Synthetic data: a holy grail to healthcare research?

Marta Beleza Costa[1]
Miguel Goulão[2]

Abstract:
Healthcare research heavily relies on patient data, raising challenges due to stringent data protection legislation. Synthetic data is presented as a solution, offering privacy while facilitating research. This essay delves into the intersection of synthetic data and data protection law, analysing its regulatory implications, benefits, and disadvantages. Synthetic data shows value in various healthcare applications, from medical imaging to epidemiological studies. However, drawbacks like bias amplification and data quality assessment are still a concern. While synthetic data holds potential, regulatory frameworks need refinement to fully leverage its capability to be considered a "Holy Grail".

Keywords: Synthetic data; scientific research; healthcare; data protection law

Resumo:
A investigação cientifica no sector da saúde encontra-se fortemente dependente do tratamento dos dados de doentes. As exigências

[1] Graduated in Law from the Faculty of Law of the University of Lisbon and Master's student of Law and Tech at NOVA School of Law. Postgraduate in Digital Services Law from the Private Law Research Center of the Faculty of Law of the University of Lisbon. martabpcosta@gmail.com

[2] Data Privacy Consultant. Graduated in Law from the Faculty of Law of the University of Lisbon and Master's student of Law and Tech at NOVA School of Law. Postgraduate in Digital Services Law and in Data Protection Law, both from the Private Law Research Center of the Faculty of Law of the University of Lisbon. goulao.miguel@gmail.com

rigorosas em matéria de legislação de proteção de dados a que este tratamento está sujeito colocam vários desafios ao desenvolvimento das atividades de investigação. Assim, os dados sintéticos são apresentados como uma solução, assegurando a privacidade e, consequentemente, facilitando a investigação. Este artigo explora a relação entre os dados sintéticos e o direito da proteção de dados, analisando as implicações regulatórias, os seus benefícios e as suas desvantagens. Os dados sintéticos têm demonstrado utilidade no domínio dos cuidados de saúde, desde a imagiologia médica a estudos epidemiológicos, embora apresentando, igualmente, algumas desvantagens, como a amplificação de *bias* ou a fraca qualidade dos dados. Embora a utilização de dados sintéticos apresente potencial, é necessário introduzir reformas legislativas que possibilitem que as suas capacidades sejam plenamente aproveitadas e, consequentemente, que estes dados se tornem um verdadeiro *Holy Grail*.

Palavras-chave: Dados sintéticos; investigação científica; cuidados de saúde; direito da proteção de dados

I. Introduction

The processing of personal data for the purposes of healthcare scientific research has faced significant obstacles. Insofar as sharing knowledge and other information about health data (such as patient health records or diagnostics) is crucial for healthcare research, there is the necessity to overcome the limitations established in data protection legislation[3-4], in order to foster innovative and proper deployment of research[5]. This is especially relevant, since health data is regulated in a more stringent way by data protection legislation, having the need for assuring informed consent by the data subjects[6]. Also, given the fact that the processing of high quantities of special categories of data by new technologies is at stake, in a situation where it is possible to have indirect data collection of vulnerable data subjects, there might be the necessity to perform a Data Privacy Impact Assessment (DPIA)[7].

Nevertheless, even when complying with data protection legislation[8], researchers might face other regulatory restrictions, mostly related to ethical aspects for sharing and securing data (especially with regard to scientific research on humans). Other limitations include costly access

[3] It will be considered the General Data Protection Regulation par excellence: REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR).

[4] KUO, Nicholas; PEREZ-CONCHA, Oscar; HANLY, Mark; MNATZAGANIAN, Emmanuel; HAO, Brandon; DI SIPIO, Marcus; YU, Guolin; VANJARA, Jash; VALERIE, Ivy; OLIVEIRA COSTA, Juliana; CHURCHES, Timothy; LUJIC, Sanja Lujic; HEGARTY, Jo; JORM, Louisa; BARBIERI, Sebastiano, "Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project", *JMIR Medical Education,* 10, 2024; available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10828942/

[5] GONZALES, Aldren; GURUSWAMY, Guruprabha; SMITH, Scott, "Synthetic data in health care: A narrative review", *PLOS Digit Health,* II, 1, 2023, available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082

[6] Note, for example, art. 13 and 14, GDPR.

[7] As set out in art. 35, GDPR.

[8] Note, for example art. 5(2), GDPR.

to datasets that are outside of the public domain[9] and data fragmentation that can result in gathering an incomprehensive dataset for exhaustive analysis, consequently, hindering researcher's capability to draw significant conclusions and data imbalance[10], and leading to misleading conclusions[11]. In addition, the lack of high-quality datasets has also been a significant barrier, especially to comprehend disease patterns, the identification of effective treatments and patient outcomes improvement[12].

The emergence of Artificial Intelligence (AI) – namely, Machine Learning (ML) models, such as Generative Adversarial Networks (GAN)[13] – has been proven to allow progress in the various medical fields, while maintaining efficient and dependable procedures when it comes to data access[14]. However, the biggest contribution that AI has brought for this purpose lies on the possibility of generating synthetic data – which has been presented as a game changer for solving the aforementioned limitations. In fact, data protection authorities, such as the Information Commissioner Office (ICO)[15] and the *Commission Nationale de l'Informatique et des Libertés* (CNIL)[16], have been

[9] GONZALES, Aldren; GURUSWAMY, Guruprabha; SMITH, Scott, "Synthetic data in health care: A narrative review", *PLOS Digit Health,* II, 1, 2023, available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082

[10] This occurs when there is misrepresentation of determined samples.

[11] ALI, Zahid, "The Dawn of a New Era: How Synthetic Data is Transforming Medical Research", 2023, available at: https://www.linkedin.com/pulse/dawn-new-era-how-synthetic-data-transforming-medical-research-ali-pmnjc/

[12] LAMBERTI, Aldo, "Lifting Data Barriers: Exploring Synthetic Data in Healthcare Research", 2023, available at: https://syntheticus.ai/blog/lifting-data-barriers-exploring-synthetic-data-in-healthcare-research

[13] "GANs work by employing two neural networks: one creates fake samples, and the other assesses how close they are to real data. These networks collaborate to refine the generated samples until they closely resemble real data", KAABACHI, Bayrem; DESPRAZ Jérémie, MEURERS Thierry; OTTE, Karen; HALILOVIC, Mehmed; PRASSER, Fabian; RAISARO, Jean Louis, "Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics", *MedRxiv,* 2023, available at: https://www.medrxiv.org/content/10.1101/2023.11.28.23299124v1.full

[14] *Ibidem*.

