15

AI and Healthcare Data

Griet Verhenneman

15.1 INTRODUCTION

A strict regulatory trajectory must be followed to introduce artificial intelligence in healthcare. Each stage in the development and improvement of AI for healthcare is characterized by its own regulatory framework. Let us consider AI-assisted cancer detection in medical images. Typically, the development and testing of the algorithms indicating suspicious zones requires setting up one or more clinical trials. During the clinical research stage, regulations such as the Clinical Trials Regulation apply.[1] When the results are good, the AI-assisted cancer detection software may be deployed in products such as MRI scanners. At that moment, the use of AI-assisted cancer detection software becomes standard-of-care and (national) regulatory frameworks on patients' rights must be considered. However, after the introduction of the AI-assisted cancer detection software to the market, post-market rules will require further follow-up of product safety. These regulatory instruments are just a few examples. Other identified risks, such as violations of medical secrecy or fundamental rights to the protection of private life and personal data, have led regulators to include specific rights and obligations in regulatory initiatives on the processing of personal data (such as the General Data Protection Regulation, hereinafter "GDPR"),[2] trustworthy artificial intelligence (such as the AI Act),[3] fair governance of personal and nonpersonal data (such as the Data Governance Act)[4]

[1] Regulation (EU) 536/2014 of the European Parliament and of the Council of April 16, 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC 2014.

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 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 (General Data Protection Regulation) 2016.

[3] Regulation (EU) 2024/1689 of the European Parliament and the Council of June 23, 2024, laying down harmonised rules on artificial intelligence and amending Regulation (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

[4] Regulation (EU) 2022/868 of the European Parliament and of the Council of May 30, 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) 2022.

and the proposal for a Regulation on a European Health Data Space (hereinafter "EHDS").[5]

The safety of therapies, medical devices, and software is a concern everyone shares, whether or not they include AI. After all, people's lives may be at stake. Previously, incidents with more classic types of medical devices, such as metal-on-metal hip replacements[6] and PIP breast implants,[7] have led regulators to adapt the safety monitoring processes and adopt the Medical Devices Regulation and In Vitro Medical Devices Regulation.[8] When updated in 2017, these regulatory frameworks not only considered "physical" medical devices but clarified the requirements also for software as a medical device.[9] Following the increased uptake of machine learning methods and the introduction of decision-supporting and automated decision-making software in healthcare, regulators deemed it necessary to act more firmly and sharpen regulatory oversight also with respect to software as a medical device.

Throughout the development and deployment of AI in healthcare, the collection and use of data is a connecting theme. The availability of data is a condition for the development of AI. It should arrest our attention that data availability is also a regulatory requirement, especially in the healthcare sector. The collection of data to establish sufficient evidence, for example, on product safety, is not only a requirement for the development of AI but also for the permanent availability of AI-driven products on the market. Initiatives such as the Medical Devices Regulation and the AI Act have indeed enacted obligations to collect data for the purpose of establishing (evidence of) the safety of therapies, devices, and procedures.

[5] Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space, COM(2022) 197 final, May 3, 2022.

[6] Metal-on-metal hip replacements are all-metal implants whereby a metal ball replaces the femur, and a metal cup is created in the hip bone to keep the ball in place. When moving, the ball's metal surface touches the cup's metal surface, causing friction. Investigations by, amongst others, the BMJ and BBC Newsnight had raised concerns over metal hip implants causing people to be exposed to dangerously high levels of toxic metals. Despite the risks being known and documented since 2008, metal-on-metal hip implants continued to be used without having to pass any clinical trials. See www.bmj.com/press-releases/2012/02/28/joint-bmj-bbc-newsnight-investigation-raises-new-concerns-over-metal-hip-i.

[7] The PIP breast implant was a silicon gel-based breast implant used for breast augmentation or reconstruction. The company developing the implant, Poly Implant Prothèse, had used an industrial-grade silicone that later proved to cause health risks. Gaps in the approval process had made it possible for the silicone to be used for ten years after the first indications of health risks. See Victoria Martindale and Andre Menache, "The PIP scandal: An analysis of the process of quality control that failed to safeguard women from the health risks" (2013) *Journal of the Royal Society of Medicine*, 106: 173. This case is also discussed in Chapter 12 of this Handbook, authored by Nathalie A. Smuha and Karen Yeung.

[8] Regulation (EU) 2017/745 of the European Parliament and of the Council of April 5, 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002; and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC 2017; Regulation (EU) 2017/746 of the European Parliament and of the Council of April 5, 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU 2017.

[9] Article 2(1) and Recital 17 Medical Device Regulation provide that the software may be qualified as a medical device depending on its intended purpose.

