller would be entitled to further process data, as long as he had an overriding (legitimate) interest

[75] M. Oostveen, op. cit., p 4.

[76] See Article 8(2) of the EU Charter of Fundamental Rights.

[77] De Terwangne, Article 5, In Christopher Kuner, Lee A. Bygrave, Christopher Docksey, Laura Drechsler (ed.), The EU General Data Protection Regulation (GDPR) A Commentary, Oxford University Press, 2020, p 315.

[78] Ibid.

[79] Hellenic Republic National Bioethics Commission, *Opinion, Big Data in Health*, p 5, available at http://www.bioethics.gr/images/pdf/GNOMES/OPINION_Big_Data_FINAL_EN.pdf.

[80] On the same subject see M. Oostveen, op. cit., p 4.

in doing so.[81] Allegedly, this had been designed to facilitate the use of Big Data applications.[82] Nevertheless, this proposal came under severe criticism from W29, which suggested that it would render one of the fundamental principles of the data protection framework, -the purpose limitation principle- meaningless and void.[83] Ultimately, this version was not adopted, which proves that the drafters of GDPR did not consider Big Health or Genetic Data processing as compatible processing.

On the other hand, it cannot be overlooked that certain "further processing" of personal data has been considered *a priori* compatible with the primary purpose of collection and processing.[84] Article 5 (1) lit. b GDPR does not consider further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes to be incompatible with the initial purpose. This raises the question whether "scientific research purposes" exception might legitimize Big Health or Genetic Data collection and processing. The answer should be replied "no" for four reasons.

First, it should be emphasized that contrary to the repealed Directive 95/46, GDPR covers an even narrower scope as regards scientific purposes. To be more specific, while Article 6 (1) lit. b of Directive 95/46 used to refer more broadly to "scientific purposes"[85], the scope of Article 5 (1) lit. b GDPR has been limited to "scientific research purposes".[86] Second, Big Health and Genetic Data processing cannot be construed as only serving pure scientific research purposes. As it will be explained further below,[87] when the processing of Big Health and Genetic Data is carried out by private entities, it is predominantly driven by multiple underlying economic and commercial motives. Third, even if one claimed that at least Big Health and Genetic Data processing carried out by public entities might serve pure research purposes, both W29 and EDPS share the accurate view that re-using data for scientific research purposes is subject to two separate and cumulative requirements: purpose

[81] See Article 29 Working Party on the Protection of Individuals, Press release on Chapter II of the draft regulation for the March JHA Council, 15.03.2015.

[82] The Final European Union General Data Protection Regulation, Bloomberg BNA, Privacy and Security Law Report, 25.01.2016, p 6.

[83] Article 29 Working Party on the Protection of Individuals, Press release on Chapter II of the draft regulation for the March JHA Council, 15.03.2015.

[84] De Terwangne, Article 5, In Christopher Kuner, Lee A. Bygrave, Christopher Docksey, Laura Drechsler (ed.), The EU General Data Protection Regulation (GDPR) A Commentary, Oxford University Press, 2020, p 316.

[85] Cf. Article 6 §1 (b) Directive 95/46: "Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;". According to the Explanatory Memorandum to Recommendation No.R (97) 18 of the Committee of Ministers to Member States concerning the protection of personal data collected and processed for statistical purposes, the aim of "scientific purposes" is to provide researchers with information contributing to an understanding of phenomena in fields as varied as epidemiology, psychology, economics, sociology, linguistics, political science, ecology, and so on.

[86] GDPR Recital 162 explains that the processing of personal data for scientific research purposes should be interpreted in a broad manner, including for example technological development and demonstration, fundamental research, applied research and privately funded research, as well as studies conducted in the public interest in the area of public health.

[87] See Section VIII.

specification and lawfulness.[88] What this means is that the presumption of compatibility for scientific research purposes in Article 5 (1) lit. b GDPR does not suffice for Big Health or Genetic Data processing and, as it will be explained below[89], a new legal basis has to be established. Fourth, according to GDPR Recital 162 the application of the statistical purpose exception presupposes that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person. If we drew an analogy between statistical and scientific research purposes and extended the application of the criterion of the former to the latter, we would conclude that Big Health or Genetic Data processing cannot be misperceived to serve scientific research purposes, since it is likely to result in decisions affecting natural persons.