[15] MARSHALL, Valerie; MARKHAM, Charlie; AVRAMOVIC, Pavle; COMERFORD, Paul; MAPLE, Carsten; SZPRUCH, Lukasz, "Exploring Synthetic Data Validation – Privacy, Utility and Fidelity", *FCA Report*/*ICO Research Paper*, 2023, available at: https://www.fca.org.uk/publications/research-articles/exploring-synthetic-data-validation-privacy-utility-fidelity

[16] CNIL, "Artificial Intelligence: The CNIL Publishes a Set of Resources for Professionals", 2022, available at: https://www.cnil.fr/en/artificial-intelligence-cnil-publishes-set-resources-professionals

exploring their benefits as a privacy enhancing technology. In addition, the European Data Protection Supervisor (EDPS) led a Webinar focused on the use of synthetic data as a possible technology to mitigate data protection risks[17], where even synthetic data was presented as a "panacea" for the purposes of research[18].

Having said that, some underlying concepts will be provided, focusing on the notions of health data and scientific research. Then, the synthetic data conundrum in the light of the General Data Protection Regulation (GDPR) will be discussed, by exploring its definition and dogmatics. After discussing the regulatory state of the art in the EU, the benefits and disadvantages of the usage of synthetic data in healthcare research will be presented, while providing some potential solutions to the aforementioned challenges. Finally, one must be able to conclude if synthetic data should be considered a "Holy Grail" to healthcare research.

II. Data in the Healthcare Domain

i. Data Concerning Health

Data concerning health, notwithstanding its definition in art. 4(15), GDPR[19], must be interpreted in a broad sense, accordingly to the CJEU's understanding[20]: "the expression data concerning health

[17] IPEN; Webinar on the theme: "Synthetic data: what use cases as a privacy enhancing technology?", 2021; available at: https://www.edps.europa.eu/data-protection/our-work/ipen/ipen-webinar-2021-synthetic-data-what-use-cases-privacy-enhancing_en

[18] IPEN; "Synthetic data: what use cases as a privacy enhancing technology?" – Webinar; 2021; available at: https://www.edps.europa.eu/data-protection/our-work/ipen/ipen--webinar-2021-synthetic-data-what-use-cases-privacy-enhancing_en (remarks made by Wojciech Wiewiórowski).

[19] Art. 4(15) states: "Personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status".

[20] CJEU, C-101/01 Lindqvist, Bodil Lindqvist v Åklagarkammaren i Jönköping, 6/11/2003 para 50, 51

(…) must be given a wide interpretation so as to include information concerning all aspects, both physical and mental, of the health of an individual". Recital 35 provides some examples, such as information derived from the testing or examination of a body part or bodily substance, and information, amongst others, on a disease or medical history.

This type of data is considered sensitive data under art. 9(1), which leads to the prohibition of its processing, unless an exception provided for in paragraph 2 is applicable. Art. 9(2)(j) makes it possible to process sensitive data for scientific research purposes, in compliance with the provisions of art. 89(1) and provided that the following requirements, regarding the processing, are met: i) based on Union or Member State law; ii) proportionate to the aim pursued (the purpose of the data processing); iii) respect the essence of the right to data protection (the right to informational self-determination)[21]; iv) provides for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject[22]; v) that technical and organizational measures are ensured, mostly in order to comply with the data minimization principle.

ii. The Scientific Research Conundrum

a. Scientific Research in the GDPR

The concept of "scientific research" is not explicitly defined in the GDPR, although it is mentioned in several provisions[23]. However,

[21] BARRETO MENEZES CORDEIRO, António, *Direito da Proteção de Dados – À luz do RGPD e da Lei n.º 58/2019*, Almedina, 2020.

[22] Criticisms to art. 89(1) have been pointed out due to the generality and restricted scope of the safeguards stated. Regarding these considerations, PORMEISTER, Kärt, "Genetic Data and the Research Exemption. Is the GDPR Going too Far?", International Data Privacy Law, VII, 2, 2017, available at https://academic.oup.com/idpl/article-abstract/7/2/137/3798545?redirectedFrom=fulltext

[23] This is the case of art. 5(1)(b) and (e), art. 9(2)(j), art. 17(3)(d), art. 21(6) and art. 89.

recital 159[24] determines its concept in a broad sense, stating 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". Recital 157 offers numerous instances of the kinds of scientific research that can be conducted. These include registry-based research, medical research into cardiovascular disease, cancer, and depression[25].

The UK data protection authority has established certain criteria in order to better understand this concept, making the following classification[26-27]: i) activities: scientific research incorporates, among others, the formulation of hypotheses, the isolation of variables and the design of experiments, the observation and measurement of data, the publication of findings, data integration and analysis, and inferential statistics; ii) standards: these include technical guidelines and approval by a committee and specific rules (e.g. animal or human research, clinical trials); iii) access: it includes the publication of results or commitment to share research findings, access to which does not have to be open (it can be in a scientific journal or other type of paid publication).

[24] Recital 159 states: "Where personal data are processed for scientific research purposes, this Regulation should also apply to that processing. For the purposes of this Regulation, 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. In addition, it should take into account the Union's objective under Article 179(1) TFEU of achieving a European Research Area. Scientific research purposes should also include studies conducted in the public interest in the area of public health. To meet the specificities of processing personal data for scientific research purposes, specific conditions should apply in particular as regards the publication or otherwise disclosure of personal data in the context of scientific research purposes. If the result of scientific research in particular in the health context gives reason for further measures in the interest of the data subject, the general rules of this Regulation should apply in view of those measures."

[25] WIESE SVANBERG, Christian, *The EU General Data Protection Regulation (GDPR): A Commentary*, Oxford Academic, 2020

[26] ICO, "GDPR Guidance and Resources, The Research Provisions", 2023, available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/the-research-provisions/

[27] CANTO MONIZ Graça, Lesson taught as part of "IV Pós Graduação Avançada em Direito da Proteção de Dados Pessoais", subordinated to the theme "O Tratamento de Dados Pessoais para Fins de Investigação Científica", 2022

The most common legal basis used for scientific research purposes are the consent of the data subject[28], as well as the legitimate and public interest. Regarding consent, recital 33 states that the requirement for specificity is somewhat mitigated. This is because in certain scientific research projects it is not always possible, at the time consent is collected, to identify the full purpose of the processing (since research projects change from time to time)[29-30]. However, the EDPB has already indicated that, when processing special categories of data, the approach provided for in recital 33 needs to "be subject to a stricter interpretation and requires a high degree of scrutiny", especially because of the limitations provided for in art. 9[31].

