Journal of Technology
Law & Policy

Volume XX – 2019-2020

ISSN 2164-800X (online)

DOI 10.5195/tlp.2020.235

http://tlp.law.pitt.edu

Privacy, Risk, Anonymization and Data Sharing in the Internet of Health Things

Liane Colonna

Abstract

This paper explores a specific risk-mitigation strategy to reduce privacy concerns in the Internet of Health Things (IoHT): data anonymization. It contributes to the current academic debate surrounding the role of anonymization in the IoHT by evaluating how data controllers can balance privacy risks against the quality of output data and select the appropriate privacy model that achieves the aims underlying the concept of Privacy by Design. It sets forth several approaches for identifying the risk of re-identification in the IoHT as well as explores the potential for synthetic data generation to be used as an alternative method to anonymization for data sharing.

Privacy, Risk, Anonymization and Data Sharing in the Internet of Health Things

Liane Colonna[*]

I. Introduction[1]

The Internet of Health Things promises to revolutionize healthcare in terms of improving individual health and wellness, home care, residential care, and acute care. The idea is to rely on a multiplicity of sensor-based systems to predict and prevent disease, provide personalized healthcare, offer wellness monitoring, and give support to formal and informal caregivers.[2] The IoHT has been made possible because of the concurrent development of sensor technology, data transmission, and storage technology together with new search and artificial-intelligence techniques.[3]

The IoHT are able to collect data about the most intimate details of an individual's life. For example, data about an individual's sleep patterns, health and fitness activities, social life, media consumption, and mobility patterns. These data precisely captured from any number of smart devices can create new insights about an individual concerning their specific behaviors. It can also give rise to insights about groups and society as a whole. Furthermore, it is possible, through the practice of sensor fusion, to combine information collected by different sensing devices found in this environment to create even greater insight about an individual or group of individuals. For example, data about heart rate and respiration can be combined to infer whether an individual abuses certain substances.[4]

[*] Liane Colonna is a post-doctoral fellow at the Swedish Law and Informatics Research Institute (IRI) located at Stockholm University. Her current research concerns the development of privacy aware lifelogging technologies for the frail and sick (https://paal-project.eu/).

[1] The support of the JPI More Years, Better Lives and the Swedish Research Council for Health, Working Life, and Welfare (grant number is 2017-02302) is gratefully acknowledged.

[2] Ahmad Akl et al., *Autonomous Unobtrusive Detection of Mild Cognitive Impairment in Older Adults*, 62 IEEE Transactions on Biomedical Engineering 1383 (2015); *see also* Alexandros A. Chaaroaui et al., *A Vision-Based System for Intelligent Monitoring: Human Behaviour Analysis and Privacy by Context*, 14 Sensors 8895 (2014).

[3] Cathal Gurrin et al., Digital Enlightenment Yearbook: Social Networks and Social Machines, Surveillance and Empowerment 49–73 (Kieron O'Hara et al. eds., 2014).

[4] Scott R. Peppet, *Regulating the Internet of Things: First Steps Toward Managing Discrimination, Privacy, Security, and Consent*, 93 Tex. L. Rev. 85, 145 (2014).

Journal of Technology Law & Policy

Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.235 ● http://tlp.law.pitt.edu

Within the context of the IoHT, utilizing a risk-based approach offers a strategy for data controllers to assess and manage the risks related to their data processing activities: it is a key tool to protect privacy insofar as it can identify serious risks to the right and the appropriate measures for mitigating them. Because the risk-based approach is flexible, it is particularly well suited to handle the challenges of new and emerging technologies like the IoHT. It has been noted: "The risk-based approach may provide a solution to the current data protection practices such as big data analytics or Internet of Things where the traditional compliance-based approach does not work."[5]

A core governance measure that a data controller can rely on when controlling risk includes Privacy by Design.[6] The concept entails "the philosophy and approach of embedding privacy into the design specifications of various technologies."[7] It requires the coordination of multiple stakeholders ranging from lawyers to engineers to psychologists to end-users and asks them to share responsibility for achieving privacy.[8] Privacy by Design strategies seek to respond to legal requirements and range in a degree of complexity, including both technical and organizational techniques.[9] This paper, however, will only address technical approaches, which exist on a broad spectrum. On one side of it, there is legal automation, the automatic execution of legal rules and other norms, defined in the system, in order to support proper processing of personal data.[10] On the other end of the spectrum, there are techniques like anonymization. Because hard wiring law into computer code is very difficult,[11] anonymization has been selected as the focus of this paper.

This paper will specifically examine how developers of IoHT applications can publicize their data sets for educational or research purposes consistent with laws

[5] Milda Macenaite, *The "Riskification" of European Data Protection Law through a two-fold Shift*, 8 EUROPEAN J. OF RISK REG. 506 (2017).

[6] Alessandro Spina, *A Regulatory* Mariage de Figaro*: Risk Regulation, Data Protection, and Data Ethics*, 8 EUROPEAN J. OF RISK REG. 88 (2017).

[7] Ann Cavoukian, *Privacy by Design, The 7 Foundational Principles, Implementation and Mapping of Fair Information Practices* (May 2010), http://www.ontla.on.ca/library/repository/mon/24005/301946.pdf.

[8] *See generally* Woodrow Hartzog & Frederic Stutzman, *Obscurity by Design*, 88 WASH. L. REV. 385 (2013).

[9] Dag Wiese Schartum, *Making Privacy by Design Operative*, 24 INT'L J.L. & INFO. TECH. 151 (2016).

[10] *Id.*; *see also* Cecilia Magnusson Sjöberg, *Legal Automation: AI and Law Revisited*, LEGAL TECH, SMART CONTRACTS AND BLOCKCHAIN 173 (Marcelo Corrales et al. eds., 2019).

[11] Bert-Jaap Koops & Ronald E. Leenes, *Privacy Regulation Cannot Be Hardcoded: A Critical Comment on the 'Privacy by Design' Provision in Data-Protection Law*, 28 INT'L REV. L. COMPUTERS & TECH. 159–71 (2014).

like the General Data Protection Regulation (GDPR). Here, data anonymization offers a risk-mitigation strategy to reduce privacy concerns where there is a general anxiety that sharing sensitive health data can lead to harmful effects if the data is leaked. The challenge is, however, to determine "what level of risk of re-identification is acceptable in order to deliver the potential benefits of data sharing."[12] This challenge is exacerbated by the fact that, it will be often be the case that the more personal data is concealed, the more likely such concealment will affect the utility of the data set.

II. The Internet of Health Things

At the turn of the twenty-first century, the Internet of Things (IoT) exploded, offering incredible, new opportunities to transform everyday objects into "smart" devices that have the potential to revolutionize society and business alike. In the words of the U.S. Federal Trade Commission, the IoT denotes "an interconnected environment where all manner of objects have a digital presence and the ability to communicate with other objects and people."[13] The kinds of devices that can be connected are virtually limitless, ranging from smartphones to utensils to vehicles to entire buildings.[14]

One particular sector where consumers, organizations, and governments have sought to install IoT devices is that of health and, indeed, smartphones, wearable devices, and mobile health apps are transforming the industry. Four central applications of IoHT devices are remote patient diagnostics and monitoring (e.g., sensors that can monitor medication intake), telehealth where doctors are accessible outside of their offices, behavioral modification (e.g., a device that can help an individual change her behavior and adopt a healthier lifestyle), and support for formal and informal caregivers.[15] These technologies have the potential to change

[12] European Medicines Agency, *Data Anonymisation—A Key Enabler for Clinical Data Sharing* (Dec. 4, 2018), https://www.ema.europa.eu/en/documents/report/report-data-anonymisation-key-enabler-clinical-data-sharing_en.pdf.

[13] FTC, FTC Staff Report: Internet of Things: Privacy and Security in a Connected World 17 (2015), https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf.

[14] O. Vermesan & P. Friess, *Internet of Things—From Research and Innovation to Market Deployment*, River Publishers Series in Communication 3–70 (2014), http://www.internet-of-things-research.eu/pdf/IERC_Cluster_Book_2014_Ch.3_SRIA_WEB.pdf.

