© University Press 2022
ISSN 2719-6550
ISSN 2719-7417 online

"Journal of Education, Technology and Computer Science"
No. 3(33)/2022
www.eti.ur.edu.pl

Received: 2.09.2022
Accepted for printing: 10.11.2022
Published: 22.12.2022
License: CC BY-SA 4.0

DOI: 10.15584/jetacomps.2022.3.14

CHRISTINE HILCENKO[1,2,3], **TARA TAUBMAN-BASSIRIAN**[4]

Medical Data, Reconciling Research and Data Protection

[1] ORCID: 0000-0002-9596-7833, Ph.D., Cambridge Institute for Medical Research, Cambridge, CB2 0XY, UK.
[2] Department of Haematology, University of Cambridge, Cambridge, CB2 0XY, UK
[3] Wellcome Trust-Medical Research Council Stem Cell Institute, University of Cambridge, Cambridge, UK
[4] https://www.datarainbow.eu

Abstract

Most of our medical records are being processed electronically, centralised and easily accessible. In this paper, we will discuss the advantages of the system for research, as well as its potential challenges.

Keywords: health data, privacy, data protection, GDPR, security, information, research, EMR, EHR, European Data Hub

I. The purpose and benefits of digital health data

1. Electronic Health Record (EHR) or Electronic Medical Record (EMR)

An electronic medical record (EMR) is a digital version of a chart with patient information stored in a computer and an electronic health record (EHR) is the systematised collection of patient and population electronically stored health information in a digital format. These records are shared through different countries and different medical departments. An EHR is a digital version of a patient's paper chart. EHRs are real-time, patient-centred records that make information available instantly and securely to authorised users. Hand-written data is replaced with electronic records to maintain the constant flow of patient-related data. Introduced in 2009, Today, over 80% of points of medical care are using this technology for billing and entry point of medical information. EHRs are built to easily share information with other health care providers and organisations – such as laboratories, specialists, medical imaging facilities, pharmacies allowing better diagnostics and patient outcomes based on the patient's medical history. It also improves patient participation, care coordination, and is cost saving. For instance, EHR alerts can be used to notify providers when a patient has visited the hospital, allowing them to proactively follow up with the patient. Every health provider can have the same accurate and up-to-date information about a patient. This is especially important with patients who are seeing multiple specialists, receiving treatment in an emergency setting or making transitions between care settings.

It can reduce medical errors and unnecessary tests and reduce the chance that one specialist will not know about an unrelated (but relevant) condition being managed by another specialist. An EHR not only keeps a record of a patient's medications or allergies, it also automatically checks for problems whenever a new medication is prescribed and alerts the clinician to potential conflicts. It can help providers quickly and systematically identify and correct operational problems. In a paper-based setting, identifying such problems is much more difficult, and correcting them can take years.

2. Individual patients' electronic records processing for wider scientific research projects

In data protection terms when data collected for one purpose is then deployed for a new, unrelated, purpose, we are engaging in secondary processing.

The secondary processing of personal health data for scientific research in the medical field is coveted for research and new treatments that improve public health. Its legal status, however, is far from unproblematic. An unclear status in law of data reuse is compounded by a shift in research towards decentralised clinical trials: deployment of new artificial intelligence technologies is leading to a rethinking of biomedical research from what was traditionally organised. This shift is all the more important as classical research tends to be questioned following various health scandals. Fewer people are willing to participate in medical trials. In digital research, AI tools are preferably deployed in the exploitation of databases to explore patient records, via medical imaging or connected medical devices. The analysis of the various data sets with more efficient computing capabilities makes it possible to increase the speed of discoveries, acting as an accelerator of research. Clinical procedures are simplified, time is spared, and, above all, the physical integrity of individuals is preserved. This shift, for instance, has allowed discovery of an antibiotic molecule capable of bypassing antibiotic resistance in February 2020 (Internet 1; Internet 2).

The primary advantage of digitalising research data is the increased ease of participation in research. Many studies struggle to recruit and retain sufficient participants. Less and less patients take part in clinical trials. A survey by CISCRP

(Internet 3) indicates that physical distance to research sites is one of the main barriers to increased participation.

The second advantage is the amount of data from a larger number of participants that a decentralized study can generate. Finally, decentralized clinical trials attract a more inclusive variety of participants. If the advantages are obvious, they are not inconvenient. Apart from the digital divide excluding the populations with less access to technology, data quality, security and strong medical ethics are required. Big data analytics, based on larger data sets, using tools or algorithms are big assets for medical and scientific research. Paolo and Bincoletto (2021) examine the secondary processing of personal health data for scientific research in the medical field. Looking at the controllers' obligations to comply with the data protection framework to safeguard fundamental rights and freedoms. After comparing the implementation of EU regulation into the French and Italian national legislations, they propose "a proactive, legal-technical e-health solution that complies with the rules and principles of the legal frameworks and empowers the individual's control over personal health data while promoting medical research. To this end, the data protection by design concept plays a central role, and an interdisciplinary approach is fundamental in combining legal and technical perspectives".