Processing of data for scientific research purposes needs to fulfil the requirements set out in art. 89(1). In this sense, the data processing needs to comply with appropriate safeguards for the rights and freedoms of the data subject, which need to ensure the implementation of technical and organisational measures to respect the principle of data minimisation[32]. Note that art. 89 must also be read alongside art. 9(4): insofar as art. 9(4) allows Member States do introduce national

[28] Consent is the most commonly used legal basis, since there is a set of ethical rules that are embodied in the Declaration of Helsinki, which state that consent to participate in scientific studies is an ethical requirement. With regard to these considerations, CANTO MONIZ Graça, Lesson taught as part of "IV Pós Graduação Avançada em Direito da Proteção de Dados Pessoais", subordinated to the theme "O Tratamento de Dados Pessoais para Fins de Investigação Científica", 2022

[29] CANTO MONIZ Graça, Lesson taught as part of "IV Pós Graduação Avançada em Direito da Proteção de Dados Pessoais", subordinated to the theme "O Tratamento de Dados Pessoais para Fins de Investigação Científica", 2022

[30] This would be the case, for example, for some research projects that involve "data-intensive longitudinal population-based research": since there is a high volume and variety of data being processed, starting of with a very detailed consent form might not be the most appropriate approach. Regarding these considerations, HO, Chih-hsing, "Challenges of the EU General Data Protection Regulation for Biobanking and Scientific Research", *Journal of Law, Information and Science*, XXV, 1, 2018, available at: https://www6.austlii.edu.au/cgi-bin/viewdoc/au/journals/JlLawInfoSci/2017/5.html

[31] EDPB, "EDPB Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research", 2021

[32] Note that, when the article mentions "public interest", this expression should only be considered when processing data for archiving purposes, which means that the provisions of art. 89 apply to public or private entities when data is processed for scientific research purposes.

legislation stating further conditions for the processing of certain types of sensitive data[33], the "safeguards" mentioned in art. 89 could be considered as such conditions[34].

Moreover, when the legislator assumes that data is processed for scientific research purposes, there is already a set of good practices, methodologies and processes that are sufficiently recognized, accepted and established in rules of ethical and professional nature, especially with regard to scientific research on humans[35]. Because of this, GDPR provides for several exceptions[36], such as the conditions of lawfulness[37], the principle of purpose limitation[38] and retention[39], as well as on the rights of data subjects[40].

Certain conditions apply to the exceptions on the rights of data subjects: i) even when said exceptions apply, the requirements of art. 89(1) still need to be verified; ii) in order to apply these derogations, exercise of these rights must be "likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes"[41]; iii) the derogations only apply to the purposes mentioned in art. 89(2) and 89(3). This means that, for example, derogations for the following use of health research data[42] for commercial purposes are not applied. Note that differentiation can be difficult between the original

[33] In particular, genetic data, biometric data or data concerning health.

[34] EDPB, "Study on the appropriate safeguards under Article 89(1) GDPR for the processing of personal data for scientific research", 2019

[35] For example, the Geneva Declaration of the World Medical Association states that "I will respect the secrets that are confided in me, even after the patient has died."; the International Code of Medical Ethics states that "The physician must respect the patient's privacy and confidentiality, even after the patient has died."

[36] CANTO MONIZ Graça, Lesson taught as part of "IV Pós Graduação Avançada em Direito da Proteção de Dados Pessoais", subordinated to the theme "O Tratamento de Dados Pessoais para Fins de Investigação Científica", 2022

[37] Art. 6(1) and 9.

[38] Art. 5(1)(b).

[39] Art. 5(1)(e).

[40] Art. 15, 16, 18 and 21, ex vi art. 89(2), as well as art. 14(5)(b)

[41] Arts. 89(2) and 89(3).

[42] Thus, collected for the purposes of scientific research.

scientific research purpose and subsequent purposes for the same data processing[43].

However, it must be highlighted that ethics committees play an important role in guaranteeing that the fundamental right to data protection (as well as other human rights) is integrated in scientific research, ensuring that research projects are already designed from the very beginning with data protection principles in mind. Ethical committees should, then, help in the understanding of which activities qualify as "scientific research", as well as defining the ethical standards mentioned in the GDPR[44].

b. Limitations regarding Data Processing for Scientific Research Purposes

The introduction of the GDPR was expected to harmonize the legal framework for data protection in the EU. However, this desired result has not been fully achieved, which is especially visible in the implementation of the "safeguards" provided for in art. 89(1). This is not only due to the fact that the definition of important concepts is reserved for the recitals[45], but there is also a favourable discussion on Member States being obliged to develop specific legislation regarding the aforementioned "safeguards"[46-47]. This circumstance can promote "different interpretations and forum shopping, which may erode the individuals' privacy, since the Member States do not want to fall back in the field scientific research and losing income"[48].

[43] WIESE SVANBERG, *Christian, The EU General Data Protection Regulation (GDPR): A Commentary*, Oxford Academic, 2020

[44] EDPS, "A Preliminary Opinion on data protection and scientific research", 2020

[45] Note the fact that just for art. 89 there are several recitals that correspond to it, in particular, recitals 156 to 163.

[46] Others defend that art. 89 only imposes obligations for the researchers as controllers to implement safeguards. Thus, it would be enough that controllers establish the safeguards, which can include implementing guidelines or codes of conduct.

[47] EDPB, "Study on the appropriate safeguards under art. 89(1) GDPR for the processing of personal data for scientific research", 2019

[48] MESZAROS, Janos, "The Conflict Between Privacy and Scientific Research in the GDPR", *2018 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, 2018, available at: https://ieeexplore.ieee.org/document/8579471

The GDPR also presents limitations regarding the processing of health data for the purposes of scientific research, including the necessity to specify several conditions for the publication or disclosure of said data. Also, the GDPR lies on the need to specify the nature and scope of the research, due to the fact that the data processing requires specified, explicit and legitimate purposes[49]. On one hand, this restricts the ability to repurpose personal data collected for a specific project for other uses[50]. On the other hand, it can be an obstacle to the development of the project being carried out, for example, in the cases where the dataset is used to formulate the scope of the actual research. Thus, this requires researchers to already have, from the beginning, a proper understanding and establishment of the research project's nature and purposes[51].

iii. Is Synthetic Data Relevant to all types of Scientific Research?

The use of synthetic data may be more relevant in certain types of scientific research. A study concluded that when many patients are used in comparison to the number of variables, there is higher accuracy and consistency of results between synthetic and the original data, while in respect of research that uses smaller populations, predictions were of moderate accuracy, yet clear trends were correctly observed[52]. For example, some scientific research situations in which synthetic data is mostly applied include simulation studies and predictive analytics,

[49] Art. 5(1)(b)

[50] WOLK DER VAN, Alex, "The (Im)Possibilities of Scientific Research Under the GDPR", *Cybersecurity Law Report*, 2017, available at: https://www.mofo.com/resources/insights/200617-scientific-research-gdpr

[51] *Ibidem.*

[52] REINER BENAIM, Anat; ALMOG, Ronit; GORELIK, Yuri; HOCKBERG, Irit; NASSAR, Laila; MASHIACH, Tanya; KHAMAISI, Mogher; LURIE, Yael; AZZAM, Zaher; KHOURY, Johad; KURNIK, Daniel; BEYAR, Rafael, "Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies", *JMIR Med Inform,* VIII, 2, 2020, available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7059086/

algorithms, hypothesis, and methods testing and epidemiological study/ public health research.