[15] David H. Roman & Kyle D. Conlee, *The Digital Revolution Comes to US Healthcare*, 5 Internet of Things 1, 15 (June 29, 2015), https://www.wur.nl/upload_mm/0/f/3/8fe8684c-2a84-4965-9dce-550584aae48c_Internet%20of%20Things%205%20-%20Digital%20Revolution%20Comes%20to%20US%20Healtcare.pdf; Francisco Flórez-Revuelta, Alex Mihailidis, Martina Ziefle, Liane Colonna, Susanna Spinsante, *Privacy-Aware and Acceptable Lifelogging Services for Older and Frail People: the*

the way health care is provided in an effort to assist healthy people to stay healthy longer as well as to support the frail and sick by enabling things like better patient monitoring, drug management, and early medical interventions.[16]

As noted at the outset, the IoHT was borne from the concurrent development of sensor technology, data transmission and storage technologies, and new search and artificial intelligence techniques.[17] Health devices typically have physical hardware that interfaces with (or in!) the human body and are designed to pervasively connect to the Internet.[18] Physiological parameters such as heart rate, respiration, or blood oxygen saturation can be collected alongside behavioral parameters linked to health and well-being.[19] These types of data are collected by the hardware, often in real time, and then, at least generally, are transferred over the open Internet to a cloud service provider, although fog computing is growing as an alternative.[20] Machine learning and AI techniques are relied upon to process these data to make decisions concerning, for example, a change of behavior. There are also associated mobile applications as well as Web applications involved, which are sometimes necessary to review data.[21] In short, IoHT devices have a "multidimensional nature," combining sensors, big data solutions, distributed data storage, machine learning, and AI.[22]

III. The Benefits and Challenges to Data Sharing in the IoHT

Data is at the core of the development of the IoHT and much of its potential depends on the ability to share large amounts of health data to develop and train AI

PAAL Project, 2018 IEEE 8th International Conference on Consumer Electronics—Berlin (ICCE-Berlin 2018).

[16] Syagnik Banerjee, Thomas Hemphill & Phil Longstreet, Wearable Devices and Healthcare: Data Sharing and Privacy, 34 The Information Society 49 (2018), https://www.tandfonline.com/doi/full/10.1080/01972243.2017.1391912.

[17] Gurrin et al., *supra* note 3.

[18] H. Michael O'Brien, *The Inevitable Collision with Product Liability*, 19 J. Internet L., no. 12 (2016).

[19] Brent Mittelstadt et al., *The Ethical Implications of Personal Health Monitoring*, 5 Int'l. J. of Technoethics 37–60 (2014).

[20] Liane Colonna, *In Search of Data Protection's Holy Grail Applying Privacy by Design to Lifelogging Technologies*, Data Protection and Privacy: Data Protection and Democracy (Ronald Leenes et al. eds., 2019).

[21] O'Brien, *supra* note 18.

[22] Charlotte A. Tschider, *Regulating the Internet of Things: Discrimination, Privacy, and Cybersecurity in the Artificial Intelligence Age*, 96 Denv. L. Rev. 87, 143 (2018) (referring to Arshdeep Bahga & Vijay Madisetti, Internet of Things: A Hands-On Approach 21 (2014)).

and machine-learning algorithms. Accessing and securing data is imperative to the success of any AI or machine-learning system because huge data sets are what fuel intelligent algorithms, allowing the software to learn from patterns or features in the data. There can be little doubt that the ability to freely transfer data sets would have a positive impact on the quality of AI/ML of the future. As Hasnain et al. puts it: "Access to (biomedical) data is increasingly important to enable data driven science in the research community."[23]

While it is clear that data access and sharing are crucial to data-driven innovation, there are nevertheless many barriers to data discovery and reuse that can limit its potential.[24] First, while there are huge amounts of data available online, the problem is first to actually *find* an appropriate data set.[25] This is because data often lack sufficiently rich metadata and may not be registered or indexed in a searchable resource that is known and accessible to potential users.[26] Accessibility is a second issue as the precise conditions by which data are accessible are not always clear.[27] In other words, there is often a lack of clarity and transparency surrounding the conditions governing access and reuse of data.[28] Even where findability and accessibility are not issues, it may be the case that data is not interoperable from either a technical or legal perspective.[29]

In order for data to be reusable it must be accurate and contain relevant attributes. However, open source data sets often contain errors, bias or

[23] Ali Hasnain & Dietrich Rebholz-Schuhmann, *Assessing FAIR Data Principles Against the 5-Star Open Data Principles*, THE SEMANTIC WEB: ESWC 2018 SATELLITE EVENTS 469, 469 (Aldo Gangemi et al. eds, 2018).

[24] Mark D. Wilkinson et al., *The FAIR Guiding Principles for Scientific Data Management and Stewardship*, SCI. DATA 3 (2016); *see also* OECD, ENHANCING ACCESS TO AND SHARING OF DATA: RECONCILING RISKS AND BENEFITS FOR DATA RE-USE ACROSS SOCIETIES (2019) (laying out how and why access and sharing of data remain below their potential).

[25] Wilkinson et al., *supra* note 24.

[26] *Final Report and Action Plan from the European Commission Expert Group on FAIR Data: Turning FAIR into Reality* (2018), https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf.

[27] *Id.*

[28] *Id.*

[29] *Id.*; *but see* REPORT ON THE EUROPEAN COMMISSION'S WORKSHOPS ON COMMON EUROPEAN DATA SPACES (2019), https://ec.europa.eu/digital-single-market/en/news/report-european-commissions-workshops-common-european-data-spaces (wherein the European Union is trying to create common European dataspaces that specifically addresses concerns about e.g. data interoperability and data quality).

misinformation that can negatively affect the AI and ML systems using it.[30] As such, datasets must be cleaned to ensure that any insights developed from it are based on high-quality, accurate and well-structured examples. Furthermore, many AI and machine-learning algorithms require curated data to train the algorithms properly. But herein lies the rub: high-quality labeled data sets are frequently subject to license restrictions and other kinds of paywalls or regulatory hurdles.[31] The Organisation for Economic Co-operation and Development (OECD) explains that businesses in the data ecosystem use a diversity of revenue models including: freemium, advertisement, subscription, usage fees, selling of goods, selling of services, licensing, and commission fees.[32]

It is also important to mention that there are strict privacy and security rules concerning the collection and processing of health data, and the devices that handle them.[33] For example, the GDPR recognizes "health data" as a "special category of personal data" which in principle should not be processed.[34] Additional legal concerns include matters like intellectual property protection, contract compliance, and avoidance of anti-competitive behavior.[35]

Pavón and Goumas explain: "Some of the biggest obstacles fueling AI and ML with the best data are not technological, but legal in nature."[36] Indeed, many research institutions simply lack the time, funding, and human resources to collect high-quality (labeled) data from a diverse population in a legally compliant manner.[37] As a result, data sets about, for example, smartwatches, are incredibly scarce.

[30] Pedro Pavón & Alexandra Goumas, Data: *The Fuel Powering Artificial Intelligence and Machine Learning*, BUSINESSLAWTODAY.ORG, https://businesslawtoday.org/2017/12/data-the-fuel-powering-artificial-intelligence-and-machine-learning/ (Dec. 14, 2017).

[31] *Id.*

[32] OECD, *supra* note 24.

[33] Brent Mittelstadt, *Designing the Health-Related Internet of Things: Ethical Principles and Guidelines*, 8 INFORMATION, Dec. 2017.

[34] OECD, *supra* note 24.

[35] *Id.*

[36] Pavón & Goumas, *supra* note 30, at 1–4.

[37] *See* Wilkinson et al., *supra* note 24. Note, however, that recently, a diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers have proposed the FAIR Guiding Principles which serve as a model for data openness.

IV. Anonymization in the IoHT

A Navigating the Semantic Muddle

Anonymization is a nebulous term and it is often conflated with other, closely related concepts, like de-identification, pseudonymization, encryption, and unlikability. The problem with the free flow and ad hoc use of these terms is that confusion arises because not all of the words mean exactly the same thing: important perceptions about the technology are lost when the different meanings are blurred together. The lack of clarity with respect to what precisely "anonymization" entails, creates a rocky foundation for the law's application and the possibility for legal loopholes. Set forth below is a taxonomy of terminology relevant in this context and a clarification of terms.

