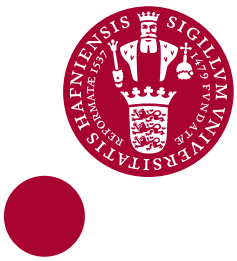

Homomorphic Encryption

The 'Holy Grail' for Big Data Analytics & Legal Compliance in the Pharmaceutical and Healthcare Sector?

Corrales Compagnucci, Marcelo; Meszaros, Janos; Minssen, Timo; Arasilango , Arasaratnam ; Ous , Talal ; Rajarajan, Muttukrshnan

Published in:
European Pharmaceutical Law Review

DOI:
10.21552/eplr/2019/4/5

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Homomorphic Encryption: The 'Holy Grail' for Big Data Analytics and Legal Compliance in the Pharmaceutical and Healthcare Sector?

*Marcelo Corrales Compagnucci, Janos Meszaros, Timo Minssen, Arasaratnam Arasilango, Talal Ous and Muttukrishnan Rajarajan**

The pharmaceutical and healthcare sector is a prime target for cybercriminals around the world. These cyber-attacks represent significant challenges in the context of data protection and data security. The General Data Protection Regulation (GDPR) imposes strict rules regarding the processing and analysis of personal data. In conventional approaches, data analysts request data from various sources. Then, they anonymise or pseudonymise the data using various tools and techniques. These methods often use powerful algorithms to ensure a high level of security. However, these methods tend to either reduce the quality of data for further analysis or they expose the data while decrypting it for analysis. Homomorphic Encryption (HE) has recently been touted as the 'Holy Grail' of cryptography since it allows the analysis of big data sets without ever needing to decrypt and thus compromising the confidentiality of the data. This provides a whole new layer of protection and at the same time allows the processing of data for secondary use and scientific research. While HE is not a new technology, it is still in the early stages of development. In this piece, we will introduce a new automated tool for searching and analysing encrypted data using HE techniques, which is being developed within the scope of the EnergyShield project.[1]

I. Introduction

Big data analytics refers to the science of analysing raw data to enhance productivity and make meaningful conclusions from data sets. Data are extracted and categorised using various techniques to identify and analyse behavioural information and patterns.[2] In other words, big data analytics is about harnessing the power of big data[3] which can reduce costs and deliver potentially life-saving insights in the pharmaceutical and life sciences industry.[4]

Pharmaceutical and healthcare organisations are increasingly looking to take advantage of big data analytics using robust IT infrastructure and cloud-

DOI: 10.21552/eplr/2019/4/5

* Marcelo Corrales Compagnucci is Postdoctoral Fellow at the Centre for Advanced Studies in Biomedical Innovation Law (CeBIL), Faculty of Law, University of Copenhagen (UCPH). For correspondence: <marcelo.c.compagnucci@jur.ku.dk>. Janos Meszaros is a Postdoctoral Fellow at the Institute of European and American Studies, Academia Sinica. Timo Minssen is a Professor of Law at the University of Copenhagen (UCPH), Founding Director of the Centre for Advanced Studies in Biomedical Innovation Law (CeBIL). Arasaratnam Arasilango is a Technical Associate at Tech Inspire Ltd. Talal Ous is the CEO of Tech Inspire Ltd. Muttukrshnan Rajarajan is a Professor of Security Engineering, Department of Electrical and Electronic Engineering, City University of London.

1 The EnergyShield project will develop an integrated toolkit, which will combine inter alia novel security tools and allow data collectors to process and share data in cloud-based applications. This project has received funding from the European Union's H2020 research and innovation programme under the Grant Agreement No. 832907, <https://energy-shield.eu> accessed 5 November 2019.

2 David Loshin, *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph* (Morgan Kaufmann 2013).

3 Andrea Darrel et al, 'The Benefits of Big Data Analytics in the Healthcare Sector', in Baoying Wang, Ruowang Li and William Perrizo (eds), *Dig Data Analytics in Bioinformatics and Healthcare* (IGI Global 2015), 411.

4 Cio Dive 'Harnessing the Power of Big Data' (5 July 2018) <https://www.ciodive.com/spons/harnessing-the-power-of-big-data/526876/> accessed 5 November 2019.

based solutions. As a research-intensive industry, they rely largely on their ability to make strategic decisions based on the analysis elicited from big data results.[5] The cloud is, therefore, an obvious choice for applications running large workloads and storing large volumes of data. Cloud providers offer highly scalable resources that could help the pharma and healthcare sector to tap into the power of big data analytics.[6] Hyperscale cloud providers and AI technologies are, therefore, fundamental in this new form of healthcare and drug development.[7]

Nonetheless, since the General Data Protection Regulation (GDPR) came into force on the 25 May 2018, more limits have been imposed on the way organisations process and share personal data.[8] The GDPR was designed to strengthen data protection and privacy for all EU citizens and to empower individuals by granting them more control and transparency over their data when using Internet services.[9] Falling foul of GDPR can result in hefty financial penalties and data breach notifications must be reported within 72 hours.[10]

The question is how to create a framework which is secure enough, but still ensures that the data is of sufficient quality that big data analytics are useful and meaningful. Homomorphic encryption (HE) can do that and for this reason has been referred to as the 'Holy Grail' of cryptography. Personal information needs to be encrypted both at rest (when data is stored) and while in transit (when transferred from one place to another). While modern encryption algorithms are very secure because they require a lot of time and resources to break them, they also make it impossible to process and analyse the data without first decrypting it – and decrypting the data makes it again vulnerable to cyber-attacks or unauthorised third-party interception.[11] Therefore, HE might solve the vulnerability inherent in all other approaches to data protection and data security.

While HE can add an extra layer of security, it is still in an early phase of development. At present, the problem with HE is the 'noise' threshold. The noise is typically a small term added into the ciphertext while encrypting data, which considerably slows down the processing speed. The HE operations always increase the noise. However, the decryption operation does not work if the noise is larger than a certain maximum value.

In this piece we present a new automated framework which reduces the noise of HE operations. In other words, the analysed data never reveals any private information because it is still encrypted – both at rest and in transit – and at the same time remains useful for big data analytics.

The paper is structured as follows. After this introduction, Section II briefly discusses current technological trends and data security concerns with regard to the latest cyber-attacks in the pharma and healthcare sector. Section III, explains the pros and cons of anonymisation and pseudonymisation techniques with regard to big data analytics in light of the GDPR. Section IV, focuses on the new proposed HE tool and explains the reasons why this could be a good alternative framework to protect patient's data without exposing it. Finally, section V concludes the paper.