When it comes to simulation and prediction research, there is the necessity to use large datasets from real world sources, in order to properly anticipate certain behaviors and outcomes. Considering this, synthetic data emerge as a replacement or a supplement to real world sources, given the fact that researchers are allowed to scale the sample size or even to add variables not considered in the original data set[53]. Regarding epidemiological studies and public health research, datasets present limitations in terms of size and quality, as well as being expensive. This was especially evident in the Covid-19 pandemic, when the need of making data publicly available was felt[54]. To face these difficulties, synthetic data allows the improvement of the "timeliness of data release, support[s] researchers in doing real-time computational epidemiology, provide[s] a more convenient sample for sensitivity analyses, and build[s] a more extensive test set for improving disease detection algorithms"[55]. Finally, it has been proven that synthetic data shows higher potential for prediction, enhancing diagnostics, and comprehending risk factors.

[53] Note its usage, amongst others, in "disease-specific hybrid simulation and microsimulation for testing policy options and health care financing strategies evaluation. Studies also used synthetic data to validate simulation and prediction models and to improve prediction accuracy", REINER BENAIM, Anat; ALMOG, Ronit; GORELIK, Yuri; HOCHBERG, Irit; NASSAR, Laila; MASHIACH, Tanya; KHHAMAISI, Mogher; LURIE, Yael; AZZAM, Zaher: KHOURY, Johad; KURNIK, Daniel; BEYAR, Rafael, "Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies", *JMIR Med Inform*, VIII, 2, 2020, available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7059086/

[54] "Use of synthetic data in the area of clinical research is limited at the moment. However, users need to understand the source and the way the synthetic dataset was generated to evaluate its appropriateness for specific types of studies or specific stages of research", GONZALES, Aldren; GURUSWAMY, Guruprabha; SMITH, Scott, "Synthetic data in health care: A narrative review", *PLOS Digit Health,* II, 1, 2023, available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931305/

[55] *Ibidem*.

III. Synthetic Data in the Light of the GDPR

i. What is Synthetic Data?

The definition of synthetic data has been the subject of some controversy in the literature[56-57]. Scholars usually make the distinction between real and not real data, insofar as synthetic data consists mostly of not real data (that are artificially manufactured with or without the original real data)[58]. However, this approach has been somewhat – and rightfully so – criticized by the EDPS[59]: if it is considered that the "real" data is "truthful" data, it's not correct to assume that synthetic data is "not real" only because the data is artificially generated. As it will be analysed[60], not only can synthetic data be categorized based on the amount of interference the original data has in the dataset, but even data that has been entirely generated by artificial intelligence can be considered partly "real", insofar as it may be possible to re-identify the data subject. Other scholars argue that "synthetic data" only refers to

[56] Consider, for example, the definition given by the US Census Bureau (taken from a ONS UK Working Paper): "Microdata records created by statistically modeling original data and then using those models to generate new data values that reproduce the original data's statistical properties. This definition highlights the strategic use of synthetic data because it improves data utility while preserving the privacy and confidentiality of information". Regarding this definition, UK's Office for National Statistics, "ONS methodology working paper series number 16 – Synthetic data pilot", 2021, available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot;

[57] GONZALES, Aldren; GURUSWAMY, Guruprabha; SMITH, Scott, "Synthetic data in health care: A narrative review", *PLOS Digit Health,* 2023, available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082

[58] *Ibidem*.

[59] IPEN; "Synthetic data: what use cases as a privacy enhancing technology?" – Webinar; 2021; available at: https://www.edps.europa.eu/data-protection/our-work/ipen/ipen-webinar--2021-synthetic-data-what-use-cases-privacy-enhancing_en (remarks made by Wojciech Wiewiórowski).

[60] Section III, Item ii.

datasets incorporating merely fabricated data and without any original record[61].

Nonetheless, the classification of synthetic data has not been considered rigid[62]. Statistics field scholars divide synthetic data in fully synthetic, partially synthetic, and hybrid. A more detailed spectrum of synthetic data types is described in a working paper series by the United Kingdom's Office for National Statistics (UK's ONS). The spectrum features six levels under the synthetic and synthetically augmented dataset types[63].

That said, fully synthetic data are identified as a dataset that is completely synthetic and has a strong privacy control – given the fact that there is no correlation with the dataset generated, although having low analytic value[64]. Regarding partially synthetic data, a selection of the original dataset is replaced by synthetic data. Despite having a lower privacy control – since the dataset still contains original data –, this type of data has a higher analytic value[65]. At last, hybrid synthetic data is the result of both original and synthetic data. Despite its generation being more time and memory consuming, this type of data maintains privacy control characteristics with high effectiveness[66].

[61] UK's Office for National Statistics, "ONS methodology working paper series number 16 – Synthetic data pilot", 2021, available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot; SIWICKI, Billy, "Is synthetic data the key to healthcare clinical and business intelligence?", *Healthcare IT News*, 2020, available at: https://www.healthcareitnews.com/news/synthetic-data-key-healthcare-clinical-and-business-intelligence.

[62] GIUFFRÈ, Mauro; SHUNG, Dennis, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy", *npj Digital Medicine*, VI, 2023, available at: https://www.nature.com/articles/s41746-023-00927-3#citeas

[63] UK's Office for National Statistics, "ONS methodology working paper series number 16 – Synthetic data pilot", 2021, available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaper seriesnumber16syntheticdatapilot;

[64] GONZALES, Aldren; GURUSWAMY, Guruprabha; SMITH, Scott, "Synthetic data in health care: A narrative review", *PLOS Digit Health,* 2023, available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082

[65] *Ibidem*.

[66] *Ibidem*.

ii. The dogmatics of synthetic data in the light of the GDPR; (non) personal data?

Synthetic data corresponds to artificial, algorithmically generated data that, by closely mimicking the properties and relations of the source data, can be used for the same purposes[67].

The applicability of the processing of synthetic data to the GDPR depends on its dogmatic insertion in the category of "personal data", which means knowing to what extent synthetic data constitutes "information relating to an identified or identifiable natural person"[68].

Although the artificial nature of synthetic data shows that it cannot be considered data that directly identifies a natural person, it can still be information relating to an identifiable person, which makes its classification subject to the "reasonability criteria", as provided for in recital 26. If considering "all the means likely reasonably to be used", the possibility of identifying a natural person does not exist or is negligible, the information will not be considered "personal data"[69]. Having said that, one must examine the possibility of establishing a correlation between the synthetic data and the original data and, consequently, (re) identifying the data subject[70].

Despite a case-by-case basis being the key to assess whether a set of synthetic data is personal data, it is necessary to establish various factors to determine the applicability of the "reasonableness criteria", being one of them the purpose for which synthetic data is generated. Note that synthetic data is usually viewed as a form of anonymization or as a privacy-enhancing technology[10], which means that re-identification will be more difficult. It can't be said the same when synthesizing is

[67] MITCHELL, Colin; REDRUP HILL, Elizabeth, "Are Synthetic Health Data "Personal Data"?", *PHG Foundation Report*, 2023

[68] Art. 4(1)

[69] BARRETO MENEZES CORDEIRO, António, *Direito da Proteção de Dados – À luz do RGPD e da Lei n.º 58/2019*, Almedina, 2020

[70] MITCHELL, Colin; REDRUP HILL, Elizabeth, "Are Synthetic Health Data "Personal Data"?", *PHG Foundation Report*, 2023, available at: https://www.phgfoundation.org/publications/reports/are-synthetic-health-data-personal-data/

intended for filling gaps in essential information required to test products or software[11] but also to uncover biases in real-world data[12].

