A new solution to biomedical data sharing

A Discussion Paper on Synthetic Data by Edwin Morley-Fletcher

DRAFT

1. A Europe fit for the digital age?

The draft of *Europe fit for the digital age: Towards a truly European digital society*[1], sets bold objectives for the future of the Digital Single Market: by 2025, Europe should have “tripled the share of companies using AI”, “increased the number of companies using big data”, ensured a “Gigabit internet connectivity with equally fast upload and download speeds for all main socio-economic drivers, such as schools, hospitals, businesses”. By the same date there should also be “accessible electronic health records for all Europeans”. The overall goal is to “achieve an internet where citizens are in control of their data, and their online identity. But always with the clear understanding of what is given away and how their data is protected”.

The guiding principle is that “Many of the opportunities of digitalisation are still ahead and are complex to unlock. A greater willingness to share data will help address important social challenges. Europe must seize the vast potential of the exponentially growing amounts of data, particularly in areas where it is strong […] and to maintain leadership in key sectors, such as health […]. Secure availability of health data combined with responsible research and innovation, leads to better treatment for major chronic conditions that drain over 70% of health system resources, including cancer […] and rare diseases”.

In the Work Programme 2020, *A Union that strives for more*[2], the Commission also announced a Pharmaceutical Strategy for Europe “to ensure the quality and safety of medicines and to

[1] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, *Europe fit for the digital age: Towards a truly European digital society*, Brussels, Draft February 2020.

[2] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Commission Work Programme 2020, *A Union that strives for more*, Brussels, 29 January 2020.

DRAFT

consolidate the sector's global competitiveness. Europe should also make sure that all patients can benefit from innovation while resisting the pressure of increasing costs of medicines"

2. The inconvenient truth

Notwithstanding these statements, the pressing demand for data driven by artificial intelligence and big data, and the number of national and European initiatives aiming to facilitate data sharing in the biomedical sector, the friction between technology solutions and regulatory constraints keeps mounting and the ambitious goals set by the new Commission will remain at risk.

Data sharing in healthcare is still rare, it imposes high transaction costs and happens mostly under private agreements typically enacted by large corporations. Indeed, "although available data is continuously expanding, it largely sits idle"[3], i.e. fragmented in siloes[4] carefully guarded by data controllers to reduce legal exposure. The advent of truly open and competitive big data and AI environments, while rapidly expanding in other continental blocks, is delayed in Europe, with not only economic consequences, but personal and societal ones too, as in the case of rare diseases which altogether affect 30 million people in Europe, for which research data are still dramatically scarce.

There is therefore the need to critically reassess these obstacles and set a strategic focus on technologies that can practically allow Europe to effectively scale up a compliant data ecosystem for the biomedical sciences.

3. GDPR and the health data sharing challenge

With the implementation of the GDPR less than two years ago, the goal of placing individuals at the center of data ecosystems is, on the other hand, actually coming into sight. These new

[3] M. Finck, Blockchain Regulation and Governance in Europe, Cambridge University Press, 2019, p. 120.

[4] Even in the U.S.: the American Medical Informatics Association (AMIA) declares to be in support of an expansive update to the NIH's data sharing policy, remarking that a key deficiency until now has been that "grant applications are not scored on the quality of their data sharing plans" and this has "led to suboptimal and incomplete sharing plans and has likely contributed to data silos" (J. Kent, AMIA Urges NIH to Revise Proposed Data Management, Sharing Policy, Health IT Analytics, January 16, 2020).

DRAFT

standards are indeed changing the tone of business relationships within and outside Europe and setting the bar for other countries, which are debating how to craft their own legislations in response to repeated, systemic failures. They have also dramatically increased transparency in data management processes well beyond the GDPR's initial expectations, as shown by the number of data breached that are now reported since the regulation came into force, though not without difficulties in enforcing corresponding sanctions[5]. As corporations transform their infrastructures, initial estimates are also starting to emerge on the cost, especially for small businesses, of data protection requirements[6]. While these expenditures were expected, the rise of Artificial Intelligence (AI) as a disruptive innovation force in the global economy is creating unprecedented demand for heterogenous, integrated and big data sets, further increasing the friction between individuals' privacy and the information needs of artificial cognitive systems[7], relying more and more on massive correlations, rather than on specific causation hypotheses[8].

Machine learning and AI have demonstrated their ability to aid clinical decisions, increase patient safety and reduce healthcare costs. The large volumes of biomedical data needed by these systems are now available thanks to Electronic Medical Records (EMR), wearable devices and new imaging systems. A scalable system enabling private and secure patient data processing becomes therefore the missing link for boosting data-driven innovation in the medical domain. Such a system must be able to handle distributed patient consent, peer-to-peer data exchanges, strong identity and credentials management and data distribution under formal privacy guarantees.

The GDPR mandates the use of privacy protection measures, such as anonymisation and pseudoanonymisation, in an intentionally broad way to allow local innovation and yet, as of

[5] Vinocur N., 'We have a huge problem': European regulator despairs over lack of enforcement-The world's toughest privacy law proves toothless in the eyes of many critics, Politico, 27 December 2019.

[6] https://www.bitkom.org/EN/List-and-detailpages/Press/Annual-Survey-Bitkom-draws-mixed-conclusion-regarding-GDPR-implementation

[7] https://www.medtecheurope.org/wp-content/uploads/2019/11/MTE_Nov19_AI-in-MedTech-Delivering-on-the-Promise-of-Better-Healthcare-in-Europe.pdf.

[8] V. Mayer-Schonberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, 2014.

DRAFT

today, no real consensus or reference framework exists for securely exchanging medical data at scale. Data transactions are still governed by ad-hoc and private agreements.