The case of Yoti illustrates the potentials for research and dangers for the data (Internet 4)

For many years, immigration departments have struggled with physical determination of the age immigrants claiming to be minor. Such determination by physical examination have never been accurate. The algorithm the company Yoti has created claims to have achieved the goal with high accuracy. For doing so, a large world wide database of children aged between 12 and 19 years old. Has be created. Admittedly some parents have received financial compensation in exchange of pictures of their children with month and year of birth. By creating a data set of children of various world ethnic faces, the algorithm can determine the face of a minor. So far, the software has been commercialised in supermarkets for the sale of alcohol or cigarettes. They are expanding with cinemas and other places where age verification is necessary. As mentioned, the software can be useful for immigration departments. Scientific research could certainly benefit from such database to study the ageing process in various ethnic groups. The dilemma here is how to create and store large scale biometrics data allowing a sensible secondary use of the information.

3. Big data and scientific research friend or foe

As indicated above, scientific research benefits from large scale data from various sources. This has been made easier and more cost effective thanks to digitalisation of information and cheap data storage. Big Data is defined by

Zulkarnain and Anshari (2016) as “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze”. In Kaislere, Armour and Espinosa (2016), “Big data is data too big to be handled and analyzed bytraditional database protocols such as SQL”. Big Data are particularly large, complex datasets with high analytical potentials that automated processing can analyse at higher speed than traditional processing.

“Big data analytics in medicine and healthcare allows analysis of the large datasets from thousands of patients, identifying clusters and correlation between datasets.Moreover, improving predictive models using data mining techniques”.

Big Data has large potentials helping medical research to create new growth opportunities including in predictive healthcare. This is not without posing significant challenges such as the loss of privacy and confidentiality. Privacy and security are key concerns for individuals and all corporations involved.

However, open access to health data is beneficial to scientific research. We will further discuss the European initiatives. The following sections will draw data protection lessons from several European initiatives.

4. The requirement of quality data an essential parameter to integrate

Data accuracy is paramount for quality research results. Human or algorithm errors/biases or sometimes technology outcomes can alter results. This issue has been pointed out by the European Union Agency for Fundamental Rights (FRA, 2020) that raised serious concerns about the quality of medical data and the resulting risk of medical errors. “The quality of data in EMR/EHR also raises some concern. Studies where patients were shown their medical files and asked about their accuracy found that up to 50 % of information was incomplete or erroneous. Too many important data in EMR/EHR is unstructured in the form of free text, which further reduces data quality”. See also Miller (2019).

In order to be validly processed, row medical data requires to undergo initial “cleansing” (Stöger, Schneeberger, Kieseberg, Holzinger, 2021). In this paper, the authors take “an interdisciplinary look at some of the technical and legal challenges of data cleansing against the background of the European medical device law, with the key message that technical and legal aspects must always be considered together in such a sensitive context”.

These authors initially enumerate the typical data quality issues they suggest the cleansing operation should tackle: absence of data, dummy/default, noise (*a.k.a.* the “butterfly effect”, wrong data, inconsistent data, cryptic data, duplicate primary keys, non-unique identifiers, multipurpose fields, violation of (business) rules.

With reference to the ECJ and the German Supreme Court (Bundesgerichtshof) there are potential legal consequences for faulty data in medical AI, engaging the liability of the notified body towards third persons like patients in case that an assessment procedure has been carried out without sufficient diligence (Case C-219/15 *Elisabeth Schmitt v TÜV Rheinland LGA Products GmbH* ECLI:EU:C:2017:128; BGH 17 February 2020, VII ZR 151/18).

The authors then suggest five necessary steps to use the raw data and transforms it into information that can be worked with in the subsequent analytical steps:

Parsing, Correcting, Standardizing, Matching, Consolidating.

Without going into further detail into these techniques, they suggest applying standardisation and corrections to data sets sometimes from unstructured information like manual diagnosis data to then compare and match to avoid duplication to finally solve inconsistencies. The quality of this cleansing process will have legal consequences as they point that any damage caused by malfunctions or other service provider due to these "faulty" data, will create a liability for the manufacturer of the medical device mostly under the Medical Devices Regulation (MDR). It is required that the product in question "shall be safe and effective and shall not compromise the clinical condition or the safety of patients". GDPR data accuracy requirement will not apply to the training data as data is mostly anonymised.

This has led the European Parliament to propose that developers should, where feasible, implement quality checks of the external sources of data and should put oversight mechanisms in place regarding their collection, storage, processing, and use of data (Art. 17 para. 3 European Parliament, *Framework of Ethical Aspects of Artificial Intelligence* – n 46).

Also, in its report on the safety and liability implications of Artificial Intelligence, the Internet of Things and Robotics, the European Commission has noted that Union product safety legislation now "does not explicitly address the risks to safety derived from faulty data. However, according to the 'use' of the product, producers [in terms of the MDR, manufacturers] should anticipate during the design and testing phases the data accuracy and its relevance for safety functions" (Commission, '*Report on the Safety and Liability Implications of Artificial Intelligence*' – n 18). The Report and the Whitepaper also call for product safety legislation to "provide for specific requirements addressing the risks to safety of faulty data at the design stage as well as mechanisms to ensure that quality of data is maintained throughout the use of the AI products and systems" (Commission, COM (2020) 65 final – n 3 – 15).