Even though the collection of data is imposed as a legal obligation, the processing of personal data must be compliant with the GDPR. Especially in healthcare applications, AI typically requires the processing of special-category data. The GDPR considers personal data as special-category data when, due to their nature, the processing may present a higher risk to the data subject. In principle, the processing of special-category data is prohibited, while exemptions to that prohibition are specified.[10] Data concerning the health of a natural person is qualified as special-category data. Often health-related data are collected in the real world from individual data subjects.[11] Regulatory instruments such as the AI Act or the proposal for a Regulation on the EHDS explicitly mention that they shall be without prejudice to other Union legal acts, including the GDPR.[12]

Since the (re-)use of personal health-related data is key to the functioning and development of artificial intelligence for healthcare, this chapter focuses on the role of data custodians in the healthcare context. After a brief introduction to real-world data, the chapter first discusses how law distinguishes data ownership from data custodianship. How is patient autonomy embedded in the GDPR and when do patients have the right to agree or disagree via opt-in or opt-out mechanisms? Next, the chapter discusses the reuse of health-related data and more specifically how they can be shared for AI in healthcare. Federated learning is discussed as an example of a technical measure that can be adopted to enhance privacy. Transparency is discussed as an example of an organizational measure. Anonymization and pseudonymization are introduced as minimum measures to always consider before sharing health-related data for reuse.

15.2 PRE-AI: THE REQUEST FOR HEALTH-RELATED DATA

Whether private or public, hospitals and other healthcare organizations experience increasing requests to share the health-related data they collected in the "real world." "Real-world data" are relied on to produce "real-world evidence," which is subsequently relied on to support the development and evaluation of drugs, medical devices, healthcare protocols, machine learning, and AI.

Real-world data (hereinafter RWD) are collected through routine healthcare provision. Corrigan-Curay, Sacks, and Woodcock define RWD as "*data relating to patient health status or the delivery of health care routinely collected from a variety of sources, such as the [Electronic Health Record] and administrative data*."[13]

[10] Article 9, 1. of the GDPR specifies the prohibition of processing special-category personal data. Article 9, 2. of the GDPR explains that the prohibition does not apply if one of the listed exemptions can be referred to.

[11] The data are typically subject to Article 9 General Data Protection Regulation.

[12] Article 2, 7. AI Act; Article 1, 4. proposal for a Regulation on EHDS.

[13] Jacqueline Corrigan-Curay, Leonard Sacks, and Janet Woodcock, "Real-world evidence and real-world data for evaluating drug safety and effectiveness" (2018) *The Journal of the American Medical Association*, 320: 867.

The data are, in other words, collected while healthcare organizations interact with their patients following a request from the patient. RWD result from anamneses, medicinal and non-medicinal therapies, medical imaging, laboratory tests, applied research taking place in the hospital, medical devices monitoring patient parameters, and, for example, claims and billing data. Real-world evidence (hereinafter RWE) is evidence generated through the use of RWD to complement existing knowledge regarding the usage and potential benefits or risks of a therapy, medicinal product, or device.[14]

Typically healthcare providers use an electronic health record (hereinafter EHR) to collect health-related data per patient. The EHR allows healthcare providers, working solo or in a team, to access data about their patients to follow up on patient care. However, an EHR is typically not set up to satisfy data-sharing requests for a purpose other than providing healthcare. The EHR's functionalities are chosen and developed to allow a high-quality level of care, on a continuous basis, for an individual patient. These functionalities are not necessarily the same functionalities that are needed to create reliable and trustworthy AI.

First of all, AI needs structured data. Today, most EHRs contain structured data to a certain level, but apart from structured data, most EHRs also contain a high level of natural language text. This text needs interpretation before it can be translated to structured databases suitable to feed AI applications. Even today existing AI-supported tools for deciphering natural language text were once fed with structured data on, for example, medical diagnoses, medication therapies, medication components,… as well as street names, and first and second names, for instance. The need for universal coding languages, such as the standards developed by HL7, has been long-expressed in medical informatics.[15]

Secondly, AI does, in general, not need patient names. Inevitably, an EHR, however, must allow direct identification of patients. When considering safety risks in healthcare, the misidentification of a patient would be regarded as a severe failure. Therefore, internationally recognized accreditation schemes for healthcare organizations will oblige healthcare practitioners to check multiple identifiers to uniquely identify the patient before any intervention. When EHR data are used for secondary purposes, such as the development of AI, data protection requirements will encourage the removal of patient identifiers (entirely or to the extent possible).[16]

[14] Anuradha Ramamoorthy and Shiew-Mei Huang, "What does it take to transform real-world data into real-world evidence?" (2019) *Clinical Pharmacology & Therapeutics*, 106: 10.

[15] Health Level Seven International (HL7) is an organization that develops standards to allow global health data interoperability. For more information, see www.hl7.org/about/index.cfm?ref=nav; D. M. López and B. Blobel, "Architectural approaches for HL7-based health information systems implementation" (2010) *Methods of Information in Medicine*, 49: 196.

[16] This idea is capsulated in the data minimization principle and described as a technical measure to protect data from unlawful use. See, for example, Articles 5, 1. (b), 32 and 89 GDPR.

Therefore, data holders increasingly prepare the datasets they primarily collected for the provision of healthcare to allow secondary use. While doing so, data holders will "process" personal health-related data in the sense of Article 4 (2) of the GDPR. Consequently, they must consider the principles, rights, and obligations imposed by the GDPR. They must do so when preparing data for secondary purposes they define themselves and when preparing data following instructions of a third party requesting data. In the following paragraphs, it will be explained that data holders must consider technical and organizational measures to protect personal data at that moment.