In particular, anonymisation is defined in the GDPR as the process by which health data are irreversibly altered in such a way that a data subject can no longer be identified (by “all the means reasonably likely to be used”), directly or indirectly, neither by the data controller alone, nor in collaboration with any other party, and by any malicious third party. Such non-re-identifiability can, however, reduce information in the data to the point of making them unusable for scientific discovery or realistic AI-systems training. Being a subtractive technique, based on stripping away direct and indirect identifiers, anonymisation has the flaw of “incurring not only poor privacy results, but also lackluster utility”[9,10,11]. Anonymisation based on k-anonymity, i.e. the methodology, originally devised by Latanya Sweeney and Pierangela Samarati at the Harvard Privacy Lab[12], by which the information for each person contained in the release cannot be distinguished from at least k - 1 individuals whose information also appear in the release, has shown for instance to work on constrained data uses, while losing reliability as the number of aggregated records increases[13].

At the same time, the issue is further complicated by the emergence of AI-based tools for re-identification[14] which push even further the amount of information that needs to be removed from a given data set to make it actually anonymous.

[9] S. M. Bellovin, P. K. Dutta, and N. Reitinger, Privacy and Synthetic Datasets, Stanford Technology Law Review, Vol. 22:1, Winter 2019.

[10] P. Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA Law Revue, 1703 (2010).

[11] A. Narayanan and E. W. Felton, No Silver Bullet: De-Identification Still Doesn’t Work (July 9, 2014).

[12] P. Samarati, L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Harvard Data Privacy Lab, 1998; L. Sweeney, K-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10, 2002.

[13] C. C. Aggarwal, On k-Anonymity and the Curse of Dimensionality, Proceedings of the 31st International Conference on Very large Data Bases, Trondheim, 2005.

[14] Richard McPherson, Reza Shokri, Vitaly Shmatikov, Defeating Image Obfuscation with Deep Learning, September 2016, https://www.researchgate.net/publication/30760422.

DRAFT

Pseudoanonymisation, on the other hand, relates to the processing of health data in ways by which they can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that data are not attributed to an identified or identifiable natural person. As re-identifiable, even encrypted data are pseudonymous. Given their re-identifiability, all pseudonymous data require on principle an explicit personal consent for being shared with third parties. Pseudonymisation requires therefore formal relationships among actors under which data can be de-crypted.

Because of this, both anonymous and pseudonymous health data end up being either inherently inadequate or extremely hard to scale up to an aggregation level of bigdata sets that can allow an efficient development of robust AI solutions.

Neither of these approaches are technically and economically sustainable to boost data-driven R&D at industrial scales.

Compliance risks act also as a constraint in the development of a thriving data economy driven by direct incentives. It is evident that supply/demand dynamics will be key in the growth of the European Digital Single Market as they are doing in other, more permissive, jurisdictions.

In the ideal scenario, data controllers and individuals should be able to publish the largest possible data sets to the broadest possible community, while effectively protecting the identities of individual subjects and, in turn, be able to re-capture value from data transactions, fostering network effects to progressively increase the total volume of available data.

4. National public sector initiatives

France

The French Health Data Hub (HDH)[15] and the Espace Numérique de Santé (ENS) both aim to “empower citizens as actors of their healthcare throughout their life”, by creating a digital and personal health space from birth which allows citizens to manage health data for different

[15] https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/acces-aux-donnees-de-sante/article/health-data-hub.

personalized services. The ambition is to include all national electronic health records and e-prescriptions, child health records and more in a dedicated “store” consented by users.

The plan includes a revision of the law as only public interest studies are allowed currently on health data, without patient consent, but upon the agreement of a national committee (CEREES[16]) and with CNIL authorization. The perspective of crossing multiple data sources in a truly open approach will increase the risk of identification and the Hub therefore plans to drastically increase transparency on the portal and to collect fully informed consent.

Germany

A less ambitious case is the German Medical Informatics Initiative (MII)[17], which focuses on university hospitals and medical centers caring for 1.8 million patients annually (10% of all inpatient cases in the country) on which the German Federal Ministry of Education and Research (BMBF) is investing around 160 million euros through 2021.

Whereas the French HDH project is based on a top-down approach and a shared computational infrastructure, the German MII project is based on a bottom-up approach relying on four consortia which include teaching hospitals, universities, and private partners tasked with developing strategies for shared data use and exchange. They will subsequently establish the data integration centres and create IT solutions for concrete use cases.

A National Steering Committee (NSG) ensures the interoperability of consent practices. The overarching goal is to capture hospital data and make these available for research by making it possible also for individual citizens to volunteer their anonymised data. In this case too, to address concerns over privacy and personal data protection, the government will introduce a special law on the protection of health data in electronic patient records.

DRAFT

[16] Comité d’expertise pour les recherches, les études et les évaluations dans le domaine de la santé, appointed in 2017.

[17] https://www.medizininformatik-initiative.de/en/about-initiative; M. Cuggia and S. Combe, The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare, Yearbook of Medical Informatics, 2019; 28(1): 195–202.

United Kingdom

The UK government, building on the "NIH Big Data to Knowledge (BD2K) initiative"[18], has launched a national strategy, to support AI development and health data sharing, captured in the 2018 "The future of healthcare: our vision for digital, data and technology in health and care"[19]. The document includes a code of conduct for data-driven healthcare technology, data collections, and a communication campaign, "Your Data Matters to the NHS"[20], in conjunction with a national data opt-out scheme in line with the recommendations of the National Data Guardian in her Review of Data Security, Consent and Opt-Outs[21].