II. New Technologies Disrupting the Pharmaceutical Sector and Raising Cybersecurity Concerns

Technological advances in recent decades have been characterised by exponential growth. As a result of this growth, the pharmaceutical and health science industry has become more digitalised. Pharma and healthcare organisations are becoming more like interconnected platforms where they can store, share and analyse large amounts of data in unprecedented

5 Suresh Kumar Peddoju, 'Big Data Analytics for Childhood Pneumonia Monitoring' in Chintan Bhatt and Suresh Kumar Peddoju (eds) *Cloud Computing Systems and Applications in Healthcare* (IGI Global 2016), 87.

6 Nick Ismail, 'How Cloud Computing Can Transform the Pharmaceutical Industry' (14 October 2016) <https://www.information-age.com/cloud-computing-pharmaceutical-industry-123462676/> accessed 5 November 2019

7 Veronica Combs, 'Pharma Companies are Counting on Cloud Computing and AI to Make Drug Development Faster and Cheaper' (12 August 2019) <https://www.zdnet.com/article/pharma-companies-are-counting-on-cloud-computing-and-ai-to-make-drug-development-faster-and-cheaper/> accessed 5 November 2019

8 Nidhish Dhru, *Office 365 for Healthcare Professionals: Improving Patient Care Through Collaboration, Compliance, and Productivity* (Apress 2018), 62.

9 Kerina Jones, 'Incongruities and Dilemmas in Data Donation: Juggling Our 1s and 0s' in Jenny Krutzinna and Luciano Floridi (eds), *The Ethics of Medical Data Donation*, Philosophical Studies Series Vol. 137 (Springer 2019), 79.

10 Sanjay Sharma, *Data Privacy and GDPR Handbook* (Wiley & Sons 2019), 103.

11 Casey Crane, 'What is Homomorphic Encryption?' (4 October 2019) <https://www.experfy.com/blog/what-is-homomorphic-encryption> accessed 5 November 2019

ways.[12] Cloud computing, big data analytics,[13] artificial intelligence (AI), machine learning (ML)[14] and Internet of Things (IoT)[15] are amongst the innovations that have started to disrupt the market.

Among these new technologies, the cloud[16] seems to be a clear choice for applications running large workloads and processing large volumes of data. Public clouds are, therefore, the future of big data analytics, and the pharmaceutical sector – after years of cautious consideration – has decided to take the plunge.[17]

The pharma and healthcare sector has always been heavily regulated and adopting cloud computing services seemed to involve risky operations. Recent studies from a Business Communications Company (BCC) Research Report, however, suggested that the global market for cloud computing technologies in healthcare will reach $35bn by 2022.[18] Hospitals and health systems are using more cloud computing services to hasten their digital transformation. Computational drug discovery uses a combination of cloud-based applications and AI technologies to improve the odds of making new drug developments cheaper and faster.[19] These developments will allow them to deliver more personalised medicine and patient care, use data-driven algorithms for better decision-making and engage more fully with doctors and patients.[20]

However, the pharma and healthcare sector's previous scepticism to cloud adoptions stemmed from security and privacy concerns. Healthcare data can fetch a high price on the 'Dark Web' and coupled with other personal and financial data, it can be used to conduct medical fraud.[21] A research study at Vanderbilt University revealed that more than 2,100 patient deaths are related to hospital data breaches each year.[22] Security researchers have proven that it is possible to infiltrate hospital equipment, steal patient data, fake results and tamper with life-saving medical equipment.[23]

A cyber security analyst at Kaspersky Lab has warned there is a reasonable danger of hacked medical devices resulting in patient deaths.[24] Furthermore, cybersecurity experts at Ben-Gurion University suggested that medical imaging devices, such as computed tomography (CT) scanners, are potentially vulnerable to cyber-threats and that manufacturers as well as healthcare providers must take actions in protecting them.[25]

In 2016, more than 16 million medical records were stolen from healthcare organisations in the US. In

12 Daniel Burns and Scott Sundseth, 'Applications of Pharmacogenetics in Pharmaceutical Research and Development' in Anke-Hilse Maitland-van der Zee and Ann K. Daly (eds) *Pharmacogenetics and Individualized Therapy* (Wiley & Sons 2012), 452.

13 See, generally, Christoph Thuemmler and Chunxue Bai, *Health 4.0: How Virtualization and Big Data are Revolutionizing Healthcare* (Springer 2017).

14 For examples on AI and ML, see Claudia Rijcken 'Sequoias of Artificial Intelligence' in Claudia Rijcken (ed) *Pharmaceutical Care in Digital Revolution: Insights Towards Circular Innovation* (Academic Press 2019), 125 et seq.

15 See, generally, B, Pankajavalli and G Karthick, *Incorporating the Internet of Things in the Healthcare Applications and Wearable Devices* (IGI Global 2019).

16 The term cloud computing has been defined in various ways. In this article we adopt the definition from the US National Institute of Standards and Technology (NIST) as it embraces relevant aspects of the different cloud service and deployment models: 'Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (ie, networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.' See, Peter Mell and Tim Grance, 'The NIST Definition of Cloud Computing' (September 2011) <https://csrc.nist.gov/publications/detail/sp/800-145/final> accessed 5 November 2019.

17 Andy Newsom, 'Cloud Computing: Pharma Takes The Plunge' <https://life-sciences.cioapplications.com/cioviewpoint/cloud-computing-pharma-takes-the-plunge-nid-924.html> accessed 5 November 2019.

18 BCC Research Report, 'Healthcare Cloud Computing: Global Markets to 2022 (January 2018) <https://www.bccresearch.com/market-research/healthcare/healthcare-cloud-computing-technologies-report.html> accessed 5 November 2019.

19 Veronica Combs, 'Pharma Companies are Counting on Cloud Computing and AI to Make Drug Development Faster and Cheaper' (12 August 2019) <https://www.zdnet.com/article/pharma-companies-are-counting-on-cloud-computing-and-ai-to-make-drug-development-faster-and-cheaper/> accessed 5 November 2019.

20 David Champagne, Amy Hung and Olivier Leclerc, 'The Road to Digital Success in the Pharma Sector, McKinsey' (August 2015) <https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/the-road-to-digital-success-in-pharma> accessed 5 November 2019.