15.3 DATA OWNER- OR CUSTODIANSHIP?

Especially in discussions with laypeople, it is sometimes suggested that patients own their data. However, in the legal debate on personal and nonpersonal data, the idea of regulating the value of data in terms of ownership has been abandoned largely.[17]

First, while it is correct that individual-level health-related data are available only after patients have presented themselves, it is incorrect to assume that only patients contribute to the emergence of health-related data. Many others contribute knowledge and interpretations. Physicians, for example, build on the anamneses and add their knowledge to order tests, conclude about the diagnosis, and suggest prescriptions. Nurses observe the patient while at the hospital and while they register measurements, frequencies, and amounts. Lab technicians receive samples, run tests, and return inferred information about the sample. All of those actions generate relevant data too.

Second, from a legal perspective, it should be stressed that ownership is a right *in rem*.[18] Considering data ownership would imply that an exclusive right would rest on the data. If we were to consider the patient as the owner of their health-related data, we would have to acknowledge an exclusive right to decide who can have, hold, modify, or destroy the data (and who cannot). EU law does not support such a legal status for data. On the contrary, when considering personal data, it should be stressed that a salient characteristic of the GDPR is the balance it seeks between the individual's rights and society's interests. The fundamental right to the protection of personal data is not and has never been an absolute right. Ducuing indicates that

[17] See, for example, Kathleen Liddell, David A. Simon, and Anneke Lucassen, "Patient data ownership: Who owns your health?" (2021) *Journal of Law and the Biosciences*, 8: lsab023; Gianclaudio Malgieri, "'User-provided personal content' in the EU: Digital currency between data protection and intellectual property" (2018) *International Review of Law, Computers & Technology*. 32: 118.

[18] The European Union Intellectual Property Office describes a "right in rem" also as a "real right" or a right that reflects the absolute right to recover, possess and enjoy a specific object. Rights in rem are directed toward an object rather than a person and therefore differ from rights "in personam." See EUIPO Trade mark guidelines, Part E, Section 15.3, Chapter 2: Licenses, rights in rem, levies of execution, insolvency proceedings, entitlement proceedings or similar proceedings, ed. 2023.

more recent regulatory initiatives (such as the Data Governance Act) present "traces" of data ownership to organize the commodification and the economic value of data as a resource. The "traces," Ducuing concludes, seem to suggest a somewhat functional approach in which, through a mixture of legal sources, including ownership and the GDPR, one aims to regulate data as an economic resource.[19]

Instead, it is essential to consider data custodianship. The custodian must demonstrate a high level of knowledge and awareness about potential risks for data subjects, especially when they are in a vulnerable position, such as patients. Data custodians should be aware of and accept the responsibility connected to their role as a guardian of personal data. In healthcare organizations, the pressure is high to see to the protection of health-related data kept in an EHR and to ensure attention for the patient as the data subject behind valuable datasets, and rightfully so. Not more than patients, data custodians should consider EHR data as "their" data in terms of ownership. They are expected to consider the conditions for data sharing carefully, but they should not hinder sharing when the request is legitimate and lawful.

15.3.1 *Custodianship and Patient Autonomy*

When considering patient autonomy as a concept reflecting individuality,[20] the question arises how the GDPR allows the data subject to decide autonomously about the reuse of personal data for the development or functioning of AI. While, as explained earlier, data protection is not enacted as an absolute right, patients can decide autonomously about the processing of their data unless the law provides otherwise. In general terms, Article 8 of the European Convention on Human Rights and Article 52 of the Charter of Fundamental Rights of the European Union provide that limitations to the fundamental rights to the respect for private life and the protection of personal data shall be allowed only when necessary in a democratic society and meeting the objectives of general interest or the protection of rights and freedoms of others. A cumulative reading of Articles 6 and 9 of the GDPR can establish a more concrete interpretation of this general principle. Together, Articles 6 and 9 of the GDPR provide the limitative list of situations in which the (secondary)

[19] Charlotte Ducuing, "What can we still learn from data ownership? The traces of ownership in the regulation of data as an economic resource" (ELI Digital Law SIG Seminar, online, June 1, 2022).

[20] John Stuart Mill adopted the concept of individuality as a characteristic of the self-determining and self-ruling subject reflecting authentic subjective preferences. Many refer to the work of John Stuart Mill when discussing the patient as an autonomous individual. See, for example, Thomas Nys, Yvonne Denier, and Toon Vandevelde, *Autonomy & Paternalism: Reflections on the Theory of Health Care* (Peeter Publishers, 2007). For a more extensive discussion on the role of autonomy in relation to the processing of health-related data see also Griet Verhenneman, *The Patient, Data Protection and Changing Healthcare Models: The Impact of e-Health on Informed Consent, Anonymisation and Purpose Limitation* (Intersentia, 2021), www.cambridge.org/core/books/patient-data-protection-and-changing-healthcare-models/5B12AE59BE02759D9762B14C768E5FD5, accessed February 19, 2023, 137–140.

use of personal health-related data is allowed without the patient's consent.[21] In these situations, the data subject's wish is considered to not necessarily prevail over the interests of other parties or society. Examples include the collection of health-related data for the treatment of a patient. Depending on specifications in Member State law, the collection can be based on Article 6, 1. (b) "performance of a contract to which the data subject is party" or 6, 1. (c) "legal obligation to which the data controller is subject" on the one hand, and Article 9, 2. (h) "necessary for the provision of health" on the other hand.[22] A national cancer screening program is another example. In this case, the data collection is typically enacted in Member State law, causing Article 6, 1. (e) "performance of a task in the public interest" to apply in combination with Article 9, 2. (h) "necessary for purposes of preventive medicine."