The Scottish National Safe Haven has been operating since 2015. The computing centre of the University of Edinburgh (EPCC) has been tasked to build, maintain and operate this Scottish Safe Haven in partnership with the National Health Service and the Scottish Government, providing a secure setting for research on non-consented, de-identified datasets of clinical data and images, government data and other public sector data. EPCC has adopted the "5 safes" approach to data sharing (safe data, safe settings, safe people, safe projects, safe outputs). Accessing the Safe Haven requires approvals by care givers and the Public Benefit and Privacy Panel for Health and Social Care (PBPP). A response to a query normally takes a few months.[22]

The Netherlands

In the Netherlands, the Translational Research IT (TraIT) Infrastructure was launched in 2011 to manage a nationwide infrastructure for data and workflow management across four major domains of translational research: clinical, imaging, biobanking, and experimental (any-omics). Researchers from multi-site projects can share and disseminate data and analyses from these domains making TraIT an internationally recognized FAIR (i.e. findable, accessible,

DRAFT

[18] Bourne P E, Bonazzi V, Dunn M, Green E D, Guyer M, Komatsoulis G et al., The NIH Big Data to Knowledge (BD2K) initiative, Journal of the American Medical Informatics Association, 2015;22(06).

[19] https://www.gov.uk/government/publications/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care.

[20] https://www.sfh-tr.nhs.uk/media/1984/your-data-matters-patient-handout-version-1.pdf.

[21] https://www.gov.uk/government/publications/review-of-data-security-consent-and-opt-outs.

[22] Rob Baxter, The Scottish National Safe Haven, Update Jan 2020, presentation at the workshop "Towards pseudo/anonymised health data", EOSC Life, Paris, 22-23 January 2020.

interoperable and reusable) best practice example and data stewardship model, becoming a key components of the Dutch national biobanking infrastructure (BBMRI-NL)[23], and one of the initiators of the comprehensive Dutch national infrastructure for personalized medicine & health research, Health-RI.

Denmark

Denmark is probably the best example of end-to-end infrastructure and related policies to manage healthcare data, making the Danish population effectively a clinical study cohort at a national level. MedCom, first established already in 1994, ensures that specialised IT systems can exchange the health-related information in a secure environment when required by a patient's course of treatment. In 2016 the Ministry of Health initiated the project "Patient Reported Outcome (PRO) in General Practice" to expand the use of an electronic PRO system in general practice.

Estonia

Similar to the Danish experience is Estonia, in which the Central Health Information System EHR collects patient summaries on every clinical encounter. Data are visible to all clinicians and patients. Doctors` access to the central database happens only via personal ID-cards. All visits summaries are linked to the Medical Images Bank, the Prescription Centre and health care providers systems via the national secure internet-based data exchange layer.

Finland

Finland focused since 2011 on Knowledge management for more efficient and effective public services. This has led to the establishment of the one-stop shop system Findata, facilitated also by a new GDPR-friendly legislation. The Act on the Secondary Use of Health and Social Data, in effect since May 2019, aims at facilitating the secondary use of social welfare and healthcare data from multiple sources as well as to promote its secure use for broader purposes. Furthermore, through the nationwide service My Kanta Pages, citizens can see their health records and prescriptions, request a prescription renewal, and save their living will and organ

[23] Merino-Martinez R, Norlin L, van Enckevort D, Anton G, Schuffenhauer S, Silander K, Mook L, Holub P, Bild R, Swertz M, Litton JE. Toward Global Biobank Integration by Implementation of the Minimum Information About Biobank Data Sharing (MIABIS 2.0 Core). Biopreservation and Biobanking 2016, 14(4): 298-306.

DRAFT

donation testament. Under consent, different healthcare units can see the citizen's health records. The right to be forgotten is implemented.

Italy – Lombardy Region

Over the last two decades, the Lombardy Region has steadily grown a population database in which detailed information on the health services for its 10 million inhabitants is collected, under the guidance of Aria SpA, the regional legal entity in charge of this initiative. The database captures services provided both by hospital-based and community-based care delivery systems, including pharmacies, labs, primary care clinics and nursing homes, as well as resource consumption such as personnel, medical devices and pharmaceuticals and their costs, all in compliance with the GDPR.

Aria has set to strategically leverage this asset bringing it to fruition not only for the Region central administration and its various organisational units but also for research centres and Universities, which can generate added value, for the Region as well, by leveraging these data for innovation. To this end, Aria is implementing the Digital Information Hub, as both an organisational and technological model of operations, aimed at facilitating the governance and the centralised usage of the database by internal and external accredited subjects, balancing security with ease of access and fruition.

DRAFT

5. Legal and ethical challenges in public-private partnerships

As discussed above, in this challenging context, most commercial data exchanges happen under private or public-private agreements typically enacted by large corporations. The UK, for instance, launched a much debated project in 2016, at Moorfields Eye Hospital NHS Foundation Trust, in partnership with DeepMind, substituted, after acquisition in 2018, by Google Health. The outcome has been a non-commercial public asset[24], [25] consisting of a deep learning

[24] Powles J. and Hodson H., Google DeepMind and healthcare in an age of algorithms, Health and Technology, December 2017, Volume 7, Issue 4, 351–367.

[25] Lomas N., Google completes controversial takeover of DeepMind Health, Techcrunch, September 19, 2019.

architecture for clinically heterogeneous sets of three-dimensional optical coherence tomography scans[26]owned by Moorfields.

A parallel agreement was entered by DeepMind with Royal Free London NHS Foundation Trust, for accessing identifiable patient records, without explicit consent, for developing a clinical app for kidney injury. The app reduced the average cost of admission by 17%, demonstrating huge potential[27], while criticism on privacy protection has been escalating. The Royal Free NHS Trust, Taunton & Somerset NHS Foundation Trust, Imperial College Healthcare NHS Trust, and University College London Hospitals NHS Foundation Trust are all under contract with Google.

An important case is Sensyne Health plc, a clinical artificial intelligence technology company that raised £60 million via its IPO in 2018. The NHS Trusts allowed Sensyne to receive a part of the financial returns via equity ownership and royalties from Oxford University Hospitals NHS Foundation Trust, Chelsea & Westminster Hospital NHS Foundation Trust, and South Warwickshire NHS Foundation Trust, acting as a bridge between the NHS and pharmaceutical companies to commercialise data for accelerating drug discovery and improving patient care on a large databases of anonymised data. These data are ethically sourced, in that any analysis of (and hence the Company's access to it) must be pre-approved on a case-by-case basis by the NHS Trusts.