21 Craig Stewart, 'How the Cloud Will Change the Pharmaceutical Industry' (13 March 2019) <https://www.snaplogic.com/blog/cloud-change-the-pharmaceutical-industry> accessed 5 November 2019.

22 Robert Abel, 'Vanderbilt University Researcher's Claim Breaches Linked to Patient Deaths' (27 March 2018) <https://www.scmagazineuk.com/vanderbilt-university-researchers-claim-breaches-linked-patient-deaths/article/1472987> accessed 5 November 2019.

23 Sarah Griffiths, 'Could Hackers Kill Hospital Patients? Cyber Security Experts Prove They Can Steal Patient Data, Fake Results and Damage Equipment' (20 April 2016) <https://www.dailymail.co.uk/sciencetech/article-3549592/Could-hackers-kill-hospital-patients-Cyber-security-experts-prove-steal-patient-data-fake-results-damage-equipment.html> accessed 5 November 2019.

24 Owen Hughes, 'Patient Death From Hacked Medical Devices Plausible Says Top Kaspersky Security Researcher' (29 March 2018) <https://www.digitalhealth.net/2018/03/killer-medical-devices-not-just-hype-says-kaspersky/> accessed 5 November 2019.

25 Tom Mahler, 'Medical Imaging Devices are Vulnerable to Cyber-attacks, Israeli Team Warns' <https://cyber.bgu.ac.il/medical-imaging-devices-are-vulnerable-to-cyber-attacks-israeli-team-warns/> accessed 5 November 2019.

that same year, the healthcare sector was the fifth most targeted industry in terms of cyber-attacks. Earlier in 2017, WannaCry – a ransomware worldwide cyber-attack – also impacted the UK's National Health Service (NHS). The WannaCry ransomware cryptoworm targeted the Microsoft Windows operating system by encrypting data and asking for ransom to be paid in the Bitcoin cryptocurrency. Many other similar attacks have resulted in stolen health medical records from hospitals.[26]

According to the Protenus Breach Barometer, the healthcare sector suffered from 503 data breaches affecting 15 million patient records in 2018, which is three times more the amount seen in 2017. However, just over halfway through 2019, the number of data breaches have risen to more than 25 million patient records.[27]

Health data breaches came in many forms. Third-party vendors and phishing attacks were among the most common incidents. Particularly concerning is that most of these cyber-attacks were unreported for a long period. As it stands, the rest of the year 2019 has been peppered with massive data breaches and may prove to be the worst year for healthcare cybersecurity.[28]

Due to this growing complexity of cyber-attack and increased digitisation, the European Commission has identified the healthcare sector as a critical sector where information security should be emphasised. Digital security, privacy, data protection and accountability are at the forefront of the Horizon 2020, which is the biggest EU Research and Innovation programme. The trend towards digital health, along with platform-based and data-driven patient ecosystems has exacerbated the risks of cyber-attacks. Healthcare organisations collect information from patients and provide analysis for diagnosis and better treatment even using mobile devices.[29] The following section discusses the pros and cons of some of the typical ways to reduce these risks taking into account the legal provisions in light of the GDPR.

III. De-Identification, Anonymisation and Pseudonymisation: GDPR Compliance and the 'Achilles heel' of Data Analytics

The GDPR recognises the potential benefits of big data analytics for society and individuals in health, scientific research, the environment and other related areas. However, it also imposes stringent rules for the processing and analysis of big data due to growing threats, increasing number of cyber-attacks and numerous data leak scandals.[30] This conflict places data protection and data security principles under strain and finding a better framework is at the hub of the discussion in legal fora.[31]

The GDPR identified the privacy-enhancing effect of anonymisation and pseudonymisation methods by providing exceptions to many of the cumbersome provisions of the GDPR. Both approaches are good ways for pharmaceutical companies and healthcare organisations to mitigate the probability of data breaches and therefore reduce the risk of paying fines.

Anonymisation and pseudonymisation are highly recommended data processing techniques in the GDPR because they reduce risk and aid data processors in complying with their data protection obligations for secure processing of personal information.[32] These two methods, however, differ significantly in light of the GDPR. The main difference rests on whether the data subject can be re-identified or not.[33]

---

26 Anomali, Inc. 'Cyber Threat Landscape: The Healthcare Industry' (2019) <https://dsimg.ubm-us.net/envelope/399273/564983/The-Healthcare-Industry.pdf> accessed 5 November 2019.

27 Jessica Davis, 'The 10 Biggest Healthcare Data Breaches of 2019, So Far' (23 July 2019) <https://healthitsecurity.com/news/the-10-biggest-healthcare-data-breaches-of-2019-so-far> accessed 5 November 2019.

28 ibid.

29 European Cyber Security Organisation (ECS), 'Healthcare Sector Report: Cyber Security for the Healthcare Sector' (March 2018) <https://www.ecs-org.eu/documents/publications/5ad7266dc1cba.pdf> accessed 5 November 2019.

30 Peter Chan and Lauren Hankel, 'System for Detecting Data Protection Violations' in Noëlle van der Waag-Cowling and Louise Leenen (eds) *Proceedings of the 14th International Conference on Cyber Warfare and Security*, Stellenbosch University, South Africa, 28 February – 1 March 2019 (ACPIL 2019), 30.

31 Dan Lohrmann, 'Why You Need the Cybersecurity Framework' (20 May 2018) <https://www.govtech.com/blogs/lohrmann-on-cybersecurity/why-you-need-the-cybersecurity-framework.html> accessed 5 November 2019.

32 Sérgio Ribeiro and Emilio Nakamura, 'Pseudonymisation Approach in a Health IoT System to Strengthen Security and Privacy Results from OCARIoT Project' in Robin Doss, Selwyn Piramuthu and Wei Zhou (eds) *Future Network Systems and Security*, 5th International Conference, FNSS 2019, Melbourne, VIC, Australia, November 27-29, 2019, Proceedings (Springer 2019), 135.

33 Robert Walters, *Leon Trakman and Bruno Zeller, Data Protection Law: A Comparative Analysis of Asia-Pacific and European Approaches* (Springer 2019), 90.