2. Data minimisation and storage limitation principles

Article 5 (1) lit. c GDPR establishes the data minimisation principle, according to which data should be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed. Recital 39 GDPR specifies that personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means, whereas Article 4 (1) of Directive 2016/680 adds the element that data shall "not be excessive", a wording which does not affect the substantial content of data minimisation principle.[90] The necessity requirement relies on quantity, that is, if processing of excessively large amounts of data takes place, and quality, i.e. if the processing causes disproportionate interference in the data subjects' rights and interests. [91]

Article 5 (1) lit. e GDPR establishes the storage limitation principle, according to which data shall be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed. Recital 39 GDPR adds that in order to comply with this principle the controller must set limits for erasure or for a periodic review. Following the same logic, Article 5 of Directive 2016/680 mandates the establishment of time-limits for storage and review and the implementation of corresponding procedural measures.

Big Health and Genetic Data processing can be incompatible with both data minimisation[92] and storage limitation principles, as it not only involves the processing of large

[88] Article 29 Data Protection Working Group, Opinion 03/2013 on purpose limitation, adopted on 2 April 2013, p 12. European Data Protection Supervisor, A Preliminary Opinion on data protection and scientific research, 6 January 2020, p 22. Both bodies attempted to attach the true meaning to GDPR Recital 50, which appears to assimilate purpose specification and lawfulness: "In such a case, no legal basis separate from that which allowed the collection of the personal data is required […] Further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be considered to be compatible lawful processing operations".

[89] See Section VIII.

[90] De Terwangne, op. cit., p 317.

[91] Ibid.

[92] On the same subject see M. Oostveen, op. cit., p 4.

datasets which often exceeds the principal proportionality, but it is also structured in a way that involves their continuous use for an indeterminate period.

3. Transparency principle

According to Recital 39 GDPR it should be transparent to natural persons that personal data concerning them are collected, used, consulted, or otherwise processed and to what extent the personal data are or will be processed. Also, the principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used.

The transparency principle is seriously undermined by Big Health and Genetic Data analytics. Artificial Intelligence, Internet of Things and machine learning algorithms are deeply embedded into Big Health and Genetic Data analysis[93], rendering their collection invisible and their processing tools and techniques obscure.[94] To address this issue, the notions of transparency and by extension of accountability (Article 5 (2) GDPR)[95] need to be redefined and to be upgraded to algorithmic transparency and algorithmic accountability accordingly.[96]

This means that automated systems must provide data subjects a description of the automated decision system, the kind of data that are processed, its purpose, the risks to the privacy and security and the measures employed to mitigate it.

VIII. Lawful processing of Big Health and Genetic Data

Big Health datasets are created through correlations of primary health data collections from both biomedical and non-biomedical sources. They can be further processed for a secondary purpose that constitutes a further processing. This means that irrespective of the

[93] Ioannis Iglezakis/Theodoros Trokanas/Panagiota Kiortsi, The right not to be subject to automated individual decision-making/profiling concerning Big Health Data. Developing an algorithmic culture, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3802771.

[94] Neil Richards/Jonathan King, Three Paradoxes of Big Data, Stanford Law Review Online, Vol. 66, 3 September 2013, p 42.

[95] It is reminded that accountability is not a standalone principle, but it is substantiated by the other obligations of the data controller, as it is suggested by the direct references of Article 5 (2) to Article 5 (1) GDPR (P. Voigt/A. von dem Bussche, The EU General Data Protection Regulation (GDPR). A Practical Guide, Springer, 2017, Under 4.1, p 87).

[96] The terms "algorithmic accountability" and "algorithmic transparency" were first coined by the European Parliament (European Parliament, Committee on Civil Liberties, Justice and Home Affairs, Report on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI)), 20 February 2017, Under N) to represent a pair of basic principles of personal data processing customised to a Big Data framework. Likewise, the term "algorithmic accountability" is encountered in the U.S. Algorithmic Accountability Act of 201954, which was introduced in House on 4 November 2019, but has yet to be enacted (for a detailed analysis see Ioannis Iglezakis/Theodoros Trokanas/Panagiota Kiortsi, op. cit.).

legal basis chosen for the initial purpose of processing, in order for the secondary processing of Big Health Data to be lawful, another legal basis must apply. Furthermore, Big Health and Genetic Data as a special category of personal data, can be processed under Article 6 GDPR, but also a derogation for processing under Article 9 GDPR must be found.[97] In what follows, a clear distinction will be made between Big Health and Genetic Data processing carried out by public and private entities.