In July 2019 Sensyne announced that it signed an initial two-year collaboration agreement with Bayer to accelerate the clinical development of new treatments for cardiovascular disease using Sensyne Health’s proprietary clinical AI technology platform.

DRAFT

[26] Fauw, J., Ledsam, J.R., Romera-Paredes, B. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24, 1342–1350 (2018).

[27] Tomašev, N., Glorot, X., Rae, J.W. et al. (2019), A clinically applicable approach to continuous prediction of future acute kidney injury, Nature 572, 116–119.

DRAFT

6. European Health Data Space initiative, previous projects, and existing research infrastructures

Somewhat preceding the creation of the Big Data to Knowledge (BD2K) project launched in the U.S. by the National Institutes of Health[28], in Europe the Open PHACTS (Open Pharmacological Concept Triple Store) was funded in 2011 by the Innovative Medicines Initiative (IMI), aiming at creating a public–private partnership between academia, publishers, enterprises, pharmaceutical companies and other organisations, working to enable cheaper and faster drug discoveries.

This was followed by the European Medical Information Framework (EMIF) project, launched in January 2013 for a duration of five years, with the task of developing common technical and governance solutions and improve access and use of health data.

Other major European projects were:

- epSOS (2008-2014), a large-scale pilot testing the cross-border sharing of a summary of a patient's essential health data in case of unplanned care and the electronic prescription;
- eHealth Network, established in 2012, which launched a Joint Action (JA) in 2015 for preparing various themes for policy decisions, and is still ongoing.
- STORK 2.0 (2012-2015), fostering citizens' and business mobility through cross-border authentication and identification (eID), involving 55 organizations, both public and private, across 19 European countries;
- DECIPHER (2012-2017), with the goal of securing cross-border mobile access to patient healthcare portals supported by national bodies, in order to have patients able to use a secure mobile device safely to gain 24/7 access to their prescription data, emergency data, examination results and other health information;
- EXPAND (2014-2015), securing the epSOS pilot services up to the launch of the Connecting Europe Facility (CEF);
- Antilope (2013-2015), focussing on the dissemination and adoption of the eHealth European Interoperability Framework (eEIF);

[28] P. Bourne, V. Bonazzi, M.C. Dunn, B. Russell, The NIH big data to knowledge (BD2K) initiative, Journal of the American Medical Informatics Association, November 2015 22(6):1114-1114.

DRAFT

- AssessCT (2015-2016), analysing whether to introduce as a single medical language, SNOMED-CT, for Large Scale eHealth deployment in the EU;
- ValueHealth (2015-2017), tasked with demonstrating how interoperability of health information could justify sustainable investments across Europe;
- eStandards (2016-2017) was a coordination and support action to strengthen interoperability by alignment and wide adoption of standards in the eHealth market for products and services.
- OpenMedicine (2016) had the aim of meeting the challenge of unique identification of medicinal products enhancing the safety and continuity of cross-border (and also national level) healthcare through interoperable ePrescriptions.
- OpenNCP platform (ongoing), supported by 14 member states, by which an EU citizen can ask for medical treatment in another EU country retrieving his/her medical history in compliance with national health information systems. Aiming at integrating the European eIDAS infrastructure in the OpenNCP platform and proposing the HeID solution as an answer for the missing patient authentication in the current OpenNCP implementation;
- KONFIDO (2016-2019), consisting of 15 partners from 7 different countries, among which there were also the Danish MedCom and Sundhed. Konfido aimed at creating a scalable and holistic paradigm for secure inner- and cross-border exchange, storage and overall handling of healthcare data both at national and European levels, using privacy by design principles.

Crowning all these efforts, a European Institute for Innovation through Health Data[29] was established in 2016, arising from projects like Electronic Health Records for Clinical Research and SemanticHealthNet, with the mission of scaling up innovations that rely on high-quality and interoperable health data.

In the background, the European Research Area (ERA) has been the overall system of scientific research programmes tasked with integrating the Union's scientific resources for biomedical research, and relying on a series of European infrastructures, such as:

[29] Kalra D, Stroetmann V., Sundgren M., Dupont D., Schlünder I, Thienpont G., Coorevits P., De Moor G., The European Institute for Innovation through Health Data, Learn Health Syst. 2017 Jan; 1(1):e10008.

DRAFT

- EMBL-EBI (joint European Molecular Biology Laboratory and European Bioinformatics Institute), which has the ambition of being "the home for big data in biology" by maintaining a comprehensive range of up-to-date molecular data resources.
- ELIXIR (Distributed Infrastructure for life-science information), of which EMBL-EBI is a Node, supporting the coordination of biological data provision through ELIXIR Core Data Resources, and committed to Open Access as a core principle for publicly funded research.
- BBMRI-ERIC (Biobanking and BioMolecular resources Research Infrastructure – European Research Infrastructure Consortium)
- EATRIS (European Research Infrastructure for Translational Medicine)
- ECRIN (European Clinical Research Infrastructure Network)
- EGA (European Genome-phenome Archive), which archives a large number of datasets, the access to which is controlled by a Data Access Committee (DAC)
- HNN (Health NCP Net), which is the support network to navigate through the European Health Research & Innovation funding landscape.

More recently, in 2018, the European Commission launched an initiative on the Digital Transformation of Health and Care (Digicare) aimed at fostering secure, flexible and decentralised digital health infrastructures. The creation of a European Health Research and Innovation Cloud (HRIC) within this environment was indicated as the pre-condition "for enabling data sharing and analysis for health research across the EU in compliance with data protection legislation while preserving the full trust of the participants. Such a HRIC should learn from and build on existing data infrastructures, integrate best practices, and focus on the concrete needs of the community in terms of technologies, governance, management, regulation and ethics requirements"[30].