Another situation in which the data subject's individual wishes do not prevail over society's interest concerns scientific research. By default, data can be reused for scientific research. The data subject's consent (opt-in) is not required, and when in the public interest, the data subject does not even enjoy a right to opt out.[23] First of all, Article 5, 1. (b) of the GDPR provides a specification of the purpose limitation principle indicating that "*further processing for […] scientific […] research purposes […] shall, in accordance with article 89 (1), not be considered to be incompatible with the initial purpose*." Additionally, Article 9, 2. (j) provides that contrary to the general prohibition to process health-related data, the processing is allowed when necessary for the purpose of scientific research. The application of Article 6.4 of the GDPR to the secondary use of personal data has raised some discussions, but not in a research context. Read together with Recital 50, Article 6.4. of the GDPR indicates that a new legal basis is not required when the secondary processing can be compatible with the primary processing. A combined reading of Article 6.4. and Article 5, 1. (b) has convinced many[24] that a new legal basis is indeed not required when the purpose of the secondary processing is scientific research.[25]

[21] On the need for a legal basis to process personal data, see also Chapter 7 of this book, authored by Pierre Dewitte.

[22] Some Member States qualify the patient–doctor relationship as contractual. Some Member States impose a legal obligation for the healthcare practitioner to keep a(n) electronic health record for each patient.

[23] Article 21, 6. GDPR.

[24] Dutch Data Protection Authority "Adviesverzoek onderzoek oversterfte," February 13, 2022, available online: www.autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/advies_ap_onderzoek_oversterfte.pdf; Request for advice European Data Protection Board, "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) 03/2020 6; Evelien De Sutter et al., "Rethinking informed consent in the time of COVID-19: An exploratory survey" (2022) 9 Frontiers in Medicine 995688; G. Verhenneman et al., "How GDPR enhances transparency and fosters pseudonymisation in academic medical research" (2020) *European Journal of Health Law*, 27: 35.

[25] About other types of secondary use (such as post-market monitoring of a medical device, market authorization applications, and reimbursement dossiers submitted to national health programs), the application of especially Recital 50 seems to be somewhat more contested. See Mahsa Shabani and

It should, however, be noted that notwithstanding the intention of the GDPR to achieve a higher level of harmonization, one specific provision should not be overlooked when discussing patient autonomy in relation to health-related data. Article 9, 4. of the GDPR foresees that Member States may introduce further restrictions on the processing of health-related, genetic, and biometric data.[26] Building on this provision, some Member States have introduced the obligation to obtain informed consent from the individual as an additional measure to empower patients.[27]

15.3.2 *Informed Consent for Data Processing*

When the purpose for which data are shared cannot be covered by a legal basis available in Article 6 and an additional safeguard as laid down in Article 9 of the GDPR, the (valid) informed consent of the patient should be sought prior to the secondary processing. In that case, the requested informed consent should reflect patient autonomy. The conditions for valid informed consent, as laid down in Articles 4 (11) and 7 of the GDPR, indicate that the concept of informed consent was developed as an instrument for individuals to express their wishes and be empowered. These articles stress that consent must be freely given, specific, informed, and reflect an unambiguous indication of the data subject's wishes. The controller shall be able to demonstrate that the data subject has consented and shall respect the fact that consent can be withdrawn at any time, with no motivation required.

These requirements may sound obvious, but they are challenging to fulfill in practice. In particular, the fact that for informed consent to be freely given a valid alternative for not providing consent for the processing of personal data should be available is often an issue.[28] Typically, data processing is a consequence of a service, product, or project,… especially in the context of AI. I cannot agree to participate in a data-driven research project to develop AI for medical imaging without allowing my MRI scan to be processed. I cannot use an AI-supported meal app that provides personalized dietary suggestions while not allowing data about my eating habits to be shared. I cannot use an AI-driven screening app for skin cancer without allowing

Sami Yilmaz, "Lawfulness in secondary use of health data: Interplay between three regulatory frameworks of GDPR, DGA & EHDS" (2022) *Technology and Regulation*, 2022: 128.

[26] Article 9, 4. GDPR.