5. When Clinical Trials Go Digital, Ethics Is Needed

The pandemic accelerated the digitalisation of clinical research and widespread decentralised clinical trials have brought many improvements while privacy and ethics challenges took a back seat.

In June 2021, the World Health Organization published a guidance on *Ethics & Governance of Artificial Intelligence for Health* (Internet 5) outlining six key principles for the ethical use of artificial intelligence in health. The 20 leading experts in ethics, digital technology, law, human rights, as well as experts from Ministries of Healthrecommended the technology must put ethics and human rights at the heart of its design, deployment, and use.

Artificial intelligence's potentials in healthcare are undeniable: from AI medical imaging, scanning patient health records to predict illness, monitoring devices to systems that help track disease outbreaks. Accessing medical speciality consultations to help evaluate symptoms in remote areas. Just to mention a few (Wetsman, 2019). The outbreak of the COVID pandemic also saw the surge of technological tools that were revealed to be controversial if not unnecessary tracking. The Bluetooth technology widely used was inappropriate to detect the contact tracing. In Singapore, the government admitted that the contact tracing application collected data about health that was repurposed beyond the original goal.

Additionally, some databases were hacked, and medical data were accessed by unlawful criminals (Internet 6). A report by WHO *Ethics and governance of artificial intelligence for health* warned about AI tools developed by private technology companies (like Google and Chinese company Tencent) that have large resources but not always the necessary ethical incentives. Their focus may be toward profit, rather than the public good. "While these companies may offer innovative approaches, there is concern that they might eventually exercise too much power in relation to governments, providers and patients" (Internet 7).

The report recommended six ethical principles: protect autonomy, promote human safety, ensure transparency, foster accountability, ensure equity, and promote AI that is sustainable.

In parallel, the new European regulation on AI, the AI Act (Internet 8) which could come into effect in late 2024 is a proposed European law on artificial intelligence (AI) – the first law on AI by a major regulator anywhere. The law assigns applications of AI to three risk categories (Internet 9). The Artificial Intelligence in Healthcare report (Internet 10) supports the European Commission in identifying and addressing any issues that might be hindering the wider adoption of AI technologies in the healthcare sector. The study has highlighted six categories where the European Commission is suggested to focus to support the development and adoption of AI technologies in the healthcare sector across the EU. These include:

1) a policy and legal framework supporting the further development and adoption of AI aimed at the healthcare sector in particular;

2) initiatives supporting further investment in the area;

3) actions and initiatives that will enable the access, use and exchange of healthcare data with a view to using AI;

4) initiatives to upskill healthcare professionals and to educate AI developers on current clinical practices and needs;

5) actions addressing culture issues and building trust in the use of AI in the healthcare sector;

6) policies supporting the translation of research into clinical practice.

The GDPR allows for code of conduct to be approved by national supervisory authorities. These can strengthen clinical research and pharmacovigilance.

The Spanish Code of Conduct (Internet 11)

A first national Code of Conduct was promoted by Farmaindustria in Spain regulating the processing of personal data in the field of clinical trials and other clinical research and pharmacovigilance. The code of conduct, regulates how the promoters of clinical studies with medicines and the CROs that decide to adhere thereto must apply the data protection regulations. Data controllers and data processors that adhere to the code of conduct are obliged to comply with its provisions.

The UK Code of conduct for data-driven health and care tech

Following a consultation, the UK Government has published a code of conduct for data-driven health and care technology to enable the development and adoption of safe, ethical and effective data-driven health and care technologies: "We have some truly remarkable data-driven innovations, apps, clinical decision support tools supported by intelligent algorithms, and the widespread adoption of electronic health records. In parallel, we are seeing advancements in technology and in particular artificial intelligence techniques" (Internet 12).

The Government has set out the behaviours expected from those developing, deploying and using data-driven technologies in the health and care systems:

1. Understand users, their needs and the context.

2. Define the outcome and how the technology will contribute to it.

3. Use data that is in line with appropriate guidelines for the purpose for which it is being used.

4. Be fair, transparent and accountable about what data is being used.

5. Make use of open standards.

6. Be transparent about the limitations of the data used and algorithms deployed.

7. Show what type of algorithm is being developed or deployed, the ethical examination of how the data is used, how its performance will be validated and how it will be integrated into health and care provision.

8. Generate evidence of effectiveness for the intended use and value for money.

9. Make security integral to the design.

10. Define the commercial strategy and consider only entering into commercial terms in which the benefits of the partnerships between technology companies and health and care providers are shared fairly.

6. The case of the French Health Data Hub and the COVID research

The French government has contracted the processing of health data with Microsoft, a US corporation. Microsoft is said to have the processing capabilities currently no other European company could have. Scientific research needs data on COVID patients to study the spread of the virus, the scale of long term COVID, the category of the population mostly affected or the effects of the various vaccines administered.

Several legal limitations had to be considered. First the question of the transfer of European data outside the European Union. United States being deemed a country of non-adequate data protection since the invalidation of the Privacy Shield agreement allowing the flow of data between the two continents. The French data Protection Supervisory authority, The *Commission Nationale Informatique et Libertés* (CNIL) published an opinion expressing their concerns. They recommended this measure to be only temporary, requiring further guarantees from Microsoft. The French Health Minister requested the data to be stored in data servers located within the EU. A measure that has limited protection as regardless of the localisation of the data, US NSA Section 702, The US Cloud Act and the Executive Order 12333 would apply.