While the GDPR does not define anonymization, the International Organization for Standardization (ISO) defines it as a "process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party."[38] A personal identifier can be understood as "a specific piece of information, holding a privileged and close relationship with an individual, which allows for the identification, direct or indirect, of this individual."[39] An example of an identifier is a name, email address, picture of an individual as well as a specific device identifier like a MAC address.[40] There are a multiplicity of different methods that can be applied to achieve anonymization such as k-anonymity, suppression, aggregation, perturbation, or generalization.[41]

[38] Nat'l Inst. Stds. & Tech., ISO/TS 25237:2017, Health Informatics—Pseudonymization (2017); *see also* Nat'l Inst. Stds. & Tech., Glossary, *Anonymization* (defining anonymization as a "process that removes the association between the identifying dataset and the data subject.").

[39] Eur. Network & Info. Sec. Agency, Recommendations on Shaping Technology According to GDPR Provisions: An Overview on Data Pseudonymisation (2019), https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions.

[40] *Id.*

[41] Nicola Jentzsch, Competition and Data Protection Policies in the Era of Big Data: Privacy Guarantees as Policy Tools, https://fpf.org/wp-content/uploads/2016/11/Jentzsch_Ident_Workshop_Paper_2016_V8_FINAL-I.pdf.

Data anonymization is often considered synonymous with de-identification.[42] In fact, the Article 29 Working Party[43] defines anonymization as "a technique applied to personal data in order to achieve irreversible de-identification."[44] Here, however, it is important to note that the key word is "irreversible" as, according to the ISO, "any process of reducing the association between a set of identifying data and the data subject" is considered de-identification.[45] In order to achieve "anonymization," the process must be irreversible, at least according to the Article 29 Working Party.

Abandoning the long-held, binary distinction between anonymized data and personal data, the GDPR introduces a third category of data: pseudonymized data.[46] ENISA explains, "Broadly speaking, pseudonymisation aims at protecting personal data by hiding the identity of individuals in a dataset, e.g. by replacing one or more personal data identifiers with the so-called pseudonyms (and appropriately protecting the link between the pseudonyms and the initial identifiers)."[47] From a

[42] *See* Peerapong Vanichayavisalsakul & Krerk Piromsopa, *An Evaluation of Anonymized Models and Ensemble Classifier* (2018) (stating "Data Anonymization, also known as de-identification, is a process of concealing personal information by making each record indistinguishable from the other records.").

[43] The Article 29 Working Party (art. 29 WP) is the independent European working party that dealt with issues relating to the protection of privacy and personal data until 25 May 2018 (entry into application of the GDPR). Upon enactment of the GDPR, the Article 29 Working Party has been replaced by the European Data Protection Board.

[44] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymization Techniques*, at 3, 10, 0829/14/EN WP216 (Apr. 10, 2014), https://www.pdpjournals.com/docs/88197.pdf.

[45] NAT'L INST. STDS. & TECH., *supra* note 38; *see also* Ontario Legislative Library, Office of the Information and Privacy Commissioner, *De-identification Guidelines for Structured Data* ((2016) (explaining, "'De-identification' broadly denotes the process of removing personal information from a record or data set and, on the flip side, re-identification can be defined as "any process that re-establishes the link between identifiable information and an individual."); *see* EUROPEAN MEDICINES AGENCY, *supra* note 12 (explaining, de-identification is "a term used in the United States to describe the process used to prevent personal identifiers in a dataset from being connected with information so as to allow identification of an individual. De-identification does not describe a single technique but rather a collection of approaches (e.g. suppression, averaging, generalization, perturbation, swapping) that can be applied to data with different levels of effectiveness. De-identified data may still allow a data generator or a trusted party to retain the means (e.g. a code, algorithm or pseudonym) to identify the person.").

[46] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) art. 4(5) [hereinafter GDPR].

[47] European Union Agency for Cybersecurity (ENISA), *Recommendations on Shaping Technology According to GDPR Provisions: An Overview on Data Pseudonymisation* (2019), https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions (also stating, "In broad terms, pseudonymisation refers to the process of de-associating a data subject's identity from the personal data being processed for that data subject. Typically, such a process may be performed by replacing one or more personal identifiers, i.e. pieces of information that can allow identification (such

more technical perspective, the ISO defines pseudonymisation as a "particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms."[48]

A legal definition of "pseudonymization" is set forth in the GDPR, which defines it as:

> [T]he processing of personal data in such a manner that the personal data 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 organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.[49]

In other words, according to the GDPR, the process of pseudonymization involves the conversion of data about an identified person into data about a merely "identifiable" person with the condition that the additional data necessary for re-identification are kept safely inaccessible for the users of "pseudonymized data."[50] Here, it is important to note that the GDPR definition of pseudonymization provides a stricter framework for implementation than the ISO definition to the extent that it covers not just the protection of "the real world person identity" but also the protection of indirect identifiers relating to a data subject like online unique identifiers.[51]

Recital 26 of the GDPR makes clear that "Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person," meaning pseudonymous data is not exempt from the GDPR.[52] That

as e.g. name, email address, social security number, etc.), relating to a data subject with the so-called pseudonyms, such as a randomly generated values.").

[48] NAT'L INST. STDS. & TECH., ISO/TS 25237:2017, HEALTH INFORMATICS—PSEUDONYMIZATION (2017); *see also* NAT'L INST. STDS. & TECH., ISO/IEC 20889:2018, PRIVACY ENHANCING DATA DE-IDENTIFICATION TERMINOLOGY AND CLASSIFICATION TECHNIQUES (2018) (explaining that another technical definition of pseudonymisation is provided by the ISO/IEC 20889:2018 standard as a "de-identification technique that replaces an identifier (or identifiers) for a data principal with a pseudonym in order to hide the identity of that data principal.").

[49] GDPR art. 4(5).

[50] Waltraut Kotschy & Ludwig Boltzmann, *The New General Data Protection Regulation—Is There Sufficient Pay-Off for Taking the Trouble to Anonymize or Pseudonymize Data?*, https://fpf.org/wp-content/uploads/2016/11/Kotschy-paper-on-pseudonymisation.pdf.

[51] EUR. NETWORK & INFO. SEC. AGENCY, *supra* note 39.

[52] GDPR R. 26.

said, the law offers several key incentives to apply pseudonymization when processing personal data such as the explicit satisfaction of the "data protection by design" requirement[53] and the chance to repurpose data for another compatible use.[54] To put it differently, while the GDPR applies, the application of the rules set forth in the law is more flexible since ostensibly the risks at stake for the individual with regard to the processing of her indirectly identifiable information will most often be low, at least compared to the processing of her directly identifiable information.[55]

The Article 29 Working Party Opinion on Anonymization Techniques further clarifies that pseudonymization is "not a method of anonymization" but "merely reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure."[56] The Opinion describes the five most common pseudonymization techniques: (1) encryption with a secret key; (2) hash function; (3) keyed-hash function with stored key; (4) deterministic encryption or keyed-hash function with deletion of the key; and (5) tokenization.[57] Here, a question arises concerning the relationship between pseudonymization and encryption.

Pursuant to the GDPR, encryption seems to be included within the ambit of pseudonymization so long as the encryption "key" is kept separate and secure, and data administrators implement appropriate measures to prevent the "unauthorized reversal of pseudonymization."[58] That said, in Article 6(4)(e) the law refers to "appropriate safeguards, which may include encryption or pseudonymisation."[59] The Article 29 Working Party, as noted above, makes clear that encryption is one

[53] GDPR art. 25(1); *see also* Elizabeth A. Brasher, *Addressing the Failure of Anonymization: Guidance from the European Union's General Data Protection Regulation*, 2018 COLUM. BUS. L. REV. 209, 252 (2018).

[54] GDPR art. 6(3)(a).

[55] Khaled El Emam & Cecilia Álvarez, *A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques*, 5 INT'L DATA PRIVACY L. 73 (2015).

[56] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymization Techniques*, at 3, 10, 0829/14/EN WP216 (Apr. 10, 2014), https://www.pdpjournals.com/docs/88197.pdf.