Anonymisation refers to irreversibly severing personally identifiable information from data sets, which prevents any future re-identification of the data subject. Therefore, anonymisation would be the most desirable approach to personal data protection. It allows sharing data for secondary purposes (such as research, public health, etc.) without risking individuals privacy.[34]

Recital 26 of the GDPR states that the principles of data protection should not apply to anonymous information. Namely, information that is not related to an identified or identifiable natural person. Thus, when anonymisation is engineered appropriately, it places the processing and storage of personal data outside the scope of the GDPR. For this reason, anonymisation can be an effective method for mitigating privacy risks and protecting data subjects.[35]

The Article 29 Working Party,[36] concludes that to meet the current anonymisation standards, data must be processed in such a way that a natural person cannot be identified anymore by using 'all means likely reasonably to be used' either by the controller or a third party. This means that the prerequisites and the objectives of the anonymisation process must be clearly established from the outset in order to achieve the targeted anonymisation while producing some useful data.[37]

An important aspect is that the processing of data must use an irreversible de-identification mechanism. That is, the data must be stripped of sufficient elements such that the data subject can no longer be identified.[38] The Working Party advises that there is no one-size-fits-all solution. Instead, anonymisation should be taken on a case-by-case basis, using a variety of techniques and factoring in the opinion's recommendations. Data controllers are advised to tailor-make the anonymisation method to the specific circumstances.[39]

However, one fundamental problem with anonymisation is that the threshold to achieve it is very high and data controllers often fall short of anonymising data correctly. Data can be so distinct that it can be easily identified thus fall under the scope of the GDPR. Even if any identifying features are scrubbed from the data, research studies revealed how easy it is to re-identify data.[40]

Another major problem is that anonymisation decreases the value of data analytics. Masking techniques tend to distort the quality of data significantly. Therefore, no data analytics can be done properly. The main reason why companies – in particular pharmaceutical and healthcare organisations – collect data, is that they can search for patterns. Most companies do not care about any single individual's data, but the insights they can get from the aggregate. However, stripping away identifiers that make data useful does not always bring the same analytical value. Detailed information can give a pharmaceutical or healthcare organisation the edge for planning a better treatment or making a new drug discovery such as preparing for flu outbreaks.[41]

To understand the concept of anonymisation, consider the scenario where statistics are gathered from a group of patients (eg, patients with colon cancer in London during the last 10 years). If the group is large enough and anonymisation techniques have been applied appropriately, it would not be possible to identify these patients individually. Therefore, the authorities may publish the statistical reports and findings of the group of patients without any risks or privacy concerns. However, with these kinds of anonymisation methods, it would not be possible to contact the individuals if such data is analysed in the context of scientific research and meaningful results were revealed to treat and cure the patients.[42]

In contrast to anonymisation, pseudonymisation is a de-identification procedure by which personal

34 Balaji Raghunathan, *The Complete Book of Data Anonymisation: From Planning to Implementation* (CRC Press 2013).

35 Recital 26 of the GDPR.

36 Namely, randomization and generalization. In particular, the opinion examines noise addition, permutation, differential privacy, aggregation, k-anonymity, l-diversity and t-closeness. See Opinion 05/2014 on Anonymisation Techniques adopted on 10 April 2014.

37 Opinion 05/2014 on Anonymisation Techniques adopted on 10 April 2014.

38 Richard Morgan and Ruth Boardman, *Data Protection Strategy: Implementing Data Protection Compliance* (Sweet & Maxwell 2003), 49.

39 Opinion 05/2014 on Anonymisation Techniques adopted on 10 April 2014.

40 Kelsey Campbell-Dollaghan, 'Sorry Your Data Can Still Be Identified Even If It's Anonymized' (12 October 2018) <https://www.fastcompany.com/90278465/sorry-your-data-can-still-be-identified-even-its-anonymized> accessed 5 November 2019.

41 Peter Leihn, 'Homomorphic Encryption is Now a Reality' (9 July 2019) <https://www.cso.com.au/article/663776/homomorphic-encryption-now-reality/> accessed 5 November 2019.

42 Janos Meszaros and Chih-hsing Ho, 'Building trust and transparency? Challenges of the opt-out system and the secondary use of health data in England' (2019) Medical Law International, 19(2–3), 169–171.

data is replaced by one or more artificial identifiers or pseudonyms in such a way that data cannot be linked directly to their corresponding nominative identities.[43] Pseudonymisation is also regarded to be a strong security technique to make sensitive health data less explicit and still easy to manage.[44]

Article 40 (2) (d) of the GDPR recommends controllers and processors of personal data to implement pseudonymisation of personal data as a code of conduct or good practice.[45] Pseudonymisation is defined under the GDPR as 'the 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 organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.'[46]

Most pseudonymisation procedures in the pharma and healthcare sector use a trusted third party as an intermediary to perform the pseudonymisation process. This means that there are at least three entities involved in the process. There is the primary data source that has access to nominative personal data, the trusted third party, and the data register that uses the artificial identifiers (pseudonymised data) to protect the privacy of individuals.[47]

As a way of example, consider the situation of scientific research using a cohort of patients with diabetes or Parkinson's disease. In this case, pseudonymisation techniques could be used to reduce the 'linkability' of the dataset to the data subject. For instance, instead of a full name, a patient's name would be replaced with a code name such as '789463.' This security measure would make it (theoretically) impossible to identify the individual without knowing how the code was generated. To re-identify the patient, a 'key' (private or public) would be necessary to decrypt the information. The benefit of this approach is that it would be possible to contact the patients in case the researchers find a special treatment to cure them.[48]

While pseudonymisation is an effective security measure, pseudonymised data is still regarded as personal data. Thus, data must be protected accordingly and falls under the scope of the GDPR.[49] The crucial problem with pseudonymisation is the need to decrypt the data to use data analytics. The decryption exposes personal and sensitive data to various kinds of cyber-attacks. The reason is that a person within the organisation or a malicious third party can obtain exposure by using secondary identification tools based on the open data fields.[50]

In sum, de-identification methods include a wide array of tools and techniques to protect the data subject's privacy. At the upper end of the spectrum, there is the anonymous/aggregated data, which cannot identify particular individuals. At the other end of the scale, the availability of personal data without de-identification, which identifies the data subject directly, while pseudonymisation sits in the 'middle ground.'[51]

A complicating factor is that it is hard to tell under which category these techniques belong, which is challenging both from a technical and legal point of view. For example, in the UK, the term 'anonymisation' is defined broadly and usually falls under any of these categories. The failure to come to term with the concept of anonymisation creates confusion and it is not equally employed in the official guidelines, such as the Information Commissioner's (ICO) Code of Practice on Anonymisation. The concept of 'anonymisation' has a different meaning in the context of UK law and under the scope of the GDPR.[52]

In this article, we take the concepts of 'anonymisation' and 'pseudonymisation' in a way which is consistent with the GDPR and Article 29 Working Party

43 Sanjay Sharma, *Data Privacy and GDPR Handbook* (Wiley & Sons 2019), 88.

44 Heidelinde Hobel et al, 'Anonymity and Pseudonymity in Data-Driven Science' in John Wang (ed) *Encyclopedia of Business Analytics and Optimization* (IGI Global 2014), 128.