If these are technical necessities to comply with the General Data Protection Regulation, basically requiring the same level of data protection wherever data travels, sharing data with a foreign company represents further challenges.

This is a consequence of the loss of data sovereignty that the European countries are facing. Dependence on foreign corporations for processing health data is by itself an issue. The quasi-monopolistic position of these corporations gives them excessive power of control over data. Data is processed, possibly monetised. The security of the data can be compromised as any database created is data at risk.

The GDPR has several principles applying to lawful data processing, from data minimisation, access limitation or ensuring data integrity. Big data analytics or algorithms are by essence thirsty for data. Scientific research benefits from the largest data sets. How to conciliate the legitimate needs of research with data protection principles?

7. Technical means of protecting medical data: encryption, anonymisation or pseudonymisation

With data becoming increasingly digitalised and widely accessible, securing health data is paramount. The first supplementary measure to secure data when transferred abroad is to apply a strong encryption, keeping the encryption key out of the reach of the US internet communication services. This is in fact difficult to realise when data needs further processing.Cybersecurity attacks and ransomware against hospitals are becoming a common threat. Health data have high value and are regularly targeted by cyber criminals. Cyberattacks in the healthcare environment have tripled since 2018 to reach 45 million individual victims in 2021.

Medical data is collected from various sources:

- imaging techniques for diagnosis;
- electronic health records;
- robotics in surgical procedures;
- telehealth for efficiency or reaching patients in more remote locations;
- wearables to monitor individuals' health with various Internet connected devices with often weak security.

The use of open data sources is also instrumental in the field of genomics, where data related to genetic makeup, biomarkers and bioinformatics is used to derive better therapeutic solutions.

European healthcare requires stronger protection and security measures to protect health data.

Pseudonymised personal data are data where identifiers such as names are replaced by codes the research institutions keep. A 'code key' that is the link to the individual person is kept separately from the research data in order to protect the privacy of patients. Only fully anonymised data escape from the General Data Protection Regulation (GDPR) requirements. In the case of data for research, full anonymisation might render the data useless. Therefore, it is often not a possible option.

Pseudonymised data and even anonymised data are not exempt from re-identification. Even when name, date of birth or national security numbers are "anonymised", a full health history will reveal patients' age, gender, the places where they have lived, their family relationships and aspects of their lifestyle.

Privacy-enhancing technologies such as homomorphic encryption, differential privacy, federated analyses and use of synthetic data offer new ways for protecting the privacy of individuals. New promises are seen in synthetic data (Internet 13), still not exempt of criticism (Chen, Lu, Chen, Williamson, Mahmood, 2021). Issues of anonymisation and de-identification need to be addressed and appropriately managed (Internet 14).

Health data being more sensitive, requires extra layers of protection and appropriate security measures such as encryption. The GDPR applicability to the data set has implications in the free flow of data to countries outside the European Economic Area (EEA. This is a problem, for example, with researchers at federal research institutions in the United States. Transfers to international organisations such as the World Health Organization are similarly affected (Internet 15). The European scientific academies have recently published a report explaining the consequences of stalled data transfers and pushing for responsible solutions. (The European Academies Science Advisory Council, the Federation of European Academies of Medicine & the European Federation of Academies of Sciences and Humanities (2021) (Internet 16; Bentzen et al., 2021).

The case of the Canadian PHIPA experience

Occasionally, there are privacy investigation decisions that stand out because of their precedent setting nature (Internet 17). One example is the PHIPA Decision 175 (Internet 18) which details an investigation into the sale of de-identified data by a health information entity to a third-party corporation. The data protection supervisory became aware of the situation through a news article and launched an investigation under the Personal Health Information Protection Act (PHIPA). It is admitted that de-identification is considered to be a use under PHIPA.

Such use for de-identification does not usually necessitate individual consents. Health information custodians, who have custody and control of personal health information have to be transparent to clearly and explicitly inform patients about their practices in their public notice. The necessity of transparency is emphasised. It is safe to consider de-identification to be considered data processing under the GDPR definition. Although, patients' consent is required in most cases. We will now take a deeper dive into the specific situation of the European health data and its challenges. How to reconcile the need to share data, open free access to the database, and keep data secure.

II – The paradigm of the European data sovereignty

The last few decades, we have witnessed major developments in the fields of internet communication and data digitalisation. Europe has dragged behind US or Chinese corporations. Our dependency on US corporations for cloud storage (AWS, Microsoft, Apple iCloud, etc…) or computer operating systems (Microsoft, Apple iOs or Google) or data analytics is undeniable. It's with the COVID pandemic that EU governments have had a wakeup call, realising the complexities of creating a contact tracing app without the help of a Google App or its ID verification Captcha. The de-centralised tracing application could not be successful. Since the European Court of Justice decision in July 2020 invalidating data transfers to the US, the use of US internet communication services became problematic. European countries are re-thinking data processing and data storage solutions including in the fields of health data and research.

1. European health data initiatives promise to open new perspectives for medical research

A – EU Health Data

New initiatives are underway for the EU Health data centre and common health data to be accessed more efficiently. "Early lessons learnt with COVID-19 have shown that the current system has not ensured an optimal response at EU level to the COVID-19 pandemic".