In this context, it was acknowledged that the sharing of data, information and knowledge "represents the most important functionality in the context of a HRIC"[31]. Though – it was said – "the sharing of highly sensitive clinical and health data has thus far been much less developed

[30] C. Auffray et al., Towards a European Health Research and Innovation Cloud (HRIC), Genome Medicine, 2020 (https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-0713-z).

.

[31] Ibid.

hence representing a key future focus area, and numerous challenges remain with regard to sharing these data in a meaningful way”[32].

In 2019, with the coming into force of the new European Commission, led by Ursula von der Leyen, more ambitious goals have been indicated for the health sector, such as the European Health Data Space, the Electronic Health Record Exchange Format and a common semantic strategy.

7. Alternative modes for dealing with the issue of sharing health data

The number of initiatives mentioned above stands as testimony of the crucial and still unsolved challenge around sharing of big health data, and of the challenge of applying AI on these resources in a lawful, GDPR compliant way.

Our research led us and others to believe that there are actually two ways for solving this challenge. One is the so-called “visiting” mode, in which data are not physically accessed by third parties, but rather “algorithms are brought to the data” and only the outcomes of secure computations are released.

The “visiting” mode operates through mechanisms like homomorphic encryption[33], secure multi-party computation, and federated learning[34] with untrusted blackbox, as shown within the H2020 EU-funded *MyHealthMyData* project (2016-2019), which experimented with this solution in conjunction with a permissioned blockchain system for recording transactions, an off-chain storage of health data, a metadata catalogue to view and request available data

DRAFT

[32] Ibid.

[33] A.Vizitiu, C. I. Nita, A. Puiu, C. Suciu, L. M. Itu, Towards Privacy-Preserving Deep Learning based Medical Imaging Applications, IEEE 2019; A.Vizitiu, C. I. Nita, A. Puiu, C. Suciu, L. M. Itu, Privacy-Preserving Artificial Intelligence: Application to Precision Medicine, IEEE 2019. This homomorphic encryption solution, developed by the Transilvania University of Brasov (Romania) within the MyHealthMyData project, was awarded the E.U. Innovation Radar Prize 2019 in the category Industrial & Enabling Tech.

[34] K. Bonawitz et al, Towards Federated Learning at Scale: System Design, SysML 2019.

DRAFT

assets, and smart contracts for automatically handling individual consent and institutional permissions.

The other way, which deserves particular attention, is the generation of synthetic data.

8. Synthetic data

The use of synthetic data is attracting growing attention as a practical solution to the quandary of maintaining privacy in big data ecosystems. Not surprisingly, the UK Government Statistical Service has defined it "an unprecedented opportunity to innovate with data, while safeguarding privacy and fostering public trust"[35]; the Stanford Technology Law Review states that "synthetic data is a viable, next-step solution to the database-privacy problem"[36]; a recent comment in the Financial Times says that "improvements in machine learning and computing power make it a technology to watch"[37].

Data synthesis allows "to step away from the deidentification–reidentification arms race and focus on what really matters: useful data […] combined with differential privacy to achieve a best-of-both-worlds scenario".[38]

Synthetic data are created from real data, by machine-learning generative model to "produce realistic, yet artificial data that nevertheless has the same statistical properties […]to create an as-realistic-as-possible dataset, one that not only maintains the nuances of the original data, but does so without endangering important pieces of personal information"[39].

The difference between traditionally anonymised data and synthetic datasets is that "these datasets protect privacy through the addition of statistically similar information, rather than

[35] Government Statistical Service, Privacy and data confidentiality methods: a National Statistician's Quality Review (NSQR), 13 December 2018.

[36] S. M. Bellovin, P. K. Dutta, and N. Reitinger, Privacy and Synthetic Datasets, cit.

[37] A. Ahuja, The promise of synthetic data, The Financial Times, 4 February 2020.

[38] S. M. Bellovin, P. K. Dutta, and N. Reitinger, Privacy and Synthetic Datasets, cit.

[39] Ibid.

through the stripping away of unique identifiers[40]". The authors of *Privacy and Synthetic Datasets* have recourse to a possibly clarifying analogy: "synthetic data is like replacing the pieces of a jigsaw puzzle to create a different picture; even though all the puzzle pieces fit together in the same way (i.e., each piece has similar, yet synthetic, attributes), the overall image has changed—importantly, and hopefully, the change is not discernible but nonetheless protects privacy"[41].

Statistical characteristics of the real population are "learned" during synthetization, from the original data, while the synthesis process uncouples sensitive information from the data information content, attaining anonymity though still preserving information richness. This directly responds to the GDPR specification of "personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable"[42], while the overarching objective is to create "high-quality synthetic data that closely resemble the real data and are a suitable substitute for processing and analysis"[43].

Recent literature has highlighted differential privacy as an additional privacy protecting measure to further enhance synthetic data[44]. Information leakage from each query can, in fact, be minimal on synthetic data, but is, by definition, never zero. Overtime, with each database query, the amount of leaked information grows.

The U.S. National Institute of Standards and Technology has launched a Differential Privacy Synthetic Data Challenge in 2019, showing keen focus on developing "a mathematical theory,

DRAFT

[40] Ibid.

[41] Ibid.

[42] GDPR, Recital 26.

[43] ONS, Creating Synthetic Data, in: Our First Two Years: Data Science for Public Good, "Science Campus", March 27, 2019.

[44] H. Page, C. Cabot, K. Nissim, Differential privacy: an introduction for statistical agencies, National Statistician's Quality Review, December 2018.

and set of computational techniques, that provide a method of de-identifying data sets—under the restriction of a quantifiable level of privacy loss”[45].

DRAFT

9. A new anonymisation paradigm

Synthetic data were introduced in 1993[46,47] as a way to replicate the statistical properties of a database without exposing identifiable information, acting therefore as a statistical disclosure control (SDC) method or as an alternative to SDC[48].