[27] These conclusions are based on the work done on the secondary use of personal data for research purposes by a consortium led by MILIEU, with contributions from UNamur (CRIDS/NaDI), KULeuven (CiTIP), University of Leiden (eLaw) and Vrije Universiteit Amsterdam (CLI). The authors of the study are Teodora Lalova, Els Kindt, Eleftherios Chelioudakis, Griet Verhenneman, Antoine Delforge and Jean Herverg. National-level input was provided by Carla Barbosa, Elisabetta Biasin, Gauthier Chassang, Eleftherios Chelioudakis, Athena Christofi, Agnes Csonta, Antoine Delforge, Ivo Emanuilov, Danaja Fabcic, Nenad Georgiev, Dara Hallinan, Erik Kamenjasevic, Linda de Keyser, Karolina La Fors, Teodora Lalova, Zuzana Lukacova, Sjaak Nouwt, Domenico Orlando, Anastasia Siapka, Griet Verhenneman, and Katerina Yordanova.

[28] See, for example, reservations expressed by the European Data Protection Supervisor, "A preliminary opinion on data protection and scientific research" (2020) 19–21.

a picture of my skin to be uploaded. In this case, it should be questioned whether data can be reused or shared for secondary purposes based on informed consent.

15.4 SHARING DATA FOR AI IN HEALTHCARE

"*Data have evolved from being a scarce resource, difficult to gather, managed in a centralized way and costly to store, transmit and process, to becoming an abundant resource created in a decentralized way (by individuals or sensors) easy to replicate, and to communicate or broadcast on a global scale.*"[29] This is how the European Union Agency for Cybersecurity (ENISA) introduces her report on how to ensure privacy-preserving data sharing. The quote is illustrative not only for the naturalness with which we think about keeping data for secondary use but also for the seemingly infinite number of initiatives that can benefit from the reuse of data, including personal data. In that sense, sharing health-related data differs significantly from sharing human bodily material. While the number of projects that can benefit from one sample of bodily material is, per definition, limited to, for example, the number of cuts that can be made, the reuse of data only ends when the data itself have become irrelevant.

It is essential to stress that facilitating data sharing is also a specific intent of regulators. Policy documents on FAIR data,[30] open science initiatives, and the proposal for a European Health Data Space are just a few examples hereof. "*Sharing data is already starting to become the norm and not the exception in data processing,*" ENISA continues.[31] Even in the GDPR itself, it is stated that: "*The free movement of personal data within the Union shall be neither restricted nor prohibited for reasons connected with the protection of natural persons with regard to the processing of personal data.*"[32] Although frustrations over the rigidity of the GDPR sometimes seem to gain the upper hand, also in discussions on the secondary use of data, the goal of the Regulation is thus not to hamper but to facilitate the processing of personal data.

During the COVID-19 pandemic, several authors stressed this fundamental assumption also in relation to health-related data. Albeit specific requirements must be met, the processing of personal health-related data is not necessarily not allowed.[33] Two weeks after the outbreak, the European Data Protection Board, for example, issued a statement indicating that "*data protection rules (such as the GDPR) do not*

[29] ENISA, "Engineering personal data sharing" (2023) Report/Study www.enisa.europa.eu/publications/engineering-personal-data-sharing, accessed February 22, 2023.

[30] "FAIR" stands for Findable, Accessible, Interoperable and Reusable. The FAIR Data Principles are used as a guideline for those wishing to enhance the reusability of the data.

[31] Ibid.

[32] Article 1 GDPR.

[33] Marcello Ienca and Effy Vayena, "On the responsible use of digital data to tackle the COVID-19 pandemic" (2020) *Nature Medicine*, 26: 463.

hinder measures taken in the fight against the coronavirus pandemic."[34] Several possible exemptions that would allow the processing of health-related data in the fight against COVID-19 were stressed and explained. The European Data Protection Board (EDPB) pointed at the purpose limitation and transparency principle and the importance of adopting security measures and confidentiality policies as core principles that should be considered, even in an international emergency.

To meet these principles, so-called data protection- or privacy-enhancing measures must be considered. Different privacy-enhancing techniques can be applied to the data flows and infrastructures. At the operational level of a healthcare organization, suggestions for privacy-preserving techniques profiles such as data protection officers, compliance officers, or the chief information security officer typically suggest the implementation of measures. "*It used to be the case that if you did nothing at all, you would have privacy […]. Now, you need to take conscious, deliberate, intentional actions to attain any level of privacy. […] This is why Privacy Enhancing Technologies (PETs) exist*," writes Adams referring to technical measures that can be implemented to better protect data about individuals.[35] Examples of such PETs include pseudonymization through polymorphic encryption[36] and federated learning, but next to technical measures, organizational measures such as transparency must also be considered.

The following sections illustrate the impact and necessity of privacy-enhancing measures in health-related scenarios. Anonymization and pseudonymization are discussed first. They are considered minimum measures to consider before reusing personal data. However, because anonymous data are considered out of the material scope of the GDPR while pseudonymous data are considered in scope, it is essential to understand the difference between them. Next, by discussing two other examples of privacy-enhancing techniques, one technical and one organizational, it is illustrated how anticipating the technical and the organizational aspects of a data flow help to ensure the robust protection of personal data as an "abundant resource."