"Regarding health data, its availability and comparability, the Covid-19 pandemic revealed that the EU has no clear health data architecture. The lack of harmonisation in these practices and the absence of an EU-level centre for data analysis and use to support a better response to public health crises is the focus of this study. Through extensive desk review, interviews with key actors, and enquiry into experiences from outside the EU/EEA area, this study highlights that the EU must have the capacity to use data very effectively in order to make data-supported public health policy proposals and inform political decisions. The possible functions and characteristics of an EU health data centre are outlined. The centre can only fulfil its mandate if it has the power and competency to influence Member State public-health-relevant data ecosystems and institutionally link with their national level actors. The institutional structure, its possible activities and in particular its usage of advanced technologies such as AI are examined in detail".

Study by Henrique Martins of ISCTE-Lisbon University Institute and Faculty of Medical Sciences, UBI Portugal, was made public at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament is staggering (Internet 19).

Sadly, the finding of this study is striking : the absence of harmonisation despite new initiatives such as the 'European Health Union' umbrella, or the EC proposal for the creation of a new European Health Emergency. "There is NO comprehensive health data governance at the EU level, and very few MS [Member States] could be said to have one at the national level as well while a centralised governance structure is needed to deal with large scale data for preventive and curative actions".

The suggested model could be Preparedness and Response Authority (HERA) (Internet 20). On 3 May 2022, the Commission published the Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space Act ("the Proposal").

European Data Protection Board (EDPB) and the European Data Protection Supervisory (EDPS) were required to give their opinions on this project (Internet 21)

Since the European Data Strategy for the creation of a single European Data Space for primary use of medical data, all EU citizens will have access to their electronic health records by 2030 thanks to the EU's central eHealth platform linking national contact points to the MyHealth@EU infrastructure and efficient national digital health authorities.

B – Joint Action Towards the European Health Data Space – TEHDAS

The European Health Data Space (EHDS) could enormously impact health research if it can overcome barriers to cross-border secondary use of health data and create trust amongst citizens, according to a Finnish health data stakeholder (Internet 22). The TEHDAS project, based on the European Commission's Health Programme 2020, develops European principles for the secondary use of health data from early 2020. Carried out by 25 European countries and co-ordinated by the Finnish Innovation Fund, Sitra (Internet 23).

European health projects will ease the share of data flow, with standardisation and the creation of a central authority. Free access to this data will benefit innovation at the European and international level.

III. Access to data for research

Re-using data and open science:

1- International cooperation

1.1 – Within the National and international initiatives supporting medical data health is the International Medical Informatics Association (IMIA) Open Source Working Group (OSWG) (Internet 24), "a voluntary group supported by IMIA that brings together researchers and practitioners from multiple countries with a diverse range of informatics experience but common interest in the adoption of open approaches to advancing the use of informatics to improve healthcare". This has led to the development of an open access database of Free, Libre, and Open-Source Software (FLOSS), called MedFLOSS (Internet 25) (www.medfloss.org), to apply in the medical domain to accelerate medical research.

1.2 – Harmony Alliancepromise to accelerate scientific research in Haematology

HARMONY claims to be the Next-generation science, sharing data and knowledge "To accelerate scientific research on Hematologic Malignancies (HMs), broad cooperation between all healthcare stakeholders is urgently needed. Each single piece of the puzzle, namely each set of data, might turn out to be the missing piece that will make it possible to treat the disease faster and better.

The HARMONY Alliance (Internet 26) has created a transparent and secure repository for data from various clinical studies. Everybody can contribute and help fight HMs".

"Sharing data is crucial to further research in Hematological Malignancies because many of these cancers are heterogeneous in terms of their genetic profile. Moreover, many genetic profiles are rare; many account for < 10% cases. In order to ascertain the prognostic or predictive value of genetic biomarkers, researchers need to study large cohorts of patients. This is only achievable by sharing data," says Anthony Moorman, Newcastle University, one of the HARMONY Key

Opinion Leaders for Childhood Leukemia and Project Leader of the HARMONY Research Project entitled ‘Use of Big Data to improve outcomes for patients with Acute Lymphoblastic Leukemia (ALL) (Internet 27).

A – Call to Remove obstacles to sharing health data with researchers outside of the European Union (Internet 28)

Scientific academies in Europe (the European Academies Science Advisory Council, the Federation of European Academies of Medicine, and the European Federation of Academies of Sciences and Humanities) [The European Academies Science Advisory Council, the Federation of European Academies of Medicine & the European Federation of Academies of Sciences and Humanities (Internet 29) have joined forces to call attention to the challenges that affect not only European scientists but collaborators worldwide.

In this paper (Internet 30), the co-authors,Bentzen et al. develop the necessity to levy data sharing barriers: “COVID-19 has shown that international collaborations and global data sharing are essential for health research, but legal obstacles are preventing data sharing for non–pandemic-related research among public researchers across the world, with potentially damaging effects for citizens and patients. International sharing of pseudonymized personal data among researchers is key to the advancement of health research and is an essential prerequisite for studies ofrare diseases or subgroups of common diseases to obtain adequate statistical power”.