Methods to produce them vary[49,50,51,52,53], but the underlying principle is that values in the original data are algorithmically substituted with others taken from statistically equivalent distributions, to create entirely new records, with as little traceable relation to the originals as possible.

[45] NIST, Differential Privacy Synthetic Data Challenge: Propose an algorithm to develop differentially private synthetic datasets to enable the protection of personally identifiable information (PII) while maintaining a dataset's utility for analysis, Submission Start: October 31st 2018 - Submission End: May 20th 2019.

[46] D. Rubin, Statistical disclosure limitation, Journal of official Statistics 9.2 (1993): 461-468

[47] R.J.A. Little, Statistical Analysis of Masked Data, Journal of Official Statistics 9, 407-426, 1993.

[48] M. Elliot and J. Domingo-Ferrer, The future of statistical disclosure control, National Statistician’s Quality Review, December 2018.

[49] J. Reiter, et’ al., The multiple adaptations of multiple imputation, Journal of the American Statistical Association 102.480 (2007).

[50] J. Drechsler, et al., An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, Computational Statistics & Data Analysis, 55.12 (2011): 3232-3243.

[51] H Surendra and H. S. Mohan, A Review of Synthetic Data Generation Methods for Privacy Preserving Data Publishing, International Journal of Scientific & Technology Research 6, 3 (March 2017), 95–101.

[52] J. Hu, et al.. "Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data." Bayesian Analysis 13.1 (2018): 183-200.

[53] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney. 2018. Privacy preserving synthetic data release using deep learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 510–526.

DRAFT

In *MyHealthMyData* they have been successfully used[54] to publish clinical data and MRI cardiovascular images, to train machine learning tools, and to validate clinical decision support applications. A growing body of international research is showing evidence that they can yield equivalent analytical results as original data sets[55,56,57]. While their adoption is suffering from lack of standards and best practices, their use in specialized sectors of biomedical research, (i.e.in-silico clinical trials) is growing, especially under the auspices of the US FDA.

The value of synthetic data resides in a series of key characteristics:

1. They are used in the same way original data sets are and therefore with the same storage, maintenance and analytical infrastructures;
2. They maintain statistical characteristics of the original data, but can be expanded for imputating (replacing missing values with substitutes) and augmenting real data. They can thus fill gaps, correct skewed value distributions, or remove spurious values in the original data, addressing collection, formatting or normalization issues, which are pervasive in clinical data sets, and thus producing data that are actually more informative and realistic than the original ones.[58]
3. They can be produced at low costs, for a variety of uses and in very large volumes to jump-start AI-development in areas where data are scarce or too expensive, such as the biomedical sector which, as discussed above, suffers in this regard from both economic and legal limitations.

[54] Thanks initially to the ground-breaking contribution of one of MHMD clinical partners, namely the Queen Mary University of London team, composed by S. Petersen, A. Lee, and M. Jennings (acknowledging also support from the “SmartHeart” EPSRC programme grant - www.nihr.ac.uk EP/P001009/1).

[55] Patki N., Wedge R., Veeramachaneni K., The Synthetic Data Vault, 2016 IEEE 3rd International Conference on Data Science and Advanced Analytics, Volume: 1, 399-410: “scientists can be as productive with synthesized data as they can with control data”.

[56] Aviñó, L. et al. "Generating Synthetic but Plausible Healthcare Record Datasets." arXiv preprint arXiv:1807.01514 (2018).

[57] http://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303.

[58] B. Nowok, G. M. Raab, C. Dibben, Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the *synthpop* package for R, Statistical Journal of the IAOS, pp. 1-12, 2017.

DRAFT

4. The actual risk or re-identification can be effectively quantified in relation to the original ones with differential privacy and modulated, in the generative process, based on the intended use and distribution.

10. The Technology

Synthetic data can be generated by a diverse range of systems[59,60,61,62,63,64], including naive Bayes models[65] or statistical shape analysis[66,67] (for imaging data), according to user/customer requirements and intended dimensionality in the resulting set.

[59] Y. Park and J. Ghosh, PeGS: Perturbed Gibbs Samplers that Generate Privacy-Compliant Synthetic Data, Transactions on Data Privacy 7(3), pp. 253-282, 2014.

[60] H. Ping, J. Stoyanovich, and B. Howe, DataSynthesizer: Privacy-Preserving Synthetic Datasets, Proceedings of the 29th International Conference on Scientific and Statistical Database Management (SSDBM), ACM Press, Chicago, 2017.

[61] S. McLachlan, Realism in Synthetic Data Generation, Computer Science and Information Technology, Massey University, New Zealand, 2016.

[62] Syahaneim, R. A. Hazwani, N. Wahida, S. I. Shafikah, Zuraini, and P. N. Ellyza, “Automatic Artificial Data Generator: Framework and implementation,” in 2016 International Conference on Information and Communication Technology (ICICTM), 2016, pp. 56–60.

[63] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230–8.

[64] Chen J., Chun D., Patel M., Chiang E. and James J., The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures, BMC Medical Informatics and Decision Making (2019).

[65] P. Multani, U. Niemann, M. Cypko, J.-P. Kühn, H. Völzke, S. Oeltze-Jafra, M. Spiliopoulou, “Building a Bayesian Network to Understand the Interplay of Variables in an Epidemiological Population-Based Study,” in 'Proceedings of the 31th IEEE Int. Symposium on Computer-Based Medical Systems (CBMS18)' , pp. 88-93, 2018 .

[66] J.L. Bruse, et al., Detecting Clinically Meaningful Shape Clusters in Medical Image Data: Metrics Analysis for Hierarchical Clustering applied to Healthy and Pathological Aortic Arches, IEEE Transactions on Biomedical Engineering, 2017.

[67] B. Biffi, et al., Investigating Cardiac Motion Patterns using Synthetic High Resolution 3D Cardiovascular Magnetic Resonance images and Statistical Shape Analysis, Frontiers in Pediatrics, 2017;5:34.