The primary purpose of public entities' processing of health data is the provision of health and social care. In this case, data are collected directly from patients or their legal representatives (e.g., if minors or legally incompetent persons are concerned), either in an in-person healthcare setting (e.g., in a physician's office or care facility) or in a tele-healthcare setting (e.g., during a remote consultation using eHealth or mHealth tools).[98] It should be stressed out that sharing data for healthcare, whether across borders or not, is still considered an initial/original purpose.[99] Processing health data for the purpose of patient care takes place[100] under the legal basis of a *legal obligation* (Article 6 (1) lit. c GDPR) or for the performance of a task carried out in the *public interest* or in the exercise of official authority vested in the controller (Article 6 (1) lit. e GDPR), i.e. health purposes such as public health and social protection and the management of health care services.[101] Apart from the aforementioned legal bases, health data processing for the purpose of patient care is lawful when Article 9 (2) GDPR applies: *Medical diagnosis, provision of health or social care or treatment* (Article 9 (2) lit. h GDPR) or *public interest in the area of public health* (Article 9 (2) lit. i GDPR) will be the most appropriate derogations to process Big Health data.

At the same time, public entities (e.g., social security institutions, public universities or national centres for scientific research, ministries of health, social services) process health data for wider public health purposes, according to the authority that was vested to them by the Law. Indicatively, management of health or social security systems; ensuring high standards of quality and safety in conducting scientific research that involves health data, e.g., market approval of medical products and medical devices, pharmacovigilance, and medical device monitoring; protection against serious cross-border threats to health, e.g., prevention and control of contagious diseases; operation of National disease registries.[102] Such processing may include health data collected ab initio for the aforementioned primary purposes or constitute a different use of health data that were collected from data subjects.

[97] GDPR Recital 51: "In addition to the specific requirements for such processing, the general principles and other rules of this Regulation should apply, in particular as regards the conditions for lawful processing". See also European Parliament Resolution of 25 March 2021 on the Commission evaluation report on the implementation of the General Data Protection Regulation two years after its application, Under 5.

[98] European Commission, Assessment of the EU Member States' rules on health data in the light of GDPR, 2021, p 23.

[99] Ibid, p 24.

[100] Ibid, p 28-30.

[101] GDPR Recital 45.

[102] European Commission, Assessment of the EU Member States' rules on health data in the light of GDPR, 2021, p 42.

When data are being processed for the initial purpose the processing could be deemed to be in compliance with a legal obligation (Article 6 (1) lit. c GDPR, or be necessary for the performance of a task carried out in the public interest (Article 6 (1) lit. e GDPR) or even be necessary in order to protect the vital interests of the data subject or of another natural person where the data subject is physically or legally incapable of giving consent (Article 6 (1) lit. d GDPR), e.g., for monitoring epidemics and their spread or in situations of humanitarian emergencies, in particular in situations of natural and man-made disasters.

The application of the Article 6 (1) and Article 9 (2) GDPR regarding the initial purpose must also apply when health data are being processed for different purposes.[103] Wider public purposes are highly likely to entail Big Health Data processing. Such processing is in compliance with Article 6 (1) lit. c GDPR or Article 6 (1) lit. e GDPR or even Article 6 (1) lit. d GDPR[104], e.g., for monitoring the spread of epidemics or in situations of humanitarian crisis, in particular in situations of natural and man-made disasters[105]. Furthermore, the controller must also choose a suitable derogation, i.e., either Article 9 (2) lit. h GDPR (e.g., management of health or social care systems and services) or Article 9 (2) lit. i GDPR (public interest concerning public health issues, e.g., protecting against serious cross-border threats to health or ensuring high standards of quality and safety of health care and of medicinal products or medical devices) or even Article 9 (2) lit. j GDPR (vital interests of the data subject or of another natural person where the data subject is physically or legally incapable of giving consent).

In certain cases, public entities, such as universities, hospitals, and research institutions may process health data for scientific research purposes. Such processing may involve Big Health Data collected initially for the aforementioned purposes or a re-use of health data collected for different purposes.[106] In the first case, lawful processing of Big Health Data needs to be based on the informed and unambiguous *consent* of the data subject (Articles 6 (1) lit. a and 4 (11) GDPR) coupled with a derogation of Article 9 (2) lit. a GDPR (*explicit consent*)[107] or Article 9 (2) lit. i GDPR *(research purposes* in accordance with Article 89 (1) GDPR, on condition that they are proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject).[108] In the second case (different

[103] Ibid.