15.4.1 *Anonymization and Pseudonymization*

In the GDPR, a preference for the use of anonymized data over pseudonymized and non-pseudonymized data is expressed, for example, in the data minimization principle, as a security measure and in relation to scientific research.[37] The use of

[34] Andrea Jelinek, "Statement by the EDPB chair on the processing of personal data in the context of the COVID-19 outbreak" (2020), https://edpb.europa.eu/news/news/2020/statement-edpb-chair-processing-personal-data-context-covid-19-outbreak_en.

[35] Carlisle Adams, *Introduction to Privacy Enhancing Technologies: A Classification-Based Approach to Understanding PETs* (Springer International Publishing, 2021), https://link.springer.com/10.1007/978-3-030-81043-6, accessed February 21, 2023. p. 2.

[36] ENISA (n 29) 13.

[37] Articles 5, 1. (c); 32 and 89 GDPR in particular.

anonymized data is considered to present a sufficiently low risk for the data subject's fundamental rights to allow the processing without any further measures, and is hence excluded from the GDPR's requirements.[38] Pseudonymized data, however, fall under the GDPR because the data can still be attributed to an individual data subject.[39]

In healthcare and other data-intensive sectors, for data not to fall under the definition of personal data, as provided in Article 4(1) of the GDPR, is increasingly difficult due to enhanced data availability and data linkability.[40] Data availability relates to the number of data kept about individuals. Data are not only kept in EHRs but spread over many other datasets held by public and private organizations. Data linkability relates to the ease with which data from different datasets can be combined. Machine learning and other types of AI have a distinct impact in this sense as they facilitate this process.

Requirements on open science,[41] explainability,[42] and citizen empowerment[43] stimulate data holders to increase the level of data availability and linkability. To create innovations this is a great assumption, but there is another side to the coin. A higher level of data availability and linkability requires data holders, such as healthcare organizations, to increasingly qualify data as pseudonymous rather than anonymous.

Influential studies continue to show limitations in anonymization techniques in relation to patient data. Schwarz et al., for example, reidentified patients based on de-identified MRI head scans, which were released for research purposes. Schwarz's research team showed that in 83% of the cases, face-recognition software matched

[38] Art 4(1) and Recital 26 GDPR.

[39] Art 4(5) GDPR.

[40] For a detailed analysis of the qualification of health-related data as personal data under Article 4(1) of the GDPR, see Griet Verhenneman, *The Patient, Data Protection and Changing Healthcare Models: The Impact of e-Health on Informed Consent, Anonymisation and Purpose Limitation* (Intersentia, 2021), www.cambridge.org/core/books/patient-data-protection-and-changing-healthcare-models/5B12AE59BE02759D9762B14C768E5FD5, accessed February 19, 2023.

[41] Open science is considered the standard working method for all European Union research and innovation funding programs. Beneficiaries must make their data available, including source data, as open as possible. See European Union, Unit Research, and Innovation, "Open Science" (2019), available at: https://research-and-innovation.ec.europa.eu/system/files/2019-12/ec_rtd_factsheet-open-science_2019.pdf.

[42] Explainability refers to the idea that technology, such as AI, should not be a black box but allow transparency and traceability to enable humans to understand the decisions made through artificial intelligence. See Andreas Holzinger et al., "Causability and explainability of artificial intelligence in medicine" (2019) *WIREs Data Mining and Knowledge Discovery*, 9: e1312.

[43] Especially, while not exclusively, in the context of genomic research, the idea of returning results of scientific research to individual study participants has been suggested as an essential requirement for achieving justice, beneficence, and respect for persons. To allow the return of results, the participant must remain re-identifiable, for example, by attaching a unique code to the human bodily samples obtained for research purposes. See Emmanuelle Lévesque, Yann Joly, and Jacques Simard, "Return of research results: general principles and international perspectives" (2011) *The Journal of Law, Medicine & Ethics*, 39: 583.

an MRI with a publicly available picture. In 95% of the cases, the image of the actual patient was amongst the five selected public profiles.[44] Studies such as Schwarz's led to the development of "defacing techniques," a privacy-enhancing measure to hinder the reidentification of head scans.[45] However, is the hindrance caused by the defacing technique sufficient for the scan to qualify as nonpersonal data?

To answer that question, it is important to stress that the scope of the GDPR is not delineated based on the presence of certain specific identifiers in a particular dataset. Contrary to, for example, the US Health Insurance Portability and Accountability Act (HIPAA),[46] which provides that individually identifiable information can be de-identified by removing the listed identifiers (exhaustive account) from the dataset, the GDPR requires a more complex assessment. The possibility for the controller or another person to single out a data subject building on the information in the dataset *and any additional information* that can be obtained using all the means reasonably likely must be evaluated. When considering the MRI image, this means that account must be taken of the MRI image with defacing techniques applied, pictures available on the internet, *and* the original MRI available in the EHR even when this image is not available to the data controller.[47]

15.4.2 *Federated Learning, an Example of a Privacy-Enhancing Technical Measure*

Federated analysis allows for building knowledge from data kept in different local sources (such as various EHRs in hospitals, public health databases in countries, or potentially even individual health "pods" kept by citizens[48]) while avoiding the transfer of individual-level data.[49] Hence, federated analysis is presented as a solution to avoid the centralization of (personal) health-related data for secondary use.