Another factor is related to the type of original data used. For example, a process that relies on statistical data to feed the algorithm has lower chances of re-identification[13], in comparison with a process that uses large sums of directly identifiable personal data.

Moreover, the higher the utility of a synthetic data set, the lower its anonymity, which means that the type of synthesizing process is also relevant to determine re-identification risks. An approach that resorts to advanced machine learning techniques has a better chance of establishing relationships between the original data and the synthetic data[14], in contrast to one that starts with source data manually manipulated[15].

We also understand that the classification of synthetic data as personal data also requires an examination of the pertinent legal definitions, especially concerning the understanding of identifiability and anonymization within the realm of data protection law[16]. In this sense, "the bar of anonymization has been set very high by the European legislator"[17], which means that "synthetic data [is] capable of being considered personal, 'pseudonymous' or anonymous depending on interpretation and context[18]."

IV. Regulatory State of the Art in the EU

The last few years have demonstrated an absence of an overall "EU data architecture, harmonization, and collaboration in data-sharing practices across countries, hindering the ability to support regulatory decision making based on RWE [Real World Evidence], and efficiently address public health challenges".[71] In order to circumvent these issues,

[71] Clara ALLOZA, Bethany KNOX, Hanaya RAAD, Mireia AGUILÀ, Ciara COAKLEY, Zuzana MOHROVA, Élodie BOIN, Marc BÉNARD, Jessica DAVIES, Emmanuelle JACQUOT, Coralie LECOMTE, Alban FABRE, Michael BATECH; "A Case for Synthetic Data in Regulatory Decision-Making in Europe", *PubMed*, CXIV, 4, 2023, 795-801; available at: https://pubmed.ncbi.nlm.nih.gov/37441734/

the establishment of the European Health Data Agency (EHDA)[72] was proposed.

Even though synthetic data have not yet been considered in any regulatory process, there have been some initiatives by the EU to implement the use of this type of data for healthcare research purposes. Some initiatives include the European Medicines Agency's joint statement calling for international collaboration to enable RWE for regulatory decision-making[73]. This collaboration would mainly be carried out by the International Coalition of Medicines Regulatory Authorities.

However, two pieces of legislation are particularly important: the Data Governance Act and the European Health Data Space Act. The latter, by determining and developing health data processing across Europe, provides a framework toward a more comprehensive (Health) Data Governance Act[74], where synthetic data are pointed to have a crucial role and seen as a solution for some of the public health sector needs. Other important initiatives include a Webinar focused on the use of synthetic data as a possible technology to mitigate data protection risk led by the EDPS[75], the Horizon Europe project[76], the Data Analysis

[72] Proposed by the Panel for the Future of Science and Technology, available at: https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2021)690009

[73] This decision is available at: https://www.ema.europa.eu/en/news/global-regulators-call-international-collaboration-integrate-real-world-evidence-regulatorydecision;

[74] Clara ALLOZA, Bethany KNOX, Hanaya RAAD, Mireia AGUILÀ, Ciara COAKLEY, Zuzana MOHROVA, Élodie BOIN, Marc BÉNARD, Jessica DAVIES, Emmanuelle JACQUOT, Coralie LECOMTE, Alban FABRE, Michael BATECH; "A Case for Synthetic Data in Regulatory Decision-Making in Europe", *PubMed*, CXIV, 4, 2023, 795-801; available at: https://pubmed.ncbi.nlm.nih.gov/37441734/

[75] IPEN; "Synthetic data: what use cases as a privacy enhancing technology?" – Webinar; 2021; available at: https://www.edps.europa.eu/data-protection/our-work/ipen/ipen-webinar-2021-synthetic-data-what-use-cases-privacy-enhancing_en

[76] This project provides new methods for the effective use of real-world data and/or synthetic data in regulatory decision-making and/or in health technology assessment. It is available at: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-hlth-2022-tool-11-02

and the Real-World Interrogation Network[77]. Moreover, data protection authorities, such as ICO[78] and CNIL, have been exploring their benefits as a privacy enhancing technology[79]. All these initiatives are particularly relevant to leverage the rapid collaboration that is taking place between government institutions, the academia, and the private sector[80].

V. Benefits of the use of Synthetic Data; Applying Synthetic Data in Healthcare Research

Synthetic data has brought a turning point to the field of scientific research in the area of healthcare, as it introduces a number of innovative applications.

Some applications include the use of synthetic data in medical imaging analysis, simulation of clinical trials regarding drug discovery and development, tailoring treatment plans for personalized medicine, improving readiness and reactive measures for public health analysis and prediction, as well as in supporting clinical decisions[81]. Synthetic data can also give rise to outputs that include images, audio files, and

[77] It provides real-world evidence from across Europe on diseases, populations and the uses and performance of medicines, which allows EMA and national competent authorities in the European medicines regulatory network to use these data whenever needed throughout the lifecycle of a medicinal product. It is available at: https://www.darwin-eu.org/

[78] MARSHALL, Valerie; MARKHAM, Charlie; AVRAMOVIC, Pavle; COMERFORD, Paul; MAPLE, Carsten; SZPRUCH, Lukasz, "Exploring Synthetic Data Validation – Privacy, Utility and Fidelity", *FCA Report/ICO Research Paper*, 2023, available at: https://www.fca.org.uk/publications/research-articles/exploring-synthetic-data-validation-privacy-utility-fidelity;

[79] CNIL, "Artificial Intelligence: The CNIL Publishes a Set of Resources for Professionals", 2022, available at: https://www.cnil.fr/en/artificial-intelligence-cnil-publishes-set-resources-professionals

[80] ALLOZA, Clara; KNOX, Bethany; RAAD, Hanaya; AGUILÀ, Mireia; COAKLEY, Ciara; MOHROVA, Zuzana; BOIN, Élodie; BÉNARD, Marc; DAVIES, Jessica; JACQUOT, Emmanuelle; LECOMTE, Coralie; FABRE, Alban; BATECH, Michael; "A Case for Synthetic Data in Regulatory Decision-Making in Europe", *PubMed*, CXIV, 4, 2023, 795-801; available at: https://pubmed.ncbi.nlm.nih.gov/37441734/

[81] MEHTA, Yash; "Resolving Healthcare's Prime Challenges through Synthetic Data Generation"; available at: https://datafloq.com/read/healthcare-challenges-synthetic-data-generation/

videos[82], enabling the generation of CT-like images, as well as it allows to overcome errors, inaccuracies, and bias of real-world data, by ensuring higher data quality and variety[83]. That being said, the increasing interest in utilizing synthetic data for medical purposes has prompted the development of various tools (such as Synthea[84] or MDClone's Synthetic Data Engine).