DRAFT

In particular Generative Adversarial Networks (GANs)[68], brought renewed attention to this application, having achieved remarkable results in generating data “without the need for vast troves of painstakingly-labeled training data […], [being hailed] as one of the most important innovations in deep learning”[69].

A Generative Adversarial Network uses two models playing against each other. The Generator learns to capture and recreate the data distribution while the Discriminator estimates the probability that a generated sample belongs to the original data distribution or rather has been created by the Generator; in other words it decides whether the data is fake or not.

GANs can discover structures in the data well beyond what other techniques can do, but other unsupervised statistical techniques can be utilized, such as Monte Carlo simulations, which have been shown to reduce leakage (the accidental insertion of a priori knowledge in the synthetic set) and to perform well in reproducing statistical multi-dimensionality, especially when compared to supervised methods.

11. Interpretability

Statistical realism is of course key to valid inferences on synthetic data and discriminators are not surprisingly a crucial area of research and innovation.

Commonly used systems in this space, based on Random Forest or MMD statistics, do assess the overall statistical resemblance of two sets, but in case of discrepancies they cannot identify the underlying reasons.

While these types of discriminators remain useful as a first line of evaluation, new methods now allow to weight each original variable in the generation process, thus supporting detailed diagnostics and direct, ongoing improvements to the generative pipeline. These new approaches to algorithmic transparency, i.e. the ability for a human operator to trace the weight of original variables in the synthetic set, reduce the risk of “mode collapse”, the tendency to learn from and thus replicate only the most prevalent features in the original data.

[68] I. J. Goodfellow et al., Generative Adversarial Nets, Proceedings of Neural Information Processing Systems, 27, 2014.

[69] C. Bowles et al., GAN Augmentation: Augmenting Training Data Using Generative Adversarial Networks (Jan. 8, 2019), https://perma.cc/K9SH-73L2.

DRAFT

Such systems now allow to continuously optimize the data creation process and systematically address biases.

Among other approaches, these tools leverage iterative L1-regularized parametric model using the interpretable components as inputs[70], which identify the relative weight of each original variable in the generation of the synthetic replica and therefore allow targeted adjustment of the generative process. This direct feedback loop design has shown to drastically improve efficiency of and control over the generation process[71,72,73].

12. 3D and 4D Synthetic images

New methodologies for artificial surfaces models have been developed to support the creation of 3D and 4D images, i.e., including haemodynamic data. Vascular structures or solid organs are first initialized from original radiology images and corresponding geometric information is then reproduced to replicate vessel radius, degree of tapering, branch length, among other things. Finally, pathological aspects are modeled. Synthetic images have an additional and substantial advantage in medical-AI applications: time consuming and costly manual annotations to train algorithms on clinically relevant features, can be inserted automatically in the synthetization process by the generative pipeline reducing the overall cost of AI-systems development[74].

[70] Van Belle V., Lisboa P.. White box radial basis function classifiers with component selection for clinical prediction models, Artificial Intelligence Medicine, 2014 Jan;60(1):53-64.

[71] E. Choi, S. Biswal, B. Malin, J. Duke, W. F Stewart, and J. Sun, Generating multi-label discrete electronic health records using generative adversarial networks, arXiv preprint arXiv:1703.06490, 2017.

[72] J. Jordon, J. Yoon, and M. van der Schaar, Measuring the quality of synthetic data for use in competitions, arXiv preprint arXiv:1806.11345, 2018.

[73] J. Hui, GAN — Why it is so hard to train Generative Adversarial Networks!, https://medium.com/@jonathan_hui/gan-why-it-is-so-hard-to-train-generative-advisory-networks-819a86b3750b

[74] J. T. Guibas et al., Synthetic Medical Images from Dual Generative Adversarial Networks, Proceedings of Neural Information Processing Systems, 31, 2017.

DRAFT

13. Synthetic Genetic data (to be expanded)

These approaches can be applied to replicate synthetic distributions of gene expression too, including single nucleotide polymorphism, copy number variation or protein-protein/gene-gene interactions. Deep learning approaches (especially GAN and InfoGAN) have demonstrated both scalability and statistically robustness in this area while ongoing work, targeting the oncology drug development sector, showing promising results in the generation of both SNP-chip like genotypes (containing only common genetic variants), and sequencing data[75, 76].

14. Synthetic data and Differential Privacy

As defined in the seminal work of Cynthia Dwork and Aaron Roth in 2014[77], differential privacy provides a mathematical foundation to substantiate privacy assessment and its legal definition.

The idea of adding differential privacy to neural networks drew interest as early as 2016[78]; however, it was not until 2017-18 that researchers realized the potential implications and advantages of applying the technique to GANs[79,80], by implementing noise, i.e., applying filters and weights, into the training data.

[75] C. A. Azencott, Machine learning and genomics: precision medicine vs. patient privacy, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, The Royal Society, 2018, 376 (2128).

[76] Tom Ellis, What is synthetic genomics anyway?, June 2019, https://www.researchgate.net/publication/336417522_What_is_synthetic_genomics_anyway.

[77] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy, Foundations and Trends in Theoretical Computer Science, 9(3–4), 2014.

[78] M. Abadi et al., Deep Learning with Differential Privacy, Proceedings of the Conference on Computer and Communications Security, 2016.

[79] L. Xie et al., POSTER: A Unified Framework of Differentially Private Synthetic Data Release with Generative Adversarial Network, Conference on Computer and Communications Security, 2017.

[80] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, Differentially private generative adversarial network, arXiv preprint arXiv:1802.06739, 2018.

A recent literature has articulated the key role of differential privacy in combination with synthetic data[81,82, 83], as being the only solution to sufficiently protect privacy while maintaining utility. Differential privacy's and its robust guarantees do not only mitigate the risk information leakage, but also the risks of adversarial machine learning.