Certainly, the way forward requires a move from the US. “The United States should be encouraged to establish enforceable data subject rights and effective legal remedies for European and other non-US research participants whose data are processed by US researchers. The voice of the health-research community must be heard by decision-makers at the national level, at the EDPB, and within the EU Commission Directorates-General involved, such as in the areas of justice, health and research. Without a quick resolution, European research potential will not be realized, and European citizens will fall behind”.

We saw the importance for EU research to benefit from EU health data. Currently, European health data is targeted by US corporations. A situation that has been criticised.

B – GAFAM US Corporations access to EU health data

1 – The case of the NHS data agreement with DeepMind before Palantir

In this paper, *The Privacy and Security Implications of Open Data in Healthcare A Contribution from the IMIA Open Source Working Group*, the authors, Shinji Kobayashi, Thomas B. Kane and Chris Paton, raise the issue of data anonymisation, and other ethical, and governance challenges (George, Whitehouse, Duquenoy, 2013; Hopia, Punna, Laitinen, Latvala, 2015; Menvielle, Audrain, Menvielle, 2017).

The paper reveals how in 2003, NHS England launched care.data in order to combine all healthcare records stored by general practitioners with all information stored by social services and hospitals. All data was held into the national Health and Social Care Information Centre (HSCIC) databases.

The Hospital Episode Statistics dataset on the other hand collects and curates data from 125 million individuals in England every year. If this data set can have huge potentials for research, it raises the methodology of conservation questions.

UK Care.data has been criticised for sharing data with pharmaceutical companies, insurance companies, health charities, hospital trusts, think tanks, and other private companies. In 2014, it was disclosed that anonymous, pseudonymous, and identifiable data was sold to 160 organisations (Bahatia, 2014). In response to a Freedom of Information request, HSCIC stated: “We recognise that there will however remain a latent risk that when combined with other sources of data, the identity of the individual may be ascertained”. Care.data was closed in 2016 following general criticism and opt out.

In June 2017, Taunton and Somerset NHS Foundation Trust and DeepMind Healthcare signed a 5-year contract to develop and evaluate a system able to detect early signs of kidney-failure (Internet 31). Over 1.6 million live NHS data records were given to Google, via DeepMind.

In November 2017, the UK Information Commissioner ruled that the London’s Royal Free hospital failed to comply with the Data Protection Act when it handed over personal data of 1.6 million patients to DeepMind (Internet 32). A DeepMind spokesperson said the firm “underestimated the complexity of the NHS and of the rules around patient data”. According to the ICO, Elizabeth Denham, their “investigation found a number of shortcomings in the way patient records were shared for this trial. Patients would not have reasonably expected their information to have been used in this way, and the Trust could and should have been far more transparent with patients as to what was happening”. The ICO warned that such work should never be “a choice between privacy or innovation”.

Despite privacy advocates’ hopes, this ruling did not exclude the use of the app. ‘Streams’ that has since been rolled out to other British hospitals, and DeepMind has also branched out into other clinical trials, including a project aimed at using machine-learning techniques to improve diagnosis of diabetic retinopathy (Internet 33), and another aimed at using similar techniques to better prepare radiotherapists for treating head and neck cancers.

Google’s DeepMind Health systems have potential benefits for patients, nurses, and doctors (Internet 34). The DeepMind ‘Streams’ app allows clinicians to be informed when patient vital signs deteriorate using data from patient-monitoring technology that deliver in real-time significant patient life-sign indicators to the clinician's mobile device. If the medical potentials have major benefits, they raise data ownership, secondary usage and ethics concerns.

Julia Powel, from the Faculty of Law and Computer Laboratory, University of Cambridge, was one of the early scholars warning about the ethical issues of Stream App (Internet 35) sharing NHS data with Google DeepMind.

Google DeepMind and healthcare in an age of algorithms (Internet 36)

"Data-driven tools and techniques, particularly machine learning methods that underpin artificial intelligence, offer promise in improving healthcare systems and services. One of the companies aspiring to pioneer these advances is DeepMind Technologies Limited, a wholly-owned subsidiary of the Google conglomerate, Alphabet Inc. In 2016, DeepMind announced its first major health project: a collaboration with the Royal Free London NHS Foundation Trust, to assist in the management of acute kidney injury. Initially received with great enthusiasm, the collaboration has suffered from a lack of clarity and openness, with issues of privacy and power emerging as potent challenges as the project has unfolded. Taking the DeepMind-Royal Free case study as its pivot, this article draws a number of lessons on the transfer of population-derived datasets to large private prospectors, identifying critical questions for policy-makers, industry and individuals as healthcare moves into an algorithmic age".

In a more recent episode, the NHS body was again criticised for being responsible for delivering IT strategy that struggled to ensure patients understand that medical data held by their GPs will be copied into a central database to be shared with third parties unless they opt out by 23rd of June 2022. Conflicting messaging overshadows NHS Digital's attempts to inform the public about patient data slurp (Internet 37).

When a new UK project came up, as reported by The Guardian (Internet 38), "more than a million people opted out of NHS data-sharing in one month in a huge backlash against government plans to make patient data available to private companies, the Observer can reveal. The General Practice Data for Planning and Research scheme is now on hold with no new date for implementation, and NHS (Internet 39). Digital has made a series of concessions to campaigners to try to salvage it".