[104] Nonetheless, processing based on this legal basis should in principle take place only where the processing cannot be manifestly based on another legal basis (GDPR Recital 46).

[105] GDPR Recital 46.

[106] European Commission, Assessment of the EU Member States' rules on health data in the light of GDPR, 2021, p 57.

[107] According to GDPR Recital 161 as far as consent to participation in scientific research activities in clinical trials is concerned health data processing shall also comply with the provisions of Articles 28 et seq. of EU Regulation 536/2014 on clinical trials on medicinal products for human use.

[108] It should be stressed that the adoption of EU or Member State law is a prerequisite for the derogation of Article 9 (2) lit. i GDPR to become fully operational for Big Health Data (European Data Protection Supervisor, A Preliminary Opinion on data protection and scientific research, 6 January 2020, p 17).

purpose), it has to be examined whether the (explicit) consent given for initial purposes can also cover Big Health Data research activities.

In Recital 33 GDPR, it is evident that the full identification/understanding of the scientific research purposes of the processing is unattainable. For this reason the GDPR provides that data subjects should be allowed to give their consent to certain areas of scientific research in accordance with recognised ethical standards for scientific research.[109] As EDPB remarks, the inability to specify the purpose at the outset of a research project should not be interpreted as an attempt to reduce the importance of consent as a legal basis,[110] but rather as a recognition of the need to be flexible regarding the degree of purpose specification in the context of scientific research.[111] Evidently, Recital 33 GDPR reflects the so-called "broad consent" theory, a term which is encountered in U.S. legislation.[112] Nonetheless, when special categories of data are processed on the basis of explicit consent, as in the case of Big Health Data, in order to implement the flexible approach of Recital 33 GDPR, a stricter interpretation and high degree of scrutiny would be required.[113] In such cases, consent requirements would be better met if a step-by-step approach were adopted: data controllers could allow data subjects to consent more generally at the outset of a research project and, as research advances, they could obtain a separate consent before every next step.[114] Needless to say, the initial consent, even in its broad sense, cannot legitimise processing in cases where traditional research institutes and public bodies cooperate or enter into partnerships with private technology companies, e.g. UK National Health Service granting access of health data of 1.6 million patients to Google and its AI Company Deep Mind.[115]

On the other hand, there are private entities such as internet service providers, bankers, insurers, employers, private universities, marketers who process non-biomedical data that were collected for different initial purposes and when combined with non-personal data, they may generate Big Health Data. The lawfulness of the initial processing of non-biomedical data: (i) can be based on the *consent* of data subject under Article 6 (1) lit. a GDPR, e.g., consent to have access to information society services, like social media, broadband-

[109] On the same issue see also European Data Protection Board, Guidelines 05/2020 on consent under Regulation 2016/679, Adopted on 4 May 2020, Under 154.
[110] European Data Protection Board, Guidelines 05/2020 on consent under Regulation 2016/679, Adopted on 4 May 2020, Under 154.
[111] Ibid, Under 153.
[112] The US Code of Federal Regulations (CFR) in the Section entitled "Basic Health and Human Services Policy for Protection of Human Subjects" stipulates under §46.116: "General requirements for informed consent. Broad consent may be obtained in lieu of informed consent obtained in accordance with paragraphs (b) and (c) of this section only with respect to the storage, maintenance, and secondary research uses of identifiable private information and identifiable biospecimens".
[113] European Data Protection Board, Guidelines 05/2020 on consent under Regulation 2016/679, Adopted on 4 May 2020, Under 155.
[114] Ibid, Under 157.
[115] European Data Protection Supervisor, A Preliminary Opinion on data protection and scientific research, 6 January 2020, p 7.

enabled digital tools, or (ii) may be necessary for the performance of a *contract* or in the context of an intention to enter into a contract under Article 6 (1) lit. b GDPR, e.g., online sales contracts, banking or insurance contracts, pre-contractual relationships between employer/employee or student/private university) or (iii) for the *legitimate interests* pursued by the controller or by a third party under Article 6 (1) lit. f GDPR (e.g., direct marketing purposes). Apparently, the legal bases of the performance of a contract or legitimate interests, which applied for the initial processing of non-biomedical data, cannot justify Big Health Data processing with the use of Big Data analytics by private entities. It is extremely unlikely for a data subject to enter into a contract with the data controller for such processing. It is also unlikely that the legitimate interests of a Big Health Data controller will prevail over the interests or fundamental rights and freedoms of a data subject (Article 6 (1) lit. f GDPR), considering the latter's reasonable expectations.[116]