However, the main valuable aspect for the processing of synthetic data is in the fact that it "enables organizations to access concrete and representative insights from sensitive data, while minimizing the risk to patient privacy and limiting governance requirements.[85]". When it comes to patient privacy, the synthesizing process is usually employed as a privacy enhancing technique, which means that the risk of identifying the data subject is reduced – especially given that anonymous data can be generated.

Considering this, the implementation of these synthesis processes is considered to be a technical measure aimed at achieving compliance with the principle of minimisation and, consequently, being a prime example of an "appropriate safeguard", under the terms of art. 89(1). Nevertheless, it is important to take into account, according to recital 156, that these data minimization techniques must be "in pursuance of the proportionality and necessity principles". This means that, on a case-by-case basis, it is necessary to weigh up the advantages that data synthesis brings by minimizing the risk of

[82] GAL, Michal; "Synthetic Data: Legal Implications of the Data-Generation Revolution", *SSRN Electronic Journal*, 2023; available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4414385#

[83] LAMBERTI, Aldo, "The Benefits and Limitations of Generating Synthetic Data", 2023; available at: https://syntheticus.ai/blog/the-benefits-and-limitations-of-generating-synthetic-data

[84] WALONOSKI, Jason; KRAMER, Mark; NICHOLS, Joseph; QUINA, Andre; MOESEL, Chris; HALL, Dylan; DUFFETT, Carlton; DUBE, Kudakwashe; GALLAGHER, Thomas; MCLACHLAN, Scott; "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record"; *Journal of the American Medical Informatics Association*, XXV, 3, 2018, 230-238; available at: https://pubmed.ncbi.nlm.nih.gov/29025144/

[85] LESHIN, Jonah; JOHNS, Quinn; "The Value of Synthetic Data in Healthcare"; Available at: https://www.datavant.com/blog/synthetic-data-healthcare

identifying the data subject against other disadvantages that may arise from its application[86].

For these reasons one must consider that, within the context of a "data protection by design and by default" approach[87], the protection of the rights of data subjects is more effective. For example, research done on the usage of partially synthetic data as a proxy for original data in large-scale health surveys has shown to be successful, proving that the data, while guaranteeing patient privacy, still allowed researchers to perform analysis[88]. Moreover, facing the obstacle of limited reproducibility for clinical research findings caused by data protection legislations, synthetic data allows continuous research conduction, since the sharing of synthetic patient datasets means that clinical researchers can ensure that their results and case studies are replicable[89-90].

It could also be said that the fact that synthetic data often involves the processing of a large volume and variety of personal data makes it difficult to fulfil the various requirements for consent, particularly the fact that said consent must be "informed". However, art. 14(5)(b) already mentions a derogation from the various information duties, if "the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for (...) scientific research purposes". This derogation is "subject to the conditions and safeguards referred to in Article 89(1)" or if the information requirements are likely to render impossible or seriously impair the achievement

[86] Without prejudice to the disadvantages mentioned in the next section, it is necessary to consider (again, taking into account the individual circumstances of the case) any other risks that may exist for data subjects.

[87] Art. 25

[88] LOONG, Bronwyn; ZASLAVSKY; Alan M; HE, Yulei; P HARRINGTON David; "Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS", *Statistics in Medicine Journal*, XXXII, 24, 2013, 39-61; available at: https://pubmed.ncbi.nlm.nih.gov/23670983/

[89] DILMEGANI, Cem; "Synthetic Data for Healthcare: Benefits & Case Studies in 2024"; Available at: https://research.aimultiple.com/synthetic-data-healthcare/

[90] KAABACHI, Bayrem; DESPRAZ, Jérémie; MEURERS, Thierry; OTTE, Karen; HALILOVIC, Mehmed, PRASSER, Fabian; RAISARO Jean Louis; "Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics"; *MedRxiv*, 2023.

of the objectives of that processing. Thus, the application of this article is particularly relevant when synthetic data is processed for scientific research purposes, especially since the dataset is almost never obtained directly from the data subject.

When determining what constitutes a disproportionate effort, recital 62 refers to "the number of data subjects, the age of the data and any appropriate safeguards adopted". In other words, despite the fact that, as mentioned above, data synthesis techniques are considered an appropriate safeguard that largely reduces the risk of identifying the data subject, it is nonetheless necessary that the controller carries out a "balancing exercise to assess the effort involved to provide the information to data subjects against the impact and effects on the data subject if they are not provided with the information"[91].

Regarding governance requirements, synthetic data contributes to the data sharing between researchers and organizations, not only because it bypasses some of the current restrictions on data transfers, but also because it circumvents some of the ethical standards. For example, if a company situated in a country outside the European Economic Area (EEA) wants to access EU health data, in order to develop activities in the EU market, the sharing of patient-level data is especially hindered not only by ethical requirements but, specially, by the GDPR data transfer restrictions, imposing the need to implement, among others, appropriate safeguards[92], such as Standard Contractual Clauses (SCCs)[93]. However, if a synthetic dataset is created, the data exporter can more easily comply with the respective standards, which also allows saving resources, such as time and money[94]. In addition, concerning the saving of these two last resources, it has been pointed out that, with synthetic data, costs involved in all stages of the data lifecycle can be reduced.

[91] EDPB, "Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak", 2020

[92] Art. 46

[93] Art. 47

[94] LESHIN, Jonah; JOHNS, Quinn; "The Value of Synthetic Data in Healthcare"; Available at: https://www.datavant.com/blog/synthetic-data-healthcare

For instance, by developing a digital twin of the hospital, administrators were able to optimize staffing levels and allocate resources more efficiently[95].

Synthetic data has also shown promise in healthcare research to improve risk assessment, predictive analysis and protecting patient well-being, while maintaining ethical standards[96]. One of its uses is related to the improvement of policy implications. A study focused on demographic aging used variations regarding morbidity, disability, and doctor behavior to explore positive and negative policy outcomes on healthcare demand and resource utilization[97].

In addition, this type of data was shown to be useful in improving data scarcity and, consequently, raising data volume in imaging studies in the context of the COVID-19 pandemic[98]. Therefore, it can be applied in clinical challenges involving large populations in an epidemiological context.