[57] *Id.* at 20–21.

[58] GDPR R. 75; *compare* GDPR art. 6(4)(e)) (referring to "appropriate safeguards, which may include encryption *or* pseudonymisation"), *with GDPR* art. 32(1)(a) (referring to appropriate technical measures like "the pseudonymisation *and* encryption of personal data"); *see also* GDPR R. 29 (stating, "In order to create incentives to apply pseudonymisation when processing personal data, measures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organisational measures necessary to ensure, for the processing concerned, that this Regulation is implemented, and that additional information for attributing the personal data to a specific data subject is kept separately. The controller processing the personal data should indicate the authorised persons within the same controller.").

[59] *Cf.* GDPR art. 32(1)(a) (referring to appropriate technical measures like "the pseudonymisation *and* encryption of personal data.") (emphasis added).

technique of pseudonymization, making the reference to both terms in the GDPR appear to be redundant. ENISA echoes this sentiment and explains in detail how encryption as a pseudonymisation technique works from a technical perspective.[60]

Unlinkability is another related term. Brost explains, "Unlinkability ensures that privacy-relevant data cannot be linked across privacy domains or be used for a different purpose than originally intended. This can be achieved by, e.g., early erasure, anonymization, or pseudonymization."[61] Essentially, unlinkability is about preventing an attacker from identifying the link between two or more items in a system.[62] In other words, two or more items of interest, such as senders, recipients, and messages, cannot be connected from the perspective of an attacker.

B. The Academic Debate Surrounding Anonymization

Up to now, the academic debate over anonymization has been lively and far from being set, with the literature providing both critical perspectives as well as more positive accounts. Sometimes the debate is framed as formalist (those more interested in mathematical rigor) versus pragmatist (those more interested in finding practical solutions).[63] On the critical side, Ohm has largely deemed anonymization a failure, at least as an exclusive means to prevent privacy harm.[64] Ohm's central point is that "anonymized" data is not very "anonymizing" in the Age of Big Data because it is increasingly possible to link an individual through "reidentification."[65] To put it differently, modern database technologies now make it possible to link anonymized data with outside information through common data elements to reconstruct personally identifying profiles.[66] Ohm concludes (already in 2010!), "researchers

[60] EUR. NETWORK & INFO. SEC. AGENCY, RECOMMENDATIONS ON SHAPING TECHNOLOGY ACCORDING TO GDPR PROVISIONS: AN OVERVIEW ON DATA PSEUDONYMISATION (2019), https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions.

[61] G.S. BROST & M. HOFFMAN, REQUIREMENTS FOR DIGITAL HEALTH 133–54 (S. Fricker, C. Thümmler & A. Gavras eds., 2014).

[62] M. BRUSÓ, K. CHATZIKOKOLAKIS, S. ETALLE & J. DEN HARTOG, TRUSTWORTHY GLOBAL COMPUTING 129–44 (C. Palamidessi & M.D. Ryan eds., 2013) (stating, "Unlinkability is a privacy property which holds when an attacker cannot identify the link between two or more items in a system.").

[63] *See* Ira Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703, 714–17 (2016) (discussing the debate between formalists and pragmatists).

[64] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010).

[65] *Id.* at 1703–04.

[66] *Id.*

have learned more than enough already for us to reject anonymization as a privacy-providing panacea."[67]

Instead of viewing anonymization as "a privacy-providing panacea,"[68] Ohm suggests weighing "the benefits of unfettered information flow against the costs of privacy harms" and incorporating risk assessment strategies.[69] His idea is to methodologically consider the likelihood of harm and to weigh that risk against the benefit of unfettered information flow. He concludes that where the harm is considered to be greater than the benefit then regulation should step in, focusing on specific contexts.[70] He further contends that the distinction between PII and non-PII should be abandoned in favor of a risk-based approach that takes into account (1) the data-handling techniques used by the data controller, (2) the nature of the information release, with the public releases being subject to stricter scrutiny than private disclosures of data, (3) the quantity of data involved, (4) the motives that individuals might have to reidentify the data, and (5) the trust that exists in people or institutions handling the data.[71]

In response to Ohm, Schwartz and Solove reject the complete abandonment of anonymization approaches based on removal of personally identifiable information, and instead argue for refinement of the concept of PII.[72] They suggest that a more flexible approach to PII is needed where the "risk of identification" is tracked along a spectrum where on one end there exists no risk of identification (e.g. non-identifiable data) and, at the other end, an individual is fully identified (e.g. identifiable data). In the middle of the spectrum, the risk of identification is moderate (e.g. identifiable, but not identified).[73] They suggest that all Fair Information Principles (FIPs) should apply to information that refers to an identified person. However, if data only refers to an identifiable person then only data quality, transparency, and security should apply. Essentially, Schwartz and Solove argue for

[67] *Id.* at 1716.

[68] *Id.*

[69] *Id.*

[70] *Id.*

[71] *Id.* at 1765–68.

[72] Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814 (2011); *see also* Paul M. Schwartz & Daniel J. Solove, *Reconciling Personal Information in the United States and European Union*, 102 CAL. L. REV. 877, 907 (2014).

[73] Schwartz & Solove, *supra* note 72; *see also* Schwartz & Solove *supra* note 72.

a model where the "realistic risk of identification" is used as a benchmark against which to introduce specific measures that prevent re-identification.[74]

Bambauer (Yakowitz) also criticizes Ohm suggesting that he overstates the risk of re-identification and under-appreciates the value of public data releases.[75] That is, she contends that the immense benefits of information flowing from the data commons outweigh the deanonymization risk.[76] Ultimately, she concludes that intentional reidentification of data should be criminalized and that a safe harbor must be given to data that has been anonymized using relatively straightforward techniques in order to facilitate access to research data.[77]

In an effort to move past "the anonymization stalemate" between formalist like Ohm and pragmatist like Bambauer, Rubinstein and Hartzog suggest a process-based approach for minimizing the risk of reidentification and sensitive attribute disclosure.[78] Accepting that perfect anonymization is probably impossible, they contend data controllers should follow appropriate processes for minimizing risk.[79] These processes could include, for example, de-identification as well as other legal and administrative tools. This approach is focused on those that release data and demands that they take contextually appropriate solutions to ensure that data is shared in a responsible way.

C. Data Utility and the Problem with Anonymization in the IoHT

When data is anonymized then it is often less interesting for reuse than raw personal data or pseudonymous data.[80] Sometimes this phenomenon is described as "a negative correlation" between data privacy, on the one side, and data utility, on the other.[81] Zevenbergen summarizes:

> The utility and privacy of data are generally directly and inversely related. For many datasets, it has proven

[74] Schwartz & Solove, *supra* note 72; *see also* Schwartz & Solove, *supra* note 72.

[75]Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1 (2011).

[76] *Id.* at 4.

[77] *See id.*

[78] Ira Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703, 707–08 (2016).

[79] *Id.* at 706.

[80] Frederik Zuiderveen Borgesius, Jonathan Gray & Mireille van Eechoud, *Open Data, Privacy, and Fair Information Principles: Towards A Balancing Framework*, 30 BERKELEY TECH. L.J. 2073, 2119–20 (2015).

[81] Elizabeth A. Brasher, *Addressing the Failure of Anonymization: Guidance from the European Union's General Data Protection Regulation*, 2018 COLUM. BUS. L. REV. 209, 217 (2018).