45 Taiwo Oriola, 'Internet Laws' in Khurshid Ahmad (ed) *Social Computing and the Law: Uses and Abuses in Exceptional Circumstances* (Cambridge University Press 2018), 40.

46 Art 40 (2) (d) of the GDPR.

47 Brecht Claerhout et al, 'A Data Protection Framework for Trans-European Genetic Research Projects' (2008) 141 Studies in Health Technology and Informatics 67-72

48 Janos Meszaros and Chih-hsing Ho, 'Big Data and Scientific Research: The Secondary Use of Personal Data under the Research Exemption in the GDPR' (2018) Acta Juridica Hungarica, Vol. 59, No. 4, 403-419.

49 Ulrich Flegel et al, 'Legally Sustainable Solutions for Privacy Issues in Collaborative Fraud Detection' in Christian Probst et al (eds) *Insider Threats in Cyber Security* (Springer 2010), 166.

50 Steve Touw, 'Homomorphic Encryption Alone is Security, Not Privacy' (14 September 2018) <https://www.immuta.com/homomorphic-encryption-alone-is-security-not-privacy/> accessed 5 November 2019.

51 Janos Meszaros and Chih-hsing Ho (n 42).

52 ibid.

opinions as depicted in the preceding paragraphs as well as the definitions provided by the International Organization for Standardization (ISO 29100:2011) which are also in line with the GDPR.[53]

The 'Achilles Heel' of big data analytics is therefore two-fold. On the one hand, there is a conceptual legal problem: the definition of 'anonymisation' and 'pseudonymisation' varies in domestic laws such as in the UK in comparison to the GDPR and other official international standards and guidelines. On the other hand, there is a technical problem: anonymisation and pseudonymisation techniques are both desirable and recommended for data protection and data security, but at the expenses of data utility. Both techniques reduce the quality and value of big data. In particular areas, such as clinical trials transparency, these technical and scientific problems will have to be addressed to reconcile GDPR compliance with the goals of recently enacted regulations that aim at enhancing data transparency, accountability, trust and scientific collaboration in the clinical trials sector.[54]

One method that could potentially solve this problem is an advanced type of encryption called Homomorphic Encryption (HE). With HE there is no need to decrypt the data to use data analytsis tools. This would allow the analysis of data without putting such data at risk; thus, providing an extra layer of security as explained in the section below.

IV. Homomorphic Encryption (HE): A New Automated Framework to Reduce the 'Noise'

Encryption is the process of taking a message and scrambling its content so that only certain people can read the message. There are two types of encryption: symmetric and asymmetric. Symmetric encryption is a type of encryption where only a secret key is used by the parties to encrypt and decrypt the message. However, symmetric encryption has a disadvantage: the sender and the receiver must exchange the key so that it can be used in the decryption process.[55] For example, if Alice has a sensitive document that she wants to send to Bob, she can use an encryption program to create a password and then send the document to Bob. In order to open the message, Bob needs the same password. The problem with symmetric encryption is that Alice would need to send the password to Bob and this could be risky over the cloud.[56]

This is precisely the problem that asymmetric encryption attempts to solve. Asymmetric encryption (also known as public-key encryption) uses two different keys to perform the encryption and decryption process. One key is called the public key and the other is a private key.[57] To put into perspective, it could be compared to a mailbox located on a public street. Anyone who knows the location of the mailbox could drop a letter through the slot. However, only the owner of the mailbox with a private key could open the mailbox and read the letters.[58]

At its most basic, HE is equal to any other asymmetric encryption scheme in that it allows everyone to encrypt data by using a public key, while allowing those with the private key to decrypt it. These are the building blocks of asymmetric cryptography widely used today in digital signatures and blockchain technology. From a technical point view, when using asymmetric encryption, both Alice and Bob would have to create a key pair on their computers. An acknowledged secure way of doing this is by using the Rivest-Shamir-Adleman (RSA) algorithm. This algorithm will generate a private and public key that mathematically match each other.[59]

HE is therefore an advanced and cutting-edge type of cryptography which has a great deal of potential to improve the efficiency and security of cloud computing operations. By and large, the resource-intensive encryption/decryption process requires a message to be encrypted first. Then, it must be decrypt-

53 ibid.

54 See eg, Regulation (EU) 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC (referred hereafter as 'Regulation (EU) 536/2014'). For a more detailed analysis cf. Timo Minssen, Neethu Rajam and Marcel Bogers, 'Clinical Trial Data Transparency and GDPR Compliance: Implications for Data Sharing and Open Innovation' in Katerina Sideri and Graham Dutfield (eds), *Openness, Intellectual Property and Science Policy in the Age of Data Driven Medicine*, Special Issue of Science and Public Policy (2019 Forthcoming). Available at SSRN: <https://ssrn.com/abstract=3413035> accessed 14 November 2019.

55 Jeff Parker and Michael Gregg, *CompTIA Casp + Study Guide*, 3rd edn (John Wiley & Sons 2019), 11.

56 David Basin, Patrick Schaller and Michael Schläpfer, *Applied Information Security: A Hands-on Approach* (Springer 2011), 105.