Imagine building an AI model for cancer detection through MRI images: In a nonfederated scenario, the MRI images are requested through multiple participating hospitals, pseudonymized, and subsequently collected in a central, project-specific database. The algorithm is trained on the central database. In a federated scenario, however, the MRI images are not pooled in a central database. Instead, they remain with the local hospital. The algorithmic model, carrying out analytical tasks, visits the local databases ("nodes") and executes tasks on the locally stored

[44] Christopher G. Schwarz et al., "Identification of anonymous MRI research participants with face-recognition software" (2019) *The New England Journal of Medicine*, 381: 1684.

[45] Elizabeth EL Buimer et al., "De-identification procedures for magnetic resonance images and the impact on structural brain measures at different ages" (2021) *Human Brain Mapping*, 42: 3643.

[46] Health Insurance Portability and Accountability Act [HIPAA] of 1996, Pub. L. No. 104–191.

[47] Article 29 Working Party Opinion 05/2014 on Anonymisation Techniques, April 10, 2014, 9.

[48] Hemant Ghayvat et al., "SHARIF: Solid pod-based secured healthcare information storage and exchange solution in internet of things" (2022) *IEEE Transactions on Industrial Informatics*, 18: 5609.

[49] Felix Nikolaus Wirth et al., "Privacy-preserving data sharing infrastructures for medical research: systematization and comparison" (2021) *BMC Medical Informatics and Decision Making*, 21: 242.

MRI images.[50] Subsequently, aggregated results (the conclusions) are shared with a central node for merging and meta-analysis. On the condition of a small cell risk analysis,[51] these results can often be considered nonpersonal data because individual patients can no longer be singled out.

Avoiding centralization is particularly interesting because it can reduce the risk of illicit data usage. The control on the secondary use remains with the data holder: A data custodian (such as a hospital), the individual (such as a patient), or perhaps, as suggested in Article 17 et seq of the Data Governance Act, a recognized data altruism organization.[52] Unlike organizational measures, such as contractual arrangements on the purpose of the processing, federated learning thus allows the data holder to manage the processing independently.

The implementation of federated learning should, however, not trigger the assumption that the processing operations are not covered by the material scope of the GDPR. Federated learning does not avoid the processing of personal data for a secondary purpose. It merely avoids the transfer of personal data. In other words: the processing takes place locally, but data are reused for a purpose different from the purpose for which they were initially collected. Consequently, GDPR requirements must be complied with, including the need for a legal basis.

Following Article 4 (7) of the GDPR, the party defining the purpose and (at least the essential) means of the secondary use should be considered the data controller. Generally, the requestor, not the requestee, defines the purpose and means of the secondary processing. Therefore the requestor is considered the data controller.[53] The location of the data processing (locally or centrally) is irrelevant. Who has access to the data is equally irrelevant.[54] Consequently, although a data transfer

[50] Oya Beyan et al., "Distributed analytics on sensitive medical data: The personal health train" (2020) *Data Intelligence*, 2: 96; J. Simm et al., "Splitting chemical structure data sets for federated privacy-preserving machine learning" (2021) *Journal of Cheminformatics*, 13: 96; A. Ardeshirdavani et al., NGS-logistics: Federated analysis of NGS sequence variants across multiple locations (2014), *Genome Medicine*, 6: 17.

[51] Depending on the level of specificity in the number of individuals that are included in the aggregation, aggregated data may still allow the extraction of personal data. Hence, when considering the data that are shared with the central node as anonymous one should first assess in how far individuals can still be singled out. See also N. Truong et al., "Privacy preservation in federated learning: An insightful survey from the GDPR perspective" (2021) *Computers & Security*, 110: 102402.

[52] Article 18 Data Governance Act defines the tasks of a data altruism organization.

[53] Nevertheless, the qualification of the requestee as a joint controller is a possibility we must consider. Should the requestee and the requestor determine the secondary purpose of the processing together, they will be qualified as joint-controllers. This may be the case when the parties are setting up a research project together and determining the project's scope, research questions, work packages, and tasks within the work packages. See Brendan Van Alsenoy, *Data Protection Law in the EU: Roles, Responsibilities and Liability* (Intersentia 2019) 331–334, https://intersentia.be/nl/data-protection-law-in-the-eu-roles-responsibilities-and-liability-48825.html, accessed February 24, 2023.

[54] Parties may be qualified as data controller even if they do not have access to the data, the European Court of Justice confirmed, see *Tietosuojavaltuutettu (Supreme Administrative Court Finland) vs Jehovan todistajat – uskonnollinen yhdyskunta (Jehovah's witnesses)* [2018] European Court of Justice (Grand Chamber) EU:C:2018:551.

agreement may be avoided when sharing merely anonymous data with the central node, a data processing agreement (or joined controller agreement) must be in place before reusing the data.[55]

15.4.3 *Transparency, an Example of a Privacy-Enhancing Organizational Measure*

The importance of transparency cannot be overestimated. As indicated by the EDPB in the adopted Article 29 Working Party Guidelines on transparency under the GDPR: "*transparency is a long established feature [...] engendering trust in the processes which affect the citizen by enabling them to understand, and if necessary, challenge those processed.*"[56] The transparency principle entails an overarching obligation to ensure fairness and accountability. Therefore, data controllers must provide clear information that allows data subjects to have correct expectations.