> difficult—if not impossible—to increase data subjects' privacy without concurrently decreasing the overall utility of the dataset. Small privacy gains are generally achieved by far-reaching decreases in data utility. A small increase in data utility often requires much more personal information to be revealed.[82]

In short, there can never be zero risk in a useful data set: there will always be some kind of residual risk in all useful data.[83]

The negative correlation between privacy and utility is especially highlighted when it comes to the IoHT where personal identifiers are critically important to unlocking the potential power of the dataset.[84] Behavioral monitoring technology collects information about the user's behavior (e.g. movements, actions and sounds) and physiological monitoring systems tract and record patient physiological data (e.g. heart rate, breathing rates, blood pressure, electrocardiogram (ECG)).[85] The ability to collect and process such personal data is what makes it possible to, for example, raise alarm in case of detected abnormalities or diagnose a particular illness.[86]

Utilizing sensitive medical data is highly necessary to support AI applications, whether they support diagnosis or directly deliver instructions to a device,[87] and unfortunately, it is often the case that the efficiency of the AI is reduced where these data are de-identified.[88] Here, Inan et al. notes that, "a key question" is "whether and

[82] Borgesius et al., *supra* note 80; *see further* Sophie Stalla-Bourdillon & Alison Knight, *Anonymous Data v. Personal Data—A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data*, 34 WIS. INT'L L.J. 284, 285 (2016) (explaining, ". . . the value or knowledge that can be gained from analyzing datasets (particularly using automatic algorithmic software) is maximized by virtue of finding patterns, basically linking relationships between data points. Anonymization, by contrast, aims to delink such data point relationships where they relate to informational knowledge that can be gleaned in respect of specific persons and their identities. This leaves those in charge of processing the data with a problem: how can they ensure that anonymization is conducted effectively on the data on their possession, while retaining that data's utility for potential future disclosure to, and further processing by, a third party?).

[83] Elaine Mackey, Mark Elliot & Kieron O'Hara, *The Anonymisation Decision-making Framework*, https://fpf.org/wp-content/uploads/2016/11/Mackey-Elliot-and-OHara-Anonymisation-Decision-making-Framework-v1-Oct-2016.pdf.

[84] *See, e.g.*, André Calero Valdez & Martina Ziefle, *The Users' Perspective on the Privacy-Utility Trade-Offs in Health Recommender Systems*, 121 INT'L J. HUM.-COMPUTER STUD. 108 (2019).

[85] EDUARD FOSCH-VILLARONGA, ROBOTS, HEALTHCARE, AND THE LAW: REGULATING AUTOMATION IN PERSONAL HEALTHCARE 172 (2020).

[86] *Id.*

[87] Lianne Colonna & Alex Mihailidis, *A Methodological Approach to Privacy by Design within the Context of Lifelogging Technologies*, RUTGERS COMPUTER & TECH. L.J. (forthcoming 2020).

[88] *Id.*

how anonymized data can be effectively used for data mining."[89] Unless data controllers find that the privacy benefits of anonymization are outweighed by the loss of data utility they may decide to retain data in fully identified form.[90]

V. The General Data Protection Regulation and the Risk-Based Approach

A risk-based approach is incorporated across the entire GDPR and, more specifically, one that is focused on minimizing adverse risks to the individual.[91] Consistent with the Data Protection Directive, which only addressed risk in a small number of provisions, certain data processing activities, such as the processing of sensitive data, are considered high risk and subject to specific requirements.[92] There is also a continued emphasis on risk in the provisions concerning information security.[93]

The GDPR, going further than the Data Protection Directive and placing more stringent controls on data controllers, introduces the principle of accountability in Article 5 whereby "data controllers are requested to control, in a formal and structured way, the *risks* to the rights and freedoms of data subjects arising from data processing operations."[94] There is a clear emphasis on risks in both Article 24

[89] Ali Inan et al., *Using Anonymized Data for Classification*, 429, 429–30 (2009), https://ieeexplore.ieee.org/document/4812423.

[90] KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION, 1 *passim* (2013).

[91] Center for Information Policy Leadership, Risk, *High Risk, Risk Assessments and Data Protection Impact Assessments under the GDPR*, 1, 14 *passim* (2016), https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_gdpr_project_risk_white_paper_21_december_2016.pdf (Stating that, "While there are numerous definitions and concepts of 'risk' in the privacy and data security arena, the GDPR clearly focusses on one type of risk: adverse risk to the individual."); *see also* CNIL, *Methodology for Privacy Risk Management*, 1, 13 *passim* (2012), https://www.cnil.fr/sites/default/files/typo/document/CNIL-ManagingPrivacyRisks-Methodology.pdf (Identifying individual privacy harms as physical loss of amenity, disfigurement or economic loss related to physical integrity, material (loss incurred or lost revenue with respect to an individual's assets) and moral (physical or emotional suffering, disfigurement or loss of amenity, etc.)).

[92] *See* GDPR ch. II., art. 9.

[93] GDPR ch. IV, art. 32.

[94] Spina, *supra* note 6.

(responsibility of the controller')[95] and 25(1) (Privacy by Design).[96] These Articles require that the controller take into account "the nature, scope, context and purposes of processing" as well as "the *risks* of varying likelihood and severity for the rights and freedoms of natural persons." The GDPR further requires data controllers to notify supervisory authorities and data subjects of a data breach, to conduct a data protection impact assessment, and to consult a local supervisory authority (the prior consultation) where there is "a *high risk* to the rights and freedoms" of the data subject.[97] Recital 51 clarifies that risks may be "physical, material or non-material" and pinpoints some potential harms, such as discrimination, identity theft or fraud, financial loss, and reputational damage.

Many of the new provisions in the GDPR concerning risks can be considered "meta obligations" insofar as they regulate how controllers should interpret and apply other requirements in the GDPR.[98] For example, Birnhack describes Privacy by Design, as set forth in the GDPR, as "an all-encompassing principle" that "provides an *additional* necessary safeguard to the entire regulatory basket."[99] Spina refers to the "riskification" of EU data protection law and how these provisions

[95] GDPR Chapter IV, Article 24 reads,

> Taking into account the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons, the controller shall implement appropriate technical and organisational measures to ensure and be able to demonstrate that processing is performed in accordance with this Regulation.

[96] GDPR Article 25(1) reads,

> Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

[97] GDPR ch. IV, arts. 27, 28, 33 & 34.

[98] Claudia Quelle, *The Risk Revolution in EU Data Protection Law: We Can't Have our Cake and Eat It, Too*, 1, 8, *in* DATA PROTECTION AND PRIVACY: THE AGE OF INTELLIGENT MACHINES (Ronald Leenes et al. eds., 2017).

[99] Michael Birnhack et al., *Privacy Mindset, Technological Mindset*, 55 JURIMETRICS J. 1, 13 (2014).

represent a new "model of 'enforced self-regulation' for managing technological innovation in uncertain scenarios."[100]

While the GDPR clearly supports a risk-based approach with respect to the application of certain provisions, there are concerns that "the risk based approach will lead to an uneven protection of personal data, since the level of protection afforded to each processing precisely depends on how risky they are."[101] Here, the Article 29 Working Party has noted that compliance with the core principles of the GDPR is binary: either a processing complies with a core principle, or it does not.[102] Its position is that there should be no negative impact on the rights and freedoms of an individual, not even a slightly diminished impact. Hustinx echoes this sentiment and explains that the rights of the data subject must be guaranteed regardless of the risks that are posed to data subjects.[103] That said, provisions like Article 24 (accountability) suggests that some scaling of a data controller's obligations according to the risks posed by the relevant processing operations is permissible.[104] On this point, Hustinx adds, "more detailed obligations should apply where the risk is higher and less burdensome obligations where it is lower."[105]

The question of whether the GDPR requires the management of risks or the compliance with rules and principles is subject to debate.[106] Quelle explains that there is a conceptual difficultly in the risk-based approach insofar as it asks controllers "to go beyond law," meaning to provide some kind of enhanced compliance.[107] Others, like the Centre for Information Policy Leadership discuss the risk-based approach as a means to "bridge the gap between high-level privacy

[100] Spina, *supra* note 6.

[101] Raphael Gellert, *We Have Always Managed Risks in Data Protection Law: Understanding the Similarities and Differences between the Rights-Based and the Risk-Based Approaches to Data Protection*, 2 EUR. DATA PROTECTION L. REV. 481, 483 (2016).

[102] Article 29 Working Party, *Opinion 1/98 Platform for Privacy Preferences (P3P) and the Open Profiling Standard (OPS)* (1998) (stating that the EU data protection legal framework provides for a "minimum and non-negotiable level of protection for all individuals").

[103] Peter Hustinx, *EU Data Protection Law: The Review of Directive 95/46/EC and the Proposed General Data Protection Regulation* (2014), https://edps-europa-eu.ezp.sub.su.se/sites/edp/files/publication/14-09-15_article_eui_en.pdf.