57 Michael Whitman, Herbert Mattord and Andrew Green, *Guide to Firewalls & VPNs* (Cengage Learning 2011), 271.

58 Robert Newman, *Computer Security: Protecting Digital Resources* (Jones and Bartlett Publishers 2010), 199.

59 Jon Tate et al, *Introduction to Storage Area Networks* 9th edn (IBM Redbooks 2017), 199.

ed, or 'unscrambled' to discover the meaning of the information. This method protects data. However, it leaves a small opening for a hacker to intercept information the moment it is unlocked.[60]

With the HE scheme proposed in this paper, the process will work differently. It will allow the receiver (Bob) to read the information as if it has been decrypted, however, without removing the security layer that the encryption process placed on it. Think of the manager of an organisation who could use the HE scheme to encrypt all customers sensitive data such as: customer records, business intelligence, and trade secrets. HE schemes could keep confidential information safe and even shared without decrypting it.[61]

HE can be defined as 'the conversion of data into ciphertext that can be analysed and worked with as if it were still in its original form.'[62] To put it simple terms, HE is a special form of encryption which allows complex computation and mathematical operations on the ciphertext without decrypting and exposing the encrypted data.[63] The word 'homomorphic' comes from the Greek language and means 'same shape' or 'same form.' In cryptography, this means that the analysis of the message can be performed in the ciphertext in the same way as in the plaintext, without sharing the secret key (and thus without decrypting the data).[64]

According to the European Union Agency for Network and Information Security (ENISA) Guidelines on Privacy by Design in Big Data, HE is an emerging and promising technology in this field with a lot of interest not only for the academic research community but also for the industry. HE would fall under the general category of 'privacy preserving computations.' It is widely recognised that HE solutions would greatly help to solving security and privacy problems in big data and facilitate cloud computing transactions.[65]

The reason why HE is expected to play an important role in cloud computing and big data transformations is that it allows organisations to store encrypted data in public clouds while still reaping all the benefits and full potential of the cloud provider's data analytic tools. As discussed above in the previous section, in conventional approaches, pseudonymisation and anonymisation tools are applied to provide data security and data protection, but at the expense of usability. However, HE techniques allow one to keep data encrypted while using all the computational power to analyse data in the cloud.[66]

Typical privacy techniques, such as masking,[67] generalisation[68] and k-anonymisation,[69] are commonly considered as pseudonymisation methods since they help to preserve privacy. Yet, there is no way to quantify how well preserved. Unlike pseudonymisation, anonymisation techniques such as differential privacy[70] can allow one to quantify how much privacy is being preserved.[71] These methods rely on traditional public key encryption schemes

60 Ales Teska, 'Homomorphic Encryption for GDPR' (15 July 2018) <https://teskalabs.com/blog/personal-data-deindetification-4-homomorphic-encryption-gdpr> accessed 29 November 2019.

61 ibid.

62 Kannan Balasubramanian et al, 'Homomorphic Encryption Schemes: A Survey' in Kannan Balasubramanian and M Rajakani (eds) *Algorithmic Strategies for Solving Complex Problems in Cryptography* (IGI Global 2017), 98.

63 Nirdosh Bhatnagar, *Mathematical Principles of the Internet* (CRC Press 2018), 203-204.

64 ENISA, *Privacy by Design in Big Data: An Overview of Privacy Enhancing Technologies in the Era of Big Data Analytics*, 40 (December 2015) <https://www.enisa.europa.eu/publications/big-data-protection> accessed 29 November 2019.

65 ibid 24, 40 and 41.

66 Jun Sakuma, 'Secure Outsourcing of Data Analysis' in Fei Hu (ed) *Big Data: Storage, Sharing and Security*, (CRC Press 2016), 365.

67 *Data masking* (also known as data obfuscation) is the process by which real data is obscured by random characters and other data. That is, replacing real data with entirely random values. See Khaled El Emam and Luk Arbuckle, *Anonymizing Health Data: Case Studies and Methods To Get You Started* (O'Reilly 2013), 163.

68 *Generalization* is carried out by 'replacing the actual value of the attribute with a less specific, more general value that is faithful to the original.' See Alina Campan and Traian Marius Truta, 'Data and Structural *k*-Anonymity in Social Networks' in Francesco Bonchi et al. (eds) *Privacy, Security, and Trust in KDD* (Springer 2009), 36.

69 *K-Anonymity* is a classical model that 'divide the data in clusters of k or more elements and substitute the elements in each k-cluster with a unique synthetic one, commonly the mean of the elements.' See Flavio Lombardi and Roberto Di Pietro, 'Towards a GPU Cloud: Benefits and Security Issues' in Zaigham Mahmood (ed) *Continued Rise of the Cloud: Advances and Trends in Cloud Computing* (Springer 2014), 10.

70 *Differential privacy* has recently emerged as one of the most robust mechanism which guarantees statistical data release. Differential privacy 'guarantees that an adversary (even with arbitrary background knowledge) learns nothing more about an individual from the released data set, regardless of whether her record is present or absent in the original data.' See Noman Mohammed et al 'Private Genome Data Dissemination' in Aris Gkoulalas-Divanis and Grigorios Loukides (eds) *Medical Data Privacy Handbook* (Springer 2015), 448.

71 Steve Touw, 'Homomorphic Encryption Alone is Security, Not Privacy' (14 September 2018) <https://www.immuta.com/homomorphic-encryption-alone-is-security-not-privacy/> accessed 29 November 2019.

composed of three steps: key generation, encryption and decryption. However, a HE scheme involves a fourth step: the *evaluation* of encrypted data.[72]

To take a simple example, consider a cloud-based application run by a hospital which stores all patient health data in a public cloud. By using the cloud, the hospital does not need to absorb all the costs of buying the underlying cloud infrastructure. Instead it is managed by a third party. The cloud may produce efficiency, timing, and performance improvements.

In the context of biological and scientific research, the OPTIMIS project[73] ran a use case scenario which demonstrated how this could be implemented in a hospital. This cloud-based application was part of the programming model of a toolkit. The toolkit performed the processing and data analytic techniques in genomic sequencing applications. For instance, consider the situation of various hospitals and research organisations that want to make use of secondary data to detect the DNA of a specific disease. The genomic application implemented a workflow which performed automatic gene detection taking into account a specific genome analysis (Genewise) which was able to detect the gene patterns. The whole dataflow process involved a complex network of databases run first by the hospitals and then in the cloud-based application. The result of this analysis was based on algorithms that could predict accurate gene structures.[74]

The problem that hospitals face, however, is that in order to keep patients' personal data secure, they have to encrypt the data. This prevents the data from being disclosed or accessed by a malicious outsider or by the same hospital personnel. Nevertheless, when doctors at the hospital need to use the patients' data, they would need to either transfer the encrypted data to a trusted environment – such as a private cloud, or decrypt the data on the public cloud, use the data, then encrypt it again before storage. This makes the entire process cumbersome and exposes the data to malicious activity.