VI. Disadvantages of the use of Synthetic Data

Despite the importance that synthetic data carries in terms of data processing in the healthcare domain, several concerns arise. Some disadvantages include the difficulty to generate the datasets, bias

[95] CHENG, Weibin; LIAN, Wanmin, TIAN, Junzhang; "Building the hospital intelligent twins for all-scenario intelligence health care"; *Digit Health*; 2022; available at: https://pubmed.ncbi.nlm.nih.gov/35720617/

[96] GIUFFRÈ, Mauro; SHUNG, Dennis, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy", *npj Digital Medicine*, VI, 2023, available at: https://www.nature.com/articles/s41746-023-00927-3#citeas

[97] DAVIS, Peter; LAY-YEE, Roy; PEARSON, Janet; "Using micro-simulation to create a synthesised data set and test policy options: the case of health service effects under demographic ageing"; *Health Policy*, XCVII, 2, 2023, 267-274; available at: https://pubmed.ncbi.nlm.nih.gov/20800762/

[98] JIANG, Yifan; CHEN, Han; LOEW, Murray; KO, Hanseok; "COVID-19 CT Image Synthesis with a Conditional Generative Adversarial Network", *IEEE journal of biomedical and health informatics*, XXV, 2, 2021, 441 – 452

amplification and the complexity and lack of rigorous methods for assessing data quality[99].

First of all, complex data are difficult to generate, since the effectiveness of synthetic data generation techniques is most pronounced when the generated data is simple and can be defined by a set of rules or patterns. Generating intricate data, such as natural language text or more realistic images, poses a greater challenge and necessitates the application of more advanced techniques[100]. Furthermore, since the effectiveness of the model is tied to the source data quality, there is a risk of amplifying the bias inherent in the original dataset. This problem mostly results from the techniques used for data collection. Not only that, but one must also need to consider that the (generative) methods employed to obtain the new dataset can also lead to problems regarding automation bias[101]. Consequently, the risk inherent in (clinical) decisions made on the basis of this synthetic data can be wrongly assessed, perpetuating inequalities and contributing to the discrimination of vulnerable populations. For example, when the algorithm is trained with data that mostly considers people from a determined ethnicity, the synthetic dataset generated will end up mirroring this disequilibrium. Although this problem could be solved by oversampling the less

[99] GIUFFRÈ, Mauro; SHUNG, Dennis, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy", *npj Digital Medicine*, VI, 2023, available at: https://www.nature.com/articles/s41746-023-00927-3#citeas

[100] Note the case when "the original data to be synthesized (e.g., data acquired in Living Labs) may consist of subjects' metadata (static) and a longitudinal component (set of time-dependent measurements), making it challenging to produce coherent synthetic counterparts." Regarding these considerations, ISASA, Imanol; HERNANDEZ, Mikel; EPELDE, Gorka; LONDOÑO, Francisco; BERISTAIN, Andoni; LARREA, Xabat, ALBERDI, Ane; BAMIDIS, Panagiotis; KONSTANTINIDIS Evdokimos, "Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis", *BMC Medical Informatics and Decision Making*, XXIV, 27, 2024; available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02427-0

[101] "There are considerable challenges associated with interpreting synthetic data generation models, including the black-box nature of generation algorithms, limitations in the evaluation metrics, and the potential for overfitting or underfitting". Regarding these considerations: GIUFFRÈ, Mauro; SHUNG, Dennis, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy", *npj Digital Medicine*, VI, 2023, available at: https://www.nature.com/articles/s41746-023-00927-3#citeas

present characteristics, there is still a "risk of overgeneralization and potential creation of non-existent or incorrect correlations"[102], which can lead to the misrepresentation of the vulnerable population samples and the corresponding clinical profiles.

Another challenge associated with synthetic data is in the complexity of validating its accuracy[103]. Even if a synthetic dataset appears realistic and precise, determining whether it faithfully represents the underlying patterns of real-world data is challenging[104]. Consequently, it is not possible to assure that a model trained on synthetic data will demonstrate accuracy when deployed in the real-world context[105], since there is a lack of robust methods for assessing data quality. The main issue is related to the fact that the majority of the techniques employed for synthetic data generation do not consider the complexity and diversity of the possible different medical scenarios. Furthermore, since machine learning models have the tendency to excessively adapt the original data, leaks concerning individual records could end up being memorized, which means a higher reidentification risk[106].

Finally, another disadvantage of synthetic data results from the fact that the model used has a huge influence in the quality of the outcome

[102] GIUFFRÈ, Mauro; SHUNG, Dennis, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy", *npj Digital Medicine*, VI, 2023, available at: https://www.nature.com/articles/s41746-023-00927-3#citeas

[103] GONZALES, Aldren; GURUSWAMY, Guruprabha; SMITH, Scott, "Synthetic data in health care: A narrative review", *PLOS Digit Health*, II, 1, 2023, available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082

[104] Not only that, but note that the black-box nature of GANs makes it harder to foresee which data utility is maintained/lost in the creation process (this is especially relevant when it comes to sensitive data). Regarding these considerations, KAABACHI, Bayrem; DESPRAZ Jérémie, MEURERS Thierry; OTTE, Karen; HALILOVIC, Mehmed; PRASSER, Fabian; RAISARO, Jean Louis, "Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics", *MedRxiv*, 2023, available at: https://www.medrxiv.org/content/10.1101/2023.11.28.23299124v1.full

[105] LAMBERTI, Aldo, "The Benefits and Limitations of Generating Synthetic Data"; available at: https://syntheticus.ai/blog/the-benefits-and-limitations-of-generating-synthetic-data

[106] KAABACHI, Bayrem; DESPRAZ Jérémie, MEURERS Thierry; OTTE, Karen; HALILOVIC, Mehmed; PRASSER, Fabian; RAISARO, Jean Louis, "Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics", *MedRxiv*, 2023, available at: https://www.medrxiv.org/content/10.1101/2023.11.28.23299124v1.full

data. Considering this, it is important to understand that it may be vulnerable to statistical noise, which can lead to incorrect classification of the data and the production of strongly unreliable outputs[107].

VII. Conclusion

As previously mentioned, processing of personal data for purposes of healthcare scientific research has faced significant obstacles, namely, the necessity to overcome limitations on data protection legislation and other regulatory restrictions, mostly related to ethical aspects for sharing and securing data. Other limitations include data fragmentation, monetary constraints on accessing data outside the public domain and the lack of high-quality datasets.

Synthetic data has been one of the solutions presented, introducing several innovative applications. This type of data (considered a "privacy enhancing technique"[108]) makes it difficult to identify the data subject since it allows the risk associated with processing activities to be mitigated to a large extent. Consequently, this can even lead to the non-application of data protection legislation when anonymous data is at stake (note that the GDPR framework does not apply to anonymous data[109]). Remember that the classification of synthetic data as personal data depends on the compliance with the "reasonableness criteria"[110]. This is particularly evident since the synthesis processes should be considered a technical measure for the purposes of achieving data minimization and, consequently, an "appropriate safeguard" in the terms of art. 89(1). That being said, compliance with data protection legislation is greatly facilitated, which will consequently make it easier to

[107] GONZALES, Aldren; GURUSWAMY, Guruprabha; SMITH, Scott, "Synthetic data in health care: A narrative review", *PLOS Digit Health*, II, 1, 2023, available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082

[108] Section V.

[109] Note recital 26

[110] Section III, Item ii

comply with ethical rules or requirements and, by that, allowing for researchers to easily share data between them, especially when dealing with international transfers.