The Guardian published an article by EerkeBoiten titled *Our personal health history is too valuable to be harvested by the tech giants* (Internet 40).

In December 2019 The Observer revealed (Internet 41) how UK medical data was allegedly sold to American drug companies with little transparency or accountability around the process. "US drugs giants, including Merck (referred to outside the US and Canada as MSD, Merck Sharp and Dohme), Bristol-Myers Squibb and Eli Lilly, have paid the Department of Health and Social Care, which holds data derived from GPs' surgeries, for licences costing up to £330,000 each in return for anonymised data to be used for research".

These revelations raise big questions over the transparency and claims of anonymity in NHS data transfers through the research scheme used by the health service. It appears that individual-level. NHS Digital has announced GP medical records in England would be collected via a new service called General Practice Data for Planning and Research (GPDPR) (Interned 42). It will replace the General Practice Extraction Service (GPES), which has operated for over 10 years.

More recently, it was revealed that Palantir awarded £23m deal to continue work on NHS Covid-19 Data Store (Internet 43) "The two-year contract was first reported today by (Internet 44) who have campaigned for transparency surrounding deals between the NHS and big tech firms". openDemocracy and Foxglove claim the contract was "secretly" signed "in apparent violation of their [the government's] prior promise to conduct future contracts between the NHS and big tech via a full and open public tender".

Medical data harvesting in France

The US Palantir's projects are not limited to the UK. It has been reported that US data giant Palantir is on a mission to seduce France's start-ups". Fears might not be unfounded as Palantir is said to be one of the most secretive companies in the world. Palantir (Internet 45) has expertise in big data analytics having initially worked for the US armed forces and intelligence services.

More recently, French investigative journalists (Internet 46) revealed how medical data was sold including drugs sold by pharmacists to Iqvia (Internet 47). The group had been tracking each patient via a unique identifier number to carry out "analyses of sales of health products aggregated by typologies of pharmacies, by main types of prescribers and by geographical areas".

Following the broadcast of the program, the French data protection authority, CNIL, referred the case to the Paris Judicial Court (Internet 48) asking Internet Service Providers (ISPs) to block access to a site hosting health data of nearly 500,000 people. The CNIL, which has already carried out three controls on this data leak, is continuing its investigations.

In April 2022, the French CNIL fined Dedalus 15 million Euros for their lack of security measures having led to the breach of medical data of 500.000 individuals online (Internet 49).

AMAZON, US online retailer coming big in the health data market medical data

The acquisition of One Medical gives Amazon access to more data. One Medical built its own electronic medical records system, and it has 15 years' worth of medical and health-system data that Amazon could tap.

In the US, it is reported that Amazon Pharmacy (Internet 50) has partnered with Blue Cross Blue Shield plans in five states and pharmacy benefit manager Prime Therapeutics to offer a prescription discount savings card (Internet 51).

Amazon's access to these medical data has to be considered in the context of the mass of various data the corporation has access to (Internet 52).

Medical data have high value. They are naturally targeted by big pharmaceutical companies.

As increasingly pseudonymised or even anonymised data have the potential to be re-identified, the share of such sensitive data is not without controversy. Article 29 working party (Internet 53) has specified that "to identify if a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person". This is at the centre of the question of patients' data security. Patients might have personal reasons to remain anonymous therefore it is paramount the security and confidentiality of often sensitive medical data to be respected.

Conclusion

A balance must be struck between protecting individual patients' fundamental rights of privacy and dignity, the need for research to access data, and for the industry to study the impact of medical data on health improvement. Applying privacy by design, privacy enhancing technologies, anonymisation and encryption in full transparency and respect of individual needs and wishes is paramount. Health data has huge value. "Regulators" vigilance to monitor data handling is essential. EU health data sovereignty will not only improve the privacy of patients and protection of their data, it will also impact the European research capabilities.

References

Bentzen, H.B., Castro, R., Fears, R., Griffin, G., Meulen, V., Ursin, G. (2021). *Remove obstacles to sharing health data with researchers outside of the European Union*. *Nature Medicine, 27,* 1329–1333.

Bhatia, N. (2014). *Register of approved data releases – a Freedom of Information request to NHS Digital*. What Do They Know. Retrieved from: Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering, 5*, 493–497.

FRA (2020). *Getting the Future Right, Artificial Intelligence and Fundamental Rights*. European Union Agency for Fundamental Rights, 39.

George, C., Whitehouse, D., Duquenoy, P. (eds.) (2013). *eHealth: legal, ethical and governance challenges. 1st ed Heidelberg*. New York: Springer.

Hopia, H., Punna, M., Laitinen, T., Latvala, E. (2015). A patient as a self-manager of their personal data on health and disease with new technology--challenges for nursing education. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/privacy-enhancing-technologies-report.pdf.

https://www.wired.co.uk/article/deepmind-ethics-and-society-artificial-intelligence?ref=hacker-noon.com

Internet 1: https://www.fondationlejeune.org/le-numerique-au-service-de-la-recherche/.

Internet 2: https://www.egora.fr/actus-medicales/sante-publique/39040-big-data-l-explosion-de-la--production-de-donnees.