Encryption schemes that support operations on encrypted data have a very wide range of applications in cryptography. The purpose of HE is to allow computation on encrypted data. Thus, data can remain confidential while it is processed, enabling useful tasks to be accomplished with data residing in untrusted environments.[75] In other words, HE is a technique of performing secure arbitrary computations on the ciphertext without revealing the plaintext.[76]

The properties of HE could have valuable application in enabling pharmaceuticals and healthcare organizations to search their data privately in an encrypted domain. HE could be used in a private cloud medical records storage system (eg, Patient Controlled Encryption), in which all data for a patient's medical record is encrypted by the healthcare providers before being uploaded to the patient's record in the cloud storage system.[77]

According to the Art. 29 Working Party, there are various anonymisation and pseudonymisation techniques with different degrees of robustness. One of such techniques is 'noise' addition. Noise addition is a complementary measure that makes it more difficult for an attacker to retrieve the personal data.[78] It is especially useful when 'attributes may have an important adverse effect on individuals and consists of modifying attributes in the dataset such that they are less accurate whilst retaining the overall distribution.'[79] For example, if the height of the data subject was originally measured to the nearest centimetre, the anonymised dataset may contain a height accurate to only + - 10 cm.[80]

An important element and one of the issues hindering efficiency in HE operations is the 'noise.' Noise takes place every time an operation in the ciphertext

72 Antoine Guellier, 'Can Homomorphic Cryptography Ensure Privacy?' 40 (July 2014) <https://hal.inria.fr/hal-01052509v2/document> accessed 29 November 2019.

73 Optimized Infrastructure Services (OPTIMIS) was an EU funded project within the 7th Framework Program under contract ICT-257115. The OPTIMIS consortium included different European cloud providers as well as other universities and research institutions such as Atos Origin (Spain), Umea University (Sweden), The 451 Group (UK), Universität Stuttgart (Germany), ICCS (Greece), Barcelona Supercomputing Center (Spain), SAP (UK), Fraunhofer-Gesellschaft (Germany), University of Leeds (UK), Leibniz Universtät Hannover (Germany), Flexiant (UK), BT Group (UK) and City University London (UK).

74 Marcelo Corrales Compagnucci, *Big Data, Databases and 'Ownership' Rights in the Cloud* (Springer 2019), 231.

75 Frederik Armknecht et al, 'A Guide to Fully Homomorphic Encryption', 1 <https://eprint.iacr.org/2015/1192.pdf> accessed 5 November 2019.

76 Srinivas Divya Papisetty, 'Homomorphic Encryption: Working and Analytical Assessment' MSc Thesis, 21 (2017) <http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1082551&dswid=-318> accessed 5 November 2019.

77 ibid.

78 Opinion 05/2014 on Anonymisation Techniques adopted on 10 April 2014, 3, 12.

79 ibid 12.

80 ibid.

is performed. Although there are noise-free HE schemes available today,[81] noise in encryption schemes is added to ensure semantic security of the cryptosystems. It is basically a set of small terms[82] added exponentially inside the ciphertext while encrypting data. This process makes the noise grow as HE operations proceed.[83]

Noise in HE may be viewed as a bounded element of randomness in the scheme added to the message in some way. In mathematical terms, this means that small terms are purposely added in a way that it is easy to remove and unveil the original message for trusted third parties holding some extra information.[84] If this technique is applied correctly, a third party will not be able to identify nor detect how the data have been modified.[85]

HE operations usually increase the noise by default. The more a ciphertext is processed, the more noise is added to it. However, due to its inherent complexity, the problem with noise is that it slows down the computer processing speed and this has become a major bottleneck for the overall efficiency of HE operations. The risk is that if too much noise is added, the ciphertext will no longer be correctly decrypted, making it worthless. How to decide if the noise is too much or not depends on the security and correctness properties of each system.[86] However, if the randomness added to the message were to exceed the given bound, decryption might not be correct.[87]

In sum, noise may be viewed as a 'somewhat brittle form of randomness: if too much noise is used to cover a message, it might collapse and make it impossible to recover the said message.'[88] This means that a message may have several different forms of encryption and the level of noise should be kept as a minimum. Therefore, the objective of this research paper is to develop a searchable encryption tool that can allow data analysts to pseudonymise and search data in the encrypted domain using the HE technique and at the same time reducing the amount of noise.

In order to reduce the noise, the authors describe an RSA use case example. As explained above, RSA is one of the first and most popular public-key cryptosystems, which is used widely for secure data transmission. An encryption scheme using noise in its encryption will be presented, and also the problems this may cause if the scheme is to be fully homomorphic.[89]

In all HE schemes, ciphertexts contain noise that grows during homomorphic evaluation operations. In practice, managing the noise to ensure it is always below the threshold can be done in two ways:[90]

a) Using the bootstrapping procedure. Bootstrapping in computer science means that the computer system improves in increments by itself. In the context of HE, it means that it takes an input ciphertext with a large amount of noise, and outputs a new ciphertext which has less noise and can be used for further computation. Therefore, by bootstrapping at appropriate points, the entire evaluation can be performed.[91]
b) Pre-determining the function to be evaluated and then setting the specific parameters to allow for the noise growth to occur in a controlled manner. By using this method, we ensure that the output ciphertext at the end of the evaluation will have noise below the threshold. Thus, no bootstrapping will be necessary and correct decryption is ensured.[92]

In either case, good understanding of the noise growth behaviour is essential to achieve correctness and optimal performance.[93] The authors are current-

81 Jing Li and Licheng Wang, 'Noiseless Fully Homomorphic Encryption' <https://eprint.iacr.org/2017/839.pdf> accessed 29 November 2019.

82 These terms may be based on 'integers' (if the scheme is based on 'integers'). An 'integer' in mathematics is a whole number that can be written without a fractional component. For example, 10, 21, 4, etc. Or, these terms may be based on 'polynomials' (if the scheme is based on 'polynomials'). A 'polynomial' is an expression that can contain variables and coefficients. For example, a 'polynomial' of a single indeterminate, x, is x2 – 4x + 7.

83 David Archer, 'Ramparts: A Programmer-Friendly System for Building Homomorphic Encryption Applications' <https://eprint.iacr.org/2019/988.pdf> accessed 29 November 2019.

84 Martha Norberg Hovd et al, 'The Handling of Noise and Security of Two Fully Homomorphic Encryption Schemes' Master Thesis, Department of Mathematical Sciences (Norwegian University of Science and Technology 2017), 4 <https://pdfs.semanticscholar.org/4f46/1e38c25dbbeb4b8de170c6a736372ad368f2.pdf> accessed 5 November 2019.