[104] Milda MaCenaite, *The "Riskification" of European Data Protection Law Through A Two-Fold Shift*, 8 EUR. J. RISK REG. 506, 524–25 (2017).

[105] Hustinx, *supra* note 103.

[106] Claudia Quelle, *Enhancing Compliance Under the General Data Protection Regulation: The Risky Upshot of the Accountability- and Risk-Based Approach*, 9 EUR. J. RISK REG. 502, 502–04 (2018).

[107] *Id.*

principles on one hand, and compliance on the ground on the other."[108] In the WP29's *Statement on the role of a risk-based approach in data protection legal frameworks*, it explains "the risk-based approach is increasingly and wrongly presented as an alternative to well-established data protection rights and principles, rather than as a scalable and proportionate approach to compliance."[109]

The relationship between risk and anonymization is especially unclear in the GDPR. Does the GDPR perceive anonymization as a risk management process? Recital 26 of the GDPR explains that the principles of data protection should not apply to anonymous information, which should be namely understood as "information which does not relate to an identified or identifiable natural person."[110] It reiterates further in the same recital that, "anonymized data" are "personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."[111] Here, a serious tension arises about whether there is a "reasonableness" standard that would allow for risk-based assessments or an "impossibility" standard, according to which data may be rendered anonymous only when it is zero (or near-zero) probability of reidentification.[112]

The Article 29 Working Party recognizes that there is an "inherent residual risk of re-identification linked to any technical-organizational measure aimed at rendering data 'anonymous' yet, in the very next line, it states, anonymization must 'achieve *irreversible* deidentification.'" This suggests the Article 29 Working Party supports an oblivion approach to anonymization. In 2012, the UK Information Commissioner's Office explained that anonymization requires that the risk of re-

[108] CENTRE FOR INFORMATION POLICY LEADERSHIP, A RISK-BASED APPROACH TO PRIVACY: IMPROVING EFFECTIVENESS IN PRACTICE 1, 4 (2014), https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/white_paper_1-a_risk_based_approach_to_privacy_improving_effectiveness_in_practice.pdf (stating, "The risk-based approach is not meant to replace or negate existing privacy regulation and data protection principles. The approach and risk framework methodology primarily aim to: a) complement the existing laws and regulations and facilitate the application of existing data protection principles and requirements; b) help implement the existing legal requirements and privacy principles in a particular context, with greater flexibility and more agility that is required in the new information age, by taking into account the risks to individuals; and c) improve the delivery of effective data protection in practice—benefitting individuals and organisations seeking more effective, systematic and demonstrable compliance.").

[109] Article 29 Data Protection Working Party, *Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks*, WP 2018 (2014), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf.

[110] GDRP art. IV, R. 26.

[111] GDPR R. 26.

[112] Ira Rubinstein, *Brussels Privacy Symposium on Identifiability: Policy and Practical Solutions for Anonymisation and Pseudonymisation: Framing the Discussion*, FUTURE OF PRIVACY FORUM (Nov. 8, 2016), https://fpf.org/wp-content/uploads/2016/11/Rubinstein_framing-paper.pdf.

identification is mitigated until it is remote (i.e. there is no requirement to remove risk entirely).[113] While ICO is currently updating these guidelines, its website explains that true anonymization under the GDPR requires that personal data is stripped of "sufficient elements that mean the individual can no longer be identified."[114] This suggests that ICO, unlike the Working Party, supports a reasonableness approach to anonymization.[115]

VI. Methods for Identifying the Risk of Re-Identification in the IoHT

At the outset, it must be made clear that there are (at least) two major challenges to any risk-based approach for data anonymization. The first challenge is the inevitable data-utility tradeoff. All known anonymization and pseudonymization techniques have their own intrinsic properties and attendant strengths and weaknesses and none of them can absolutely guarantee privacy, especially without destroying the utility of the dataset at hand. As the UK Anonymi(s)ation Network states, "[a]nonymisation is about risk management, nothing more and nothing less; accepting that there is a residual risk in all useful data inevitably puts you in the realms of balancing risk and utility."[116] To put it differently, there will always be a need to consider both the required level of privacy protection and the relevant utility of the dataset at the same time.

The second challenge with any method to evaluate the risk of re-identification is the fact that it is impossible to know whether, at some point in the future, there is a potential to link data to an identified or identifiable human being. Even if a risk assessment reveals there is a marginal chance of re-identification at the time data is disclosed, there nevertheless remains a limitless potential to link data, particularly in

[113] INFORMATION COMMISSIONER'S OFFICE (ICO), *Anonymisation: Managing Data Protection Risk (Code of Practice)*, 6 (2012), https://ico.org.uk/media/1061/anonymisation-code.pdf (stating "The DPA does not require anonymisation to be completely risk free—you must be able to mitigate the risk of identification until it is remote. If the risk of identification is reasonably likely the information should be regarded as personal data—these tests have been confirmed in binding case law from the High Court. Clearly, 100% anonymisation is the most desirable position, and in some cases this is possible, but it is not the test the DPA requires.").

[114] INFORMATION COMMISSIONER'S OFFICE, GUIDE TO THE GENERAL DATA PROTECTION REGULATION (GDPR), PRINCIPLE (E): STORAGE LIMITATION (2019), https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/storage-limitation/?q=anonymisation.

[115] INFORMATION COMMISSIONER'S OFFICE (ICO), *What Is Personal Data?*, https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/.

[116] MARK ELLIOT ET AL., THE ANONYMISATION DECISION-MAKING FRAMEWORK 5 (2016), https://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf.

the era of big data.[117] Here, important questions are: Can the residual risk of re-identification ever be defined?[118] How should re-identification be understood under the GDPR where data has been subsequently linked back to an individual in a completely unanticipated manner?[119] In short, is there any way to fully comprehend what data is already in the public domain (or may emerge in it) and how that information may correspond to a newly released, so-called "anonymous" data set?

In order to confront these realities, new approaches are needed to quantify the risk of re-identification in an effort to understand the impact of different options for data release.[120] These approaches should be flexible and principle orientated[121] and must take into consideration, at the very least, four factors: (1) the dataset and the elements it contains; (2) the robustness of an anonymization algorithm; (3) the context of data release; and (4) whether additional controls exist. Each of these will be discussed in turn.

With regard to the first factor, the dataset and the elements it contains, it is important to evaluate whether there are direct or indirect identifies (sometimes referred to as quasi identifiers) in order to determine the nature, sensitivity, and linkability of the information. Direct identifiers are "data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain."[122] They are data points that correspond directly to a person's identity, such as a name, social security number, or contact information. Indirect identifiers, on the other hand, are data that do not identify an individual in isolation but may reveal individual identities if combined with additional data points. For example, one frequently-cited study found that 87% of Americans can be uniquely identified by combining three indirect identifiers: date of birth, gender, and ZIP code.[123] In other words, while no individual can be singled out based on just a date of birth, when combined with gender and ZIP code, the lens

[117] EUROPEAN MEDICINES AGENCY, *supra* note 12.

[118] *Id.*

[119] *Id.*

[120] *Id.*

[121] Center for Information Policy Leadership, *Risk, High Risk, Risk Assessments and Data Protection Impact Assessments under the GDPR*, 1, 14 *passim* (2016), https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_gdpr_project_risk_white_paper_21_december_2016.pdf.

[122] NAT'L INST. STDS. & TECH., ISO/TS 25237:2008(EN), HEALTH INFORMATICS—PSEUDONYMIZATION (2017), https: // www.iso.org/obp/ui/#iso:std:iso:ts:25237:ed-1:v1:en.

[123] Latanya Sweeney, *Simple Demographics Often Identify People Uniquely, Carnegie Mellon University Laboratory for International Data Privacy* (2000) (discussing the likelihood of uniquely identifying individuals from basic information).

focuses on a specific identity. Both direct and indirect identifiers must be addressed during anonymization.