Similarly, the initial consent of the data subject cannot apply as a legal basis for Big Health Data processing, since data controllers tend to misuse it. Based on the consent that were given for different purposes, data controllers claim that they conduct scientific research, when, in reality, they collect all kind of data, which can subsequently be further processed for other purposes[117], primarily commercial ones. An example of this customary practice is when companies offer direct-to-consumer genetic tests, both health and non-health related. Admittedly, in practical terms it is often difficult for a data controller of Big Health Data to obtain a data subject's consent due to either the huge volume of datasets or the inability to identify the data subjects concerned. However, only by having data subjects provide a new, *explicit consent* (Article 6 (1) lit. a, Article 9 (2) lit. a GDPR) for the purpose of Big Health Data analysis, can private entities lawfully process them.

IX. Data Protection Impact Assessment

The processing of Big Health and Genetic Data involves large scale use of special categories of data and, generally, the use of Artificial Intelligence or other technologies, including profiling, in order to draw conclusions and generate new data. In addition, this processing is likely to pose a high risk to the rights and freedoms of natural persons, and in particular the health of the population. For these two reasons a Data Protection Impact Assessment (DPIA) for Big Health and Genetic Data processing is required, in accordance with Article 35 (3) lit. a and b GDPR[118].

[116] GDPR Recital 47.

[117] European Data Protection Supervisor, A Preliminary Opinion on data protection and scientific research, 6 January 2020, p 5.

[118] ICO, Data Protection Act and General Data Protection Regulation, Big data, artificial intelligence, machine learning and data protection pp 70 and 99 et seq. See also Georgiou D, Lambrinoudakis C*,* Data Protection Impact Assessment (DPIA) for Cloud-Based Health Organizations. Future Internet. 2021, Volume 13 issue 3, p 66, available at: https://doi.org/10.3390/fi13030066.

Carrying out a DPIA is challenging for both private and public organisations.[119] Article 35 GDRP does not provide a concrete methodology on how to carry out a DPIA[120], but it sets out[121] certain basic criteria, which must be included at a minimum in every DPIA, i.e., the description of the processing operations envisaged and the purposes of the processing, identification of the necessity and proportionality of the processing, tracking of the potential risks to the rights and freedoms of data subjects, assessment of risk management, evaluation of the compliance level of the Organisation with GDPR.

Although the issue of conducting a DPIA has not been yet addressed as far as Big Health and Genetic Data are concerned, the procedure is not different as regards the above mentioned criteria.[122] Specifically, a DPIA requires a high compliance level with GDPR, including the safeguarding of data subject rights, consulting a Data Protection Officer, adopting organisational and technical measures that enhance privacy and maintaining records of processing activities and data flows[123], as well as a clear understanding of the roles of data controller and data processor or joint controller and their interaction. This is important as according to GDPR data controllers have the obligation to conduct a DPIA.

After identifying the need to carry out a DPIA, a controller must determine the legal bases of processing, as analysed in Section VIII. In any case, the controller should be able to justify the purpose of processing and the type of data to be processed. It should be guaranteed that Big Health and Genetic Data are accurate and appropriate for the purposes for which they are processed. The storage period should be specific. Data controllers should also identify and evaluate if data processing may result in negative outcomes on the rights and fundamental freedoms of data subjects.[124] To mitigate such negative outcomes, data controllers must provide appropriate measures and monitor their effectiveness.

---

[119] Georgiou D/Lambrinoudakis C, op. cit..

[120] Since 2009 E.U. had called for a multidisciplinary co-operation between Member States and stakeholders, such as service providers and civil society associations, in order to develop risk assessment methodologies and evaluate the impact of certain technologies in privacy and personal data. Despite the fact that W29 Working Party was delegated to ratify an official impact assessment methodology, no such document was approved. For a widely accepted methodology to conduct a DPIA see CNIL, Privacy Impact Assessment (PIA), Knowledge bases, February 2018 edition, available at https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-3-en-knowledgebases.pdf.