85 Opinion 05/2014 on Anonymisation Techniques adopted on 10 April 2014, 12.

86 For instance, a polynomial is typically considered small if all its coefficients are small.

87 Martha Norberg Hovd et al, (n 84).

88 ibid.

89 ibid.

90 Anamaria Costache, Kim Laine and Rachel Player, 'Homomorphic Noise Growth in Practice: Comparing BGV and FV, 2 <https://eprint.iacr.org/2019/493.pdf> accessed 5 November 2019.

91 ibid.

92 ibid.

93 ibid.

ly working on a Java programming language[94] based on a Paillier implementation scheme.

A Paillier scheme is a probabilistic asymmetric algorithm for public key cryptography. Probabilistic encryption refers to the use of randomness in an encryption algorithm by introducing an element of chance. For example, a simple message like 'hello world' will not always correspond to the same ciphertext. Instead, that message would yield different ciphertexts each time it is encrypted. In other words, an encryption technique is probabilistic if the same cleartext can encrypt to many different ciphertexts.[95]

Although the Paillier scheme is not fully homomorphic, it is a good source to use as it allows flexibility to modify and implement for other requirements. The scheme is a perfect example of HE.[96] The authors will work on a modified and extended version to the scheme in order to support more computations. This approach will be more suitable for data protection and data security in healthcare and pharmaceutical cloud-based applications. The HE component of the EnergyShield toolkit has been tested already using a Java user interface which connects to a database. Further implementations of the HE searchable data component will be carried on the MySQL database[97] on a later stage.

V. Conclusions

In light of the exponential growth of cyber-attacks in the pharmaceutical and healthcare sector, various conventional approaches are used to protect personal and sensitive data. These approaches include advanced anonymisation and pseudonymisation techniques. Nevertheless, anonymisation reduces the quality of data analytics and pseudonymisation needs the data to be decrypted for further analyses, which exposes the data to potential cyber-attacks.

Data security and data analytics are frequently seen as two conflicting areas. However, HE tools allow the possibility to query and analyse encrypted data without ever decrypting and compromising it. Therefore, we suggest raising the threshold of data protection and data security by using enhanced cryptographic techniques such as the proposed searchable HE tool. This new framework could be used effectively by pharmaceutical and healthcare organizations to reduce the data security threats, while at the same allowing them to time tap into the power of big data analytics and to comply with regulatory frameworks that require more data transparency, data portability and utility.[98]

Yet, the current problem with HE operations is the amount of 'noise' they create which slows down computing power. The risk is that if the 'noise' is greater than a certain value, the decryption function does not work. The major challenge from a data protection and data security perspective is to be able to generate the right amount of noise, so as to protect data subject's privacy (and potentially also commercially confidential information) while preserving the usefulness of the released responses. The proposed HE framework attempts to provide a high level of security with the proper amount of noise. This new framework will aid data analysts to glean more data and search for more meaningful pattern results, while mitigating data security risks and appeasing privacy concerns.

The balancing act that is required is highly delicate and it is important to early identify and proactively address new challenges, such as the vast promises and risks of quantum computing technologies and data fusing that could enable decryption and data re-identification even of HE secured data.[99] An

94 Java language is 'a general-purpose programming language that is class-based, concurrent, object-oriented, and designed to have as few implementation dependencies as possible.' See Jeffrey Strickland, *Simulation Conceptual Modeling* (Lulu.com 2011), 105.

95 Gregory Neven, Frank Piessens and Bart De Decker, 'On the Practical Feasibility of Secure Distributed Computing: A Case Study' in Sihan Qing and Jan Eloff (eds) *Information Security for Global Information Infrastructures* (Springer 2010), 366.

96 In this project, we are using the following algorithms for the elaboration of the HE searchable tool. Key-Gen Algorithm: 1) Choose two prime numbers p & q and calculate n=p*q and λ = lcm(p-1,q-1) such that gcd(p*q,(p-1)*(q-1)) = 1; 2) Select g ∈ Z*n^2 and calculate μ = (L(g^λ mod n^2))^(-1) mod n where L(x) = x-1/n; 3) n, g acts as a public key; 4) λ, μ acts as a private key. Encryption Algorithm: 1) Let m ∈ Zn be the message; 2) Choose r ∈ Z*n; 3) Required Cipher text is c = g^m *r^n mod n^2. Decryption Algorithm: 1) Compute m = L(c^λ mod n^2)*μ mod n.

97 MySQL database is an open source Relational Database Management System which uses Structured Query Language (SQL). SQL is the most widely used language for managing the content of a database. See, generally, Matthew Stucky, *MySQL: Building Using Interfaces* (New Riders Publishing 2002); Russell Dyer, *MySQL in a Nutshell: A Desktop Quick Reference* (O'Reilly 2005).

98 Cf Minssen, Rajam and Bogers, (n 54).

99 Alevtina Dubovitskaya et al, 'Intelligent Health Care Data Management Using Blockchain: Current Limitation and Future Research Agenda' in Vijay Gadepally et al, (eds) *Heterogeneous Data Management, Polystores, and Analytics for Healthcare* (Poly 2019). Lecture Notes in Computer Science, vol 11721 <https://link.springer.com/chapter/10.1007/978-3-030-33752-0_20> accessed 10 November 2019.

impenetrable privacy protection might turn out to be an illusion. It is therefore crucial to not only improve the encryption methods, but also to strengthen the legal frameworks for effective remedies, harsh damages and severe penalties that should be available and enforceable if breaches occur. At the same time, however, it should neither be forgotten that the future of healthcare and drug development is increasingly depending on effective cross-border data sharing. Such improvements can have life-saving effects on a global scale. Hence, it is equally important that the technological possibilities are not unduly hampered by misguided implementations or overly strict interpretations of data protection. While the perfect balance between data utility, portability, transparency and protection might never be achieved, well-informed research that works across interdisciplinary boundaries should strive to help minimise the range within which the scale will surely continue to swing.

Acknowledgement: This research is supported by a Novo Nordisk Foundation grant for a scientifically independent Collaborative Research Program in Biomedical Innovation Law (grant agreement number NNF17SA0027784) and by the EU EnergyShield project under the H2020 research and innovation programme (Grant Agreement No. 832907).