Although the technique is relatively new (and the optimal means of applying differential privacy to synthetic data is not yet settled), differential privacy nonetheless provides a better way of assuring privacy given chance identification. Significant further advances are also to be found in recent work on building DP Bayes Nets[84].

15. Caveats

Some careful analysis needs to be further developed with regard to possible limitations of the synthetic data solution. Researchers from one of MyHealthMyData partners, SBA, have in particular, highlighted some potential attribute disclosure risks, to be countered with

DRAFT

[81] M. Langarizadeh, A. Orooji, A. Sheikhtaheri, 12th Annual Conference on Health Informatics Meets eHealth, eHealth 2018. "Effectiveness of anonymization methods in preserving patients' privacy: A systematic literature review," Studies in Health Technology and Informatics, 248, pp. 80-87, 2018.

[82] L. Rocher, J. M. Hendrickx & Y. A. de Montjoye, Estimating the success of re-identifications in incomplete datasets using generative models, NATURE COMMUNICATIONS | (2019) 10:3069

[83] J. Jordon, J. Yoon, M. van der Schaar, PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees, ICLR 2019.

[84] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, X. Xiao, PrivBayes: Private Data Release via Bayesian Networks, ACM Transactions on Database Systems. 1, September 2017.

DRAFT

differential privacy, though globally validating synthetic data robustness in terms of identity protection[85,86,8788].

16. Synthetic data and AI biases

As AI capabilities are boosted by innovation in algorithm design, the old saying that “lots of data beat algorithms and good data beat lots of data” remains, more than ever, true.

Yet, the quality of biomedical data sets is traditionally poor, due to the complexities and cost of measuring biological and clinical parameters. This leads to two sets of issues: AI systems trained on poor-quality data struggle to achieve sufficient accuracy, and even when they reach it, they carry the risk of incorporating systemic biases that skew their behavior in detrimental or unethical ways.

Recent cases have brought to the public’s attention worrisome examples of AI biased in important social functions such as human resource management or judiciary processes. The issue affects well curated data too. In medicine, for instance, rare diseases or uncommon presentation of more frequent ones offer unique challenges in terms of data availability. Their low prevalence leads to systematic understudy. A diagnostic algorithm trained on commonly available data will tend to underdiagnose them, compounding the underestimation of their actual incidence.

Synthetic data in such a case can be used to extend underrepresented populations and rebalance outputs. They can also be used to influence AI behavior in ways that even the most

[85] M. Hittmeir, A. Ekelhart, and R. Mayer, “On the utility of synthetic data: An empirical evaluation on machine learning tasks,” in 14th International Conference on Availability, Reliability and Security (ARES 2019), Canterbury, United Kingdom, August 26–29 2019.

[86] M. Hittmeir, A. Ekelhart, and R. Mayer, Utility and Privacy Assessments of Synthetic Data for Regression Tasks, Proceedings of the IEEE International Conference on Big Data (IEEE BigData 2019).

[87] J. Taub, M. Elliot, M. Pampaka, and D. Smith, Differential Correct Attribution Probability for Synthetic Data: An Exploration, in: Privacy in Statistical Databases (Lecture Notes in Computer Science), J. Domingo-Ferrer and F. Montes (Eds.), Springer International Publishing, Valencia, Spain, 122–137, 2018.

[88] M. Hittmeir, R. Mayer, and A. Ekelhart, A Baseline for Attribute Disclosure Risk in Synthetic Data, Proceedings of 10th ACM Conference on Data and Application Security and Privacy (CODASPY 2020), ACM, New York, 2019.

exhaustive training data set cannot, as in the case of social inequalities which are indeed factual and thus expressed in even the most accurate snapshots of general population and, if not corrected, are simply perpetrated by AI-systems.

When direct data manipulation, such as down-weighting or rebalancing key variables with ethical dimensions (e.g., race), is not feasible or cost-effective, synthetic data represent a low-cost solution to correct algorithms' behaviors toward socially responsible decisions.

Ethically desirable scenarios can be, in this sense, deterministically implemented in the data. Artificial intelligence is then trained on enriched, real-life information reflecting a consciously chosen digital structure in which "data is still king but ruling as benevolent monarch rather than prejudiced patriarch"[89].

17. A call to action

Synthetic data show a promising path toward a secure and scalable biomedical data economy, which deserves to be fostered and framed in depth, if it is to grow. This solution, in our view, stands to become the most valuable currency in the healthcare and pharmaceutical, artificial intelligence ecosystem.

Non-European institutions, such as the US NIST, are taking actions to foster this technology in main-stream industrial applications and research, but a scientific and strategic direction is missing in terms of technological best practices, legal interpretations and market development.

Technology advancements have opened the way to real-world production and use of synthetic data at industrial levels, but in absence of validated frameworks for their creation, validation and use, commercial and institutional players are not likely to adopt them in time for placing Europe at the fore front of this potential revolution. Standards and user workflows for the selection and management of most appropriate data generators, for their configuration, would allow unskilled data analysts to confidently assemble generative pipelines to serve their organizations' R&D goals. Similarly, robust proof of statistical reliability will foster their adoption by key decision makers.

This Discussion Paper draft will be extensively discussed and improved upon with contributions from experts and stakeholders. This is just a first step, in the belief that the time is ripe for an

[89] S. M. Bellovin, P. K. Dutta, and N. Reitinger, Privacy and Synthetic Datasets, cit.

DRAFT

initiative by the European Commission that will fill the void of legal and technical definitions, as well as of policy, and ease with this solution the difficulties with health data sharing while fully implementing both the spirit and the letter of the GDPR.

A Coordination and Support Action call, bringing together leading private and institutional centers around the demonstrated potential of synthetic data, to define a general framework and roadmap for their main stream implementation and market adoption, would be the stepping stone to activate a thriving Digital Single Market for the biomedical sciences.

DRAFT