Synthetic data carries several other benefits, such as reducing operational and data accessing costs (especially when accessing health related datasets outside of the public domain), providing medical data resources that have higher quality and variety (namely, images, audio-files, videos) – what, consequently, helps in overcoming inaccuracies, bias of real-world, as well as data fragmentation – and, lastly, improving policy implications. Thus, synthetic data allows for continuous research conduction, since, among other factors, the results and case studies are easier to replicate[111-112].

Although there is no regulatory landscape specifically aimed at the processing of synthetic data, there have been some initiatives to implement its use for healthcare research purposes. However, it is still necessary to create legal guidelines, best practices and other recommendations that ensure consistency and reliability of the synthetic data created, which would require the intervention of major stakeholders – such as the EDPS and the European Data Protection Board (EDPB) –, as well of national data protection authorities and the European Health Data Space. This means that, on a regulatory level, it is still important to establish consistency around operational, methodological, and technical matters[113].

Nevertheless, the usage of synthetic data also presents some relevant disadvantages, such as the difficulty to generate the dataset, bias

[111] DILMEGANI, Cem; "Synthetic Data for Healthcare: Benefits & Case Studies in 2024"; Available at: https://research.aimultiple.com/synthetic-data-healthcare/

[112] KAABACHI, Bayrem; DESPRAZ, Jérémie; MEURERS, Thierry; OTTE, Karen; HALILOVIC, Mehmed, PRASSER, Fabian; RAISARO Jean Louis; "Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics"; *MedRxiv*, 2023; available at: https://www.medrxiv.org/content/10.1101/2023.11.28.23299124v1.full

[113] ALLOZA, Clara; KNOX, Bethany; RAAD, Hanaya; AGUILÀ, Mireia; COAKLEY, Ciara; MOHROVA, Zuzana; BOIN, Élodie; BÉNARD, Marc; DAVIES, Jessica; JACQUOT, Emmanuelle; LECOMTE Coralie; FABRE, Alban; BATECH, Michael, "A Case for Synthetic Data in Regulatory Decision-Making in Europe", CXIV, 4, 2023, 795-801; available at: https://ascpt.onlinelibrary.wiley.com/doi/10.1002/cpt.3001

amplification and the complexity and lack of rigorous methods for assessing quality of the data and of the tools used. These challenges can be overcome through some mitigation measures, such as the establishment of guidelines and policies in order to properly carry out the synthetic data generation throughout the lifecycle of the research project[114]. This also involves ensuring documentation not only on the data generation process, but also on possible limitations and the existence of bias, which is useful in identifying actual and potential errors[115]. It is also important to ensure that various public and private entities join forces to collaborate, namely public and private data owners, healthcare solution developers, and synthetic data experts, guaranteeing multidisciplinary insights[116].

In particular, note the establishment of proper synthetic data generation lifecycle frameworks[117] and the analysis of clinical quality measures as ways to provide researchers a better approach to define and describe the process of validating synthetic datasets. In order to properly validate the quality of the mechanisms for synthetic data generation, the creation of domain-specific evaluation metrics and benchmarks particularly tailored for healthcare applications should be considered. On the one hand, these evaluation metrics can be created

[114] J. CHEN, Richard; LU, Ming Y; Y. CHEN, Tiffany; F. K. WILLIAMSON, Drew; MAHMOOD, Faisal; "Synthetic data in machine learning for medicine and healthcare"; *Nature Biomedical Engineering volume*, V, 2021, 493-497.

[115] GIUFFRÈ, Mauro; SHUNG, Dennis, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy", *npj Digital Medicine*, VI, 2023, available at: https://www.nature.com/articles/s41746-023-00927-3#citeas

[116] IHI; "Maximising the potential of synthetic data generation in healthcare applications"; *A Innovative Health Initiative Call Topic*; 2023; available at: https://health.ec.europa.eu/ehealth--digital-health-and-care/european-health-data-space_en

[117] Note the case of the ATEN Framework, that is composed of other three approaches that provide for: synthetic data generation; a way to define and describe the elements of realism, and; an approach for validating synthetic data. Regarding these considerations, MCLACHLAN, Scott; DUBE, Kudakwashe; GALLAGHER, Thomas; A. WALONOSKI, Jason; "The ATEN Framework for Creating the Realistic Synthetic Electronic Health Record", *HEALTHINF, SCITEPRESS – Science and Technology Publications*, 5, 2018, 220-230. available at: https://www.researchgate.net/publication/322653777_The_ATEN_Framework_for_Creating_the_Realistic_Synthetic_Electronic_Health_Record

by the development of platforms that also provide methods for assessing utility and privacy and, consequently, streamlining the evaluation process[118]. On the other hand, a benchmark that accurately allows the representation of a wide spectrum of real-world medical scenarios, provides a way to properly compare the performance of several methods for the purposes of creating synthetic data[119]. It is also crucial that the generation of synthetic data is accompanied by consistent evaluations to minimize biases, including the implementation of auditing methods. Strategies include the implementation of anomaly detection techniques[120] and other advanced statistical methods such as distribution matching, correlation analysis, and dimensionality reduction, that allow capturing the complex correlations and patterns inherent in the original dataset, improving the data's representativeness[121].

After balancing the advantages and disadvantages of applying and using synthetic data, as well as possible solutions to overcome the latter, it is possible to conclude that the use of this type of data for the purposes of scientific healthcare research does indeed present the potential to be considered a "Holy Grail". However, there is still a long way to go, especially in regulatory terms.

[118] KAABACHI, Bayrem; DESPRAZ Jérémie, MEURERS Thierry; OTTE, Karen; HALILOVIC, Mehmed; PRASSER, Fabian; RAISARO, Jean Louis, "Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics", *MedRxiv,* 2023, available at: https://www.medrxiv.org/content/10.1101/2023.11.28.23299124v1.full

[119] GIUFFRÈ, Mauro; SHUNG, Dennis, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy", *npj Digital Medicine*, 6, 2023, available at: https://www.nature.com/articles/s41746-023-00927-3#citeas

[120] Since overconfidence problems are common in synthetic data generation, these techniques allow to "identify instances that deviate significantly from the training data distribution helping to detect and handle out-of-distribution problems". Regarding these considerations, GIUFFRÈ, Mauro; SHUNG, Dennis, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy", *npj Digital Medicine*, 6, 2023, available at: https://www.nature.com/articles/s41746-023-00927-3#citeas; MÖLLER, Felix; BOTACHE, Diego; HUSELJIC, Denis; HEIDECKER, Florian; BIESHAAR, Maarten; SICK, Bernhard; "Out-of-distribution Detection and Generation using Soft Brownian Offset Sampling and Autoencoders", *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[121] GIUFFRÈ, Mauro; SHUNG, Dennis, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy", *npj Digital Medicine*, 6, 2023, available at: https://www.nature.com/articles/s41746-023-00927-3#citeas