The size of the dataset is also important.[124] Big datasets create concerns about lack of control and transparency as well as linkability.[125] Quite simply, big data facilitates the re-identification of data subjects: the larger the dataset, the easier it is to link one data point to another and reidentify an individual.[126] Ohm explains, "Would-be reidentifiers will find it easier to match data to outside information when they can access many records indicating the personal preferences and behaviors of many people."[127]

In the IoHT context, it is also important to consider whether the dataset concerns information about rare diseases, which present a special challenge in the field of anonymization. Given the low prevalence of rare diseases, it is relatively easy to identify an individual with a rare disease from a supposedly anonymized data set, especially within a specific geographical area.[128] In other words, it is hard to prevent the re-identification of this kind of personal data, especially when such data has been linked with other data.[129]

The second factor to consider in assessing risk is the robustness of an anonymization technique. Here, the Article 29 Working Party has provided some guidance as to how a data controller should assess the robustness of an anonymization algorithm. Specifically, it considers three risks, which are essential to anonymization. The first is "singling out" or the possibility "to isolate some or all records which identify an individual in the dataset."[130] The second is "linkability"or "the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases)."[131] The third is "inference" or "the possibility to deduce, with significant probability, the

[124] Ohm, *supra* note 64, at 1765–68.

[125] Vicenc Torra et al., *Big Data Privacy and Anonymization*, *in* PRIVACY AND IDENTITY MANAGEMENT: FACING UP TO THE NEXT STEPS (Anja Lehmann et al. eds., 2016).

[126] *Id.*

[127] Ohm, *supra* note 64, at 1701.

[128] Clete A. Kushida et al., *Strategies for De-Identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies*, 50 MEDICAL CARE 82 (2012).

[129] M.T. Nguyen et al., *Model Consent Clauses for Rare Disease Research*, 20 BMC MED. ETHICS 55 (2019), https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-019-0390-x#citeas.

[130] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymization Techniques* at 3, 10, 0829/14/EN WP216 (Apr. 10, 2014), https://www.pdpjournals.com/docs/88197.pdf.

[131] *Id*

value of an attribute from the values of a set of other attributes."[132] In the technical realm, computer scientists are actively studying the strengths and weaknesses of different anonymization and pseudonymization techniques.[133]

Third, context must be taken into account. Ohm explains that this requires a contextual analysis of risk that considers factors like the environment of release, the motives that individuals might have to reidentify the data, the trust that exists in people or institutions handling the data, and, of course, the potential linkage of released data with other data.[134] It is emphasized that data cannot be evaluated in isolation because "the same dataset will have different risks under different circumstances."[135]

When considering the context of a specific data share, it is especially important to evaluate whether data is being released to the world at large or whether data is being released in a non-public setting.[136] Ohm warned in 2010, "Regulators should scrutinize data releases to the general public much more closely than they do private releases between trusted parties."[137] This makes sense because it is more difficult to implement controls when it comes to a public data share.[138] As such, the risk of re-identification must be minimal.[139] However, if data is shared in a non-public setting then a higher risk is permissible since security, privacy, and contractual controls, discussed more below, can be established.[140]

[132] *Id.*

[133] Suntherasvaran Murthy, Asmidar Abu Bakar, Fiza Abdul Rahim & Ramona Ramli, A Comparative Study of Data Anonymization Techniques, 2019 IEEE 5th Int'l Conf. on Big Data Sec. on Cloud (BigDataSecurity), IEEE Int'l Conf. on High Performance and Smart Computing (HPSC), and IEEE Int'l Conf. on Intelligent Data & Sec. (IDS), https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8819477; *see also* Ambika Pawar, Swati Ahirrao & Prathamesh P. Churi, *Anonymization Techniques for Protecting Privacy: A Survey*, 2018 IEEE Punecon, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8745425.

[134] Ohm, *supra* note 64, at 1766–67.

[135] EUROPEAN MEDICINES AGENCY, *supra* note 12.

[136] *External Guidance on the Implementation of the European Medicines Agency Policy on the Publication of Clinical Data for Medicinal Products for Human Use*, EUROPEAN MEDICINES AGENCY (Sept. 2017) [hereinafter *External Guidance*], https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-1.pdf.

[137] Ohm, *supra* note 64, at 1765.

[138] *See generally External Guidance*, *supra* note 136.

[139] *See id.*

[140] *Id.*

It is also important to consider the motives that individuals might have to reidentify the data.[141] There are many different people or entities that are interested in reidentification of private, deidentified data subjects.[142] Motives for reidentification include, for example, investigative reporting, marketing, blackmail, insurance, and political action.[143] If there is no motive to launch a re-identification attack, then the probability of an attempt is low, notwithstanding other factors.[144] For example, academic researchers rarely desire to reidentify people in their datasets, at least in comparison to financially-motivated identity thieves and, therefore, they should be provided more liberal access to anonymized data sets.[145]

The UK Anonymi(s)ation Network has developed a tool to think critically about the individual data context.[146] Within the framework, the organization sets forth ten key components[147] to consider within three broad areas (the data situation audit, disclosure risk assessment and control and impact management).[148] This tool recognizes that reidentification risk arises from a number of different factors that include the properties of the dataset but also much more.[149] It is the interaction between that dataset, people, other data and the structures that shape those interaction that are the most relevant.[150]

Finally, the existence of additional controls must be evaluated. Here, contractual controls should be included to manage the (expected or potential) consequences of sharing a particular dataset.[151] For example, data use agreements

[141] Ohm, *supra* note 64, at 1765–68.

[142] Salvador Ochoa et al., *Reidentification of Individuals in Chicago's Homicide Database a Technical and Legal Study*, MASS. INST. TECH. (2001), http://web.mit.edu/sem083/www/assignments/reidentification.html.

[143] *Id.*

[144] *See* EL EMAM, *supra* note 90.

[145] *Id.*

[146] ELLIOTT ET AL., *supra* note 116.

[147] *Id.* (setting forth the following ten components: component 1: describe your (intended) data situation; component 2: understand your legal responsibilities; component 3: know your data; component 4: understand the use case; component 5: meet your ethical obligations; component 6: identify the processes you will need to go through to assess disclosure risk; component 7: identify the disclosure control processes that are relevant to your data situation; component 8: identify your stakeholders and plan how you will communicate with them; component 9: plan what happens next once you have shared or released the data; component 10: plan what you will do if things go wrong).

[148] *Id.*

[149] Mackey et al., *supra* note 83.

[150] *Id.*

[151] *Id.*

are commonly used to outline procedures, expectations, requirements, and restrictions for using and disclosing the data.[152] Security controls should further be included. Here, it is key to limit access to the data and the kinds of analyses that be conducted on a dataset[153] as a way to anticipate and mitigate the risks associated with future releases of data.[154] These types of controls might involve, for example, restrictions on secondary researchers and limit access to certain classes of researchers, such as researchers that have pledged a duty of confidentiality.[155] The OECD suggests that independent review bodies are relied upon to evaluate data use proposals for public benefits and adequacy of data security.[156] Finally, privacy controls should also be considered. These types of measures involve everything from simple redaction approaches to robust disclosure limitation techniques like secure multiparty computation.[157]

When it comes to understanding the role that additional controls can play in facilitating the secondary use of health data, Finland provides an excellent example. In 2019, the Finnish Act on the Secondary Use of Health and Social Data made it possible to use health and social data not only in research and compilation of statistics but also in the development and innovation activities, teaching, knowledge management, supervision and steering in the social welfare and healthcare sector and in official planning tasks.[158] In addition to strictly limiting data release in pseudonymized, anonymized or statistical format, there is also a license

[152] Micah Altman et al., *Practical Approaches to Big Data Privacy Over Time*, 8 INTERNATIONAL DATA PRIVACY LAW 29, 34 (2018) (stating that "[Data use] [a]greements typically describe the contents and sensitivity of the data; the restrictions on access, use, and disclosure; the data provider's rights and responsibilities; the data confidentiality, security, and retention procedures to be followed; the assignment of liability between the parties; and relevant enforcement procedures and penalties. They are used to set forth obligations, ascribe liability and other responsibilities, and provide a means of recourse if a violation occurs.").

[153] Mackey et al., *supra* note 83.

[154] Altman et al., *supra* note 152, at 35.

[155] *Id.* at 34.

[156] ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, ENHANCING ACCESS TO AND SHARING OF DATA: RECONCILING RISKS AND BENEFITS FOR DATA RE-USE ACROSS SOCIETIES (Nov. 26, 2019), https://www.oecd-ilibrary.org/sites/276aaca8-en/index.html?itemId=/content/publication/276aaca8-en.