The transparency obligation is a general obligation isolated from any information obligations that may follow from informed consent as a legal basis. Whichever legal basis is most suitable and whether it concerns primary or secondary use, the data controller is responsible for providing transparent information actively (following Articles 13 and 14 GDPR) and passively (following a data subject access request under Article 15 GDPR). This includes the obligation to inform about (intentions to) reuse.[57]

Today data controllers often focus on the availability of general information on websites, in brochures, and in privacy notices, to comply with their transparency obligation. Unfortunately, these general information channels often prove insufficient to enable data subjects to really understand for which purposes and by whom data *about them* is used. They feel insufficiently empowered to hold the data controller accountable or to exercise control over their personal data. If other patients' rights such as the right not to know, can be respected, wouldn't it make sense to create personalized overviews of secondary data processing operations in an era where personalization is a buzzword? These overviews could be provided through consumer interfaces such as client accounts, personalized profiles, or billing platforms. In healthcare, it is no longer uncommon for healthcare providers to provide patients with a direct view of their medical records through an app or website. A patient-tailored overview of secondary use could be included in this patient viewer.

[55] Following respectively Article 28 or 26 GDPR.

[56] Article 29 Working Party, "Guidelines on Transparency under Regulation 2016/679" (2018) WP260 rev. 01 4.

[57] Article 13, 3. and Article 14, 4. explicitly foresee the obligation to inform about the processing of data for "a purpose other than that for which the data were obtained," while Article 13, 1. (e), Article 14, 1. (e) and Article 15, 1. (c) oblige controllers to inform the data subject about "recipients or categories of recipients to whom the personal data have been disclosed." See also EDPB, Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research, adopted February 2, 2021, 9.

As a side note, it must be mentioned that the EDPB announced further clarifications on the scope of the exceptions to the obligation to actively inform data subjects individually.[58] Article 14, 5. (b) of the GDPR acknowledges that when data were not obtained directly from the data subject, it may occur that "*the provision of information proves impossible or would involve a disproportionate effort*."[59] In earlier interpretations, the limitations of this exception were stressed explaining It that the data controller must demonstrate either impossibility or a disproportionate effort. In demonstrating why Article 14, 5. (b) should apply, data controllers must mention the factors that prevent them from providing the information and illustrate the impact and effects for the data subject when not provided with the information in the case of disproportionate effort.[60]

15.5 CONCLUSIONS

In Belgium, the seven university hospitals developed a methodology to see to their responsibility as the guardian of health-related data. While not exclusively intended to address the requests for the reuse of data for AI, it was noted that requests for secondary use have an "*increasing variability in purpose, scope and nature*" and include "*the support of evidence-based medicine and value-driven healthcare strategies, the development of medical devices, including those relying on machine learning and artificial intelligence.*"[61] The initiative of the Belgian university hospitals is just one illustration of the need for legal and ethical guidelines on the use of health-related data for AI. As indicated by the Belgian hospitals, the goal is "*to keep hospitals and healthcare practitioners from accepting illegitimate proposals [for the secondary use of real-world data].*"[62] The same intention can also be found in regulatory initiatives such as the Act on AI and the Proposal for a Regulation on the European Health Data Space.

Any initiative for future regulations or guidelines will build on the provisions already included in Europe's General Data Protection Regulation. Even with the need to clarify specific provisions and harmonize various interpretations of these provisions, the GDPR lays down the principles that must be considered when collecting data for AI.

[58] EDPB, "Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research, adopted February 2, 2021, 9."

[59] Article 29 Working Party (n 56) 30–31

[60] Article 29 Working Party Guidelines on transparency under Regulation 2016/679, as adopted on 29, November 2017 and last revised and adopted on April 11, 2018 and as adopted by the EDPB during its first plenary meeting on May 25, 2018, 28–29.

[61] Raad Universitaire Ziekenhuizen België, "Common position establishing a framework for secondary use of real-world data (routinely) collected in hospitals," adopted July 7, 2022, available online at: www.univ-hospitals.be/common-position-establishing-a-framework-for-secondary-use-of-real-world-data-routinely-collected-in-hospitals/

[62] Ibid.

Within the healthcare domain, the data necessary for the development and use of AI are unlikely to be qualified as anonymous data. Most likely, they will fall under the definition of pseudonymized data as provided in Article 4 (5) of the GDPR. Notwithstanding the general prohibition to process health-related data pursuant to Article 9, 1. of the GDPR, the processing of health-related data can be justified when the interests of society or other parties prevail over the interests of the individual data subject or when informed consent reflects the data subject's wish. Additionally, all other data protection principles, such as transparency, must be respected.

Despite the numerous current and future challenges arising from regulatory instruments applicable to data custodians and data users and ongoing ethical discussions, the key message should not be that we should refrain from using health-related data for AI. Rather, we should never forget that behind the data are flesh-and-blood people who deserve protection through the implementation of organizational and technical measures.