[121] Article 29 Data Protection Working Party, WP 248 rev.01, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, As last Revised and Adopted on 4 October 2017, available online, https://www.cnil.fr/sites/default/files/atoms/files/20171013_wp248_rev01_enpdf_4.pdf

[122] See indicatively the steps of conducting a DPIA for Big Data proposed in ICO, Data Protection Act and General Data Protection Regulation, Big data, artificial intelligence, machine learning and data protection p 100 et seq., https://ico.org.uk/media/fororganisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

[123] See Articles 24, 25, 30, 35 GDPR.

[124] Consultative Committee of the Convention for the protection of individuals with regard to automatic processing of personal data, (T-PD) Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data p.3 https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a

X. Data Protection by Design in Big Health and Genetic Data Processing

Apart from securing the legality of processing of Big Health and Genetic Data, more sophisticated methods are needed to ensure that the drawbacks of notice and consent mechanisms, which do not provide adequate transparency and control, are addressed. Apparently, the principle of data protection by design, which is enshrined in Article 25 GDPR, is a means to address the privacy risks inherent in Big Data processing. Since it entails the implementation of appropriate technical and organizational measures from the initial phase, but also in the course of processing, it can be a useful tool for empowering the individual, protecting its personal data by default, and also, enhancing the responsibility of the data controller.[125]

The provision of Article 25 GDPR indicates that the controller may implement appropriate technical and organisational measures, such as pseudonymisation, and implementation of data-protection principles, e.g., data minimisation, effectively and, furthermore, integrate the necessary safeguards into the processing in order to meet the GDPR requirements and protect the rights of data subjects.

ENISA developed a specific strategy to implement the privacy by design principle in Big Data processing, which may be taken into account in the development of Big Health Data analytics.[126] This can be summarized as follows: a) in the data acquisition/collection phase data minimization should be introduced to define what personal data are necessary, aggregated information should be instead of personal data, privacy enhancing technologies used, the individuals should be adequately informed about the collection of their data, consent should be obtained, while opt-out tools should be available; b) in the data analysis and curation phase anonymization should be implement and also, encryption, especially in the context of performing searches and other computations over encrypted data; c) in the data storage phase security measures such as granular access control and authentication should be used and privacy preserving analytics in distributed systems; d) finally, in the data usage phase anonymization should be used to privacy preserving data publishing and retrieval that could prevent inference of personal data.

XI. Conclusions

The revolution of Big Health and Genetic Data is seen as a double-edged sword. Heavily impacted by the use of AI, algorithms and technologies that reclaim health data for further use, Big Health and Genetic Data analysis yields ambiguous results that substantial impact on individuals. As there is no jurisprudential consensus on the definition of these categories of data, it is more expedient to describe them by means of their defining characteristics. Their potential sources are diverse, and the rapid expansion of informatics has

[125] ENISA, Privacy by design in big data. An overview of privacy enhancing technologies in the era of big data analytics, 2015, p. 21.
[126] ENISA, ibid, p. 23 et seq.

been exciting the interest of various stakeholders, with data brokers occupying a prominent position.

Big Health and Genetic Data are typical examples of mixed datasets, insofar as they may blend identifiable data (or pseudonymized data), de-identified data, non-personal or anonymous data. Apart from pseudonymized Big Health and Genetic Data, which understandably fall under the scope of GDPR, even anonymous Big Genetic Data do so, as they are regarded per se identifying. Ultimately, in this paper it has been questioned whether identification or identifiability could be the sole criteria for the application of GDPR, suggesting classification or classifiability, instead.

It has been shown that the incessant accumulation and analysis of Big Health and Genetic Data compromise some of the core GDPR principles. As far as the purpose limitation principle is concerned, it has been mentioned that the presumption of compatibility for scientific research purposes does not suffice to justify Big Health or Genetic Data processing. Moreover, in order to prevent the data controller from overriding his legal obligations, a redefinition of the principles of transparency and accountability and their upgrade to algorithmic transparency and algorithmic accountability accordingly is proposed.

Big Health and Genetic Data processing may involve data collected initially for such purposes or constitute a re-use of data collected for other, initial purposes. In either case, lawful processing of Big Health and Genetic Data necessitates both a legal basis under Article 6 GDPR and a separate derogation for processing under Article 9 GDPR. Explicit consent (both as a legal basis and a derogation) is preferable where Big Health and Genetic Data processing is carried out by private entities, but it will remain the exception where their processing is performed by public entities (e.g., scientific research purposes). In addition, DPIAs should be carried out in Big Health and Genetic Data projects and their implementation is no different from any other case of personal data processing.