[157]Altman et al., *supra* note 152, at 47.

[158] Press Release, Ministry of Social Affairs and Health, Finnish Government, New Act Enables Effective and Secure use of Health and Social Data (Apr. 24, 2019), https://valtioneuvosto .fi/en/article/-/asset_publisher/1271139/uusi-laki-mahdollistaa-sosiaali-ja-terveystietojen-tehokkaan-ja-tietoturvallisen-kayton.

requirement.[159] Additionally, data can only be processed in a closed, data-secure and access-monitored environment.[160]

After considering the four factors, there must be an attempt to measure or depict the risks associated with a particular data release. This will help organizations weigh the benefits of the data processing and determine whether the risks are proportionate to the benefits of the data processing.[161] Any privacy risk assessment requires the selection of the appropriate risk metrics as well as a suitable threshold for data release.[162] It is also necessary to compute the actual measurement of the risk if the data is to be disclosed.[163] A simple risk matrix might include the likelihood of reidentification (ranging from highly likely, moderate, unlikely, and remote possibility) as well as the consequences of the potential re-identification (ranging from minor, small, medium, large, and extreme).[164]

VII. Synthetic Data Generation as an Alternative Approach for Data Sharing?

The creation of synthetic data is one technique to deal with identifiability in high-dimensional data[165] and, in the words of Bellovin et al., "if constructed properly—may solve Professor Ohm's failure of anonymization."[166] Synthetic data, first introduced by Rubin in 1993, can be conceptualized as "artificially generated data that has approximately the same properties (i.e. values) as the raw data, but that

[159] The Finnish Innovation Fund Sitra, *A New Act Will Increase the Transparency and Impact of Using Social Welfare and Healthcare Data* (Mar. 13, 2019), https://www.sitra.fi/en/news/new-act-will-increase-transparency-impact-using-social-welfare-healthcare-data/.

[160] *Id.*

[161] Nat'l Inst. Stds. & Tech., Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management (2020), https://www.nist.gov/privacy-framework (citing Eur. Data Prot. Supervisor, Necessity & Proportionality, https://edps.europa.eu/data-protection/our-work/subjects/necessity-proportionality_en).

[162] European Medicines Agency, *supra* note 12.

[163] *Id.*

[164] For a more complex privacy risk assessment matrix concerning dataset publication, see Office of the Information Commissioner Queensland, Right to Information and Privacy Guidelines Appendix—Guideline: Dataset Publication and Risk Assessment (2013), https://www.oic.qld.gov.au/__data/assets/pdf_file/0015/16242/guideline-dataset-publication-and-risk-assessment-appendix-1.pdf.

[165] Altman et al., *supra* note 152.

[166] Steven M. Bellovin et al., *Privacy and Synthetic Datasets*, 22 Stan. Tech. L. Rev. 1, 49 (2019).

does not allow conclusions to be drawn about the individuals in the original dataset."[167] Ruiz et al. further explain,

> Synthetic data rely on a principle that is by nature similar to the imputation of missing values in a data set. The idea is to fit a model, called a synthesizer, to the original data; then values are drawn from the synthesizer to replace original data rather than merely imputing missing data.[168]

Essentially, synthetic data permits algorithms and models to be tested on artificial data instead of relying on the original data for science investigations.[169] To put it in "meta terms," synthetic data "can be thought of as "fake" data created from "real" data."[170]

It is possible to divide synthetic data into two subcategories: (1) fully synthetic data or (2) partially synthetic data.[171] Soria-Comas and Domingo-Ferrer explain:

> In partially synthetic data sets, only some of the values of the original data set are replaced by synthetic (simulated) values, usually the ones that are deemed too sensitive. On the contrary, in fully synthetic data sets a new data set is generated from scratch. When generating a fully synthetic data set, the original data set is viewed as a sample from some underlying population and the synthetic data set is generated by taking a different sample.[172]

There are a number of different methods for generating synthetic data and all of the approaches have limitations.[173] The latest innovations in the fields of ML and AL may increase the possibilities to create these kinds of data.[174] Here, it must be

[167] D.B. Rubin, *Statistical Disclosure Limitation*, 9 J. OFFICIAL STAT. 461 (1993); AIRCLOAK, DATA ANONYMIZATION IN DIGITAL BUSINESS MODELS, https://aircloak.com/wp-content/uploads/Aircloak_Data-Anonymization-in-Digital-Business-Models.pdf; *see also* MCGRAW-HILL, INC., THE MCGRAW-HILL DICTIONARY OF SCIENTIFIC AND TECHNICAL TERMS (Sybil P. Parker ed., 5th ed. 1994) (defining synthetic data as "any production data applicable to a given situation that are not obtained by direct measurement").

[168] N. Ruiz et al., *On the Privacy Guarantees of Synthetic Data: A Reassessment from the Maximum-Knowledge Attacker Perspective*, 11126 LECTURE NOTES IN COMPUTER SCI. 59, 61 (2018).

[169] EUROPEAN MEDICINES AGENCY, *supra* note 12.

[170] Bellovin et al., *supra* note 166, at 21.

[171] Jordi Soria-Comas & Josep Domingo-Ferrer, *Big Data Privacy: Challenges to Privacy Principles and Models*, 1 DATA SCIENCE AND ENGINEERING 21 (2015).

[172] *Id.*

[173] Douglas Garcia Torres, Generation of Synthetic Data with Generative Adversarial Networks (Nov. 26, 2018) (unpublished Master Thesis, Royal Institute of Technology) (on file with Eindhoven University of Technology).

[174] Jungang Xu, Hui Li & Shilong Zhou, *An Overview of Deep Generative Models*, 32 IETE TECH. REV. 131 (2014).

noted that advances in machine learning and deep learning techniques promise to deliver synthetic data with higher quality and in larger quantities than that previously thought possible.[175]

Synthetic data could solve the anonymization problem because it offers a way to share data records without sharing any individual personal data. This is because synthetic data is not personal—it is fictitious.[176] In addition, the data would still be useful, at least in theory, because the data analysts could find answers that would approximate to what they would have found from the original data.[177]

Like anonymization, however, synthetic data has its drawbacks. For example, if the synthetic data is "approximately equivalent" to the original data then there may nevertheless be privacy risks for certain subsets of the population.[178] In other words, the "fake" data may actually look so similar to the "real" data that the difference is meaningless from a privacy perspective. On the other hand, if the synthetic data fails to capture all of the important features of the original dataset then it may not be possible to discover key relationships in the same way as it would be if one was processing the original data.[179] It is also important to mention that synthetic data may also need to be verified on the real dataset.

VIII. CONCLUSION

The digital transformation of healthcare presents huge opportunities, particularly in the face of population aging, which promises to strain even the most efficient medical systems. There are risks in data-driven innovations like the IoHT but there are ways to mitigate those risks. It is now, more than ever, necessary for governments to establish regulatory environments that support data-driven innovation while strengthening trust in technology. Here, the false dichotomy between privacy and technology, one that suggests individuals must give up privacy or reap the benefits of technology, must be debunked.

This paper has sought to contribute to the current academic debate surrounding the role of anonymization in the IoHT by evaluating how data controllers can balance privacy risks against the quality of output data and select the appropriate privacy

[175] Torres, *supra* note 173, at 6–9.

[176] Douglas J. Sylvester & Sharon Lohr, *Counting on Confidentiality: Legal and Statistical Approaches to Federal Privacy Law After the USA Patriot Act*, 2005 WIS. L. REV. 1033, 1119 (2005) (citing Donald B. Rubin, *Discussion: Statistical Disclosure Limitation*, 9 J. OFFICIAL STAT. 461, 461 (1993)).

[177]*Id.*

[178] Sylvester & Lohr, *supra* note 176, at 1119–20.

[179] *Id.*

model that achieves the aims underlying the concept of Privacy by Design. It has set forth several approaches for identifying the risk of re-identification in the IoHT as well as explored the potential for synthetic data generation to be used as an alternative method to anonymization for data sharing. Ostensibly, if the benefits of various anonymization techniques and emerging technologies like synthetic data are realized then a Golden Age of privacy may emerge.