Internet 3: https://www.centerwatch.com/articles/24924-patient-willingness-to-join-clinical-trials-drops-dramatically-new-data-show.

Internet 4: https://www.bbc.com/news/technology-61606477?piano-modal.

Internet 5: https://www.who.int/publications/i/item/9789240029200.

Internet 6: https://www.securityweek.com/mass-personal-data-theft-paris-covid-tests-hospitals.

Internet 7: https://www.theverge.com/2021/6/3/22514951/pandemic-public-health-solutions-google--apple-facebook.

Internet 8: https://artificialintelligenceact.eu/the-act/.

Internet 9: https://digital-strategy.ec.europa.eu/en/library/artificial-intelligence-healthcare-report.

Internet 10: https://inplp.com/latest-news/article/first-european-code-of-conduct-for-the-pharma--industry-approved/.

Internet 11: https://leeds.tech/news/code-of-conduct-for-data-driven-health-and-care-tech/.

Internet 12: https://onlinelibrary.wiley.com/doi/full/10.1111/coin.12427.

Internet 13: The Royal Society, Protecting Privacy in Practice. The current use, development, and limits of Privacy Enhancing Technologies in data analysis.

Internet 14: European Data Protection Board. https://edpb.europa.eu/system/files/2021-05/edpb_letter_out2021-0086_un_en.pdf.

Internet 15: The European Academies Science Advisory Council, the Federation of European Academies of Medicine & the European Federation of Academies of Sciences and Humanities. https://doi.org/10.26356/IHDT.

Internet 16: https://www.ipc.on.ca/ripe-for-public-debate-legal-and-ethical-issues-around-de-identified-data.

Internet 17: https://decisions.ipc.on.ca/ipc-cipvp/phipa/en/item/520967/index.do?q=phipa+175.

Internet 18: Study 21-09-2021, https://www.europarl.europa.eu/thinktank/en/document/ EPRS_STU (2021)690009.

Internet 19: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12870-European--Health-Emergency-Preparedness-and-Response-Authority-HERA-/public-consultation_en.

Internet 20: https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb--edps-joint-opinion-032022-proposal_en.

Internet 21: https://www.euractiv.com/section/health-consumers/news/legal-issues-not-infrastructure-hampers-research-in-health-data-revolution/.

Internet 22: https://tehdas.eu/.

Internet 23: https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0038-1641201.

Internet 24: www.medfloss.org.

Internet 25: https://www.harmony-alliance.eu/en/about-us.

Internet 26: https://www.harmony-alliance.eu/projects/research-project/use-of-big-data-to-improve--outcomes-for-patients-with-acute-lymphoblastic-leukemia-all-2019.

Internet 27: https://www.nature.com/articles/s41591-021-01460-0.

Internet 28: https://doi.org/10.26356/IHDT.

Internet 29: https://www.nature.com/articles/s41591-021-01460-0#ref-CR3.

Internet 30: https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0038-1641201#ORkobayashi-42.

Internet 31: https://www.theguardian.com/technology/deepmind.

Internet 32: https://www.theguardian.com/technology/2016/jul/05/google-deepmind-nhs-machine--learning-blindness.

Internet 33: https://www.theguardian.com/technology/2016/aug/30/google-deepmind-ucl-ai-radio-therapy-treatment.

Internet 34: The Privacy and Security Implications of Open Data in Healthcare https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0038-1641201.

Internet 35: https://www.deepmind.com/blog/scaling-streams-with-google

Internet 36: https://pubmed.ncbi.nlm.nih.gov/29308344/.

Internet 37: https://www.theregister.com/2021/05/24/nhs_digital_gp_data_store/.

Internet 38: https://www.theguardian.com/society/2021/aug/22/nhs-data-grab-on-hold-as-millions-opt-out.

Internet 39: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/general-practice-data-for-planning-and-research.

Internet 40: https://www.theguardian.com/commentisfree/2020/feb/16/our-personal-health-history-is-too-valuable-to-be-harvested-by-tech-giants.

Internet 41: https://www.theguardian.com/politics/2019/dec/07/nhs-medical-data-sales-american-pharma-lack-transparency.

Internet 42: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/general-practice-data-for-planning-and-research.

Internet 43: https://www.digitalhealth.net/2020/12/palantir-awarded-23m-deal-to-continue-work-on-nhs-covid-19-data-store/.

Internet 44: https://www.opendemocracy.net/en/ournhs/controversial-tech-firm-palantir-23m-nhs-data-deal/.

Internet 45: https://www.pearltrees.com/idigital2/palantir/id52210664.

Internet 46: https://www.francetvinfo.fr/replay-magazine/france-2/cash-investigation.

Internet 47: https://www.nextinpact.com/article/45660/donnees-pharmacies-et-iqvia-cnil-sexplique-et-va-mener-controles.

Internet 48: https://www.cnil.fr/en/node/120954.

Internet 49: https://www.cnil.fr/fr/fuite-de-donnees-de-sante-sanction-de-15-million-deuros-lencontre-de-la-societe-dedalus-biologie.

Internet 50: https://www.mobihealthnews.com/tag/amazon-pharmacy.

Internet 51: https://www.mobihealthnews.com/news/amazon-pharmacy-partners-five-blues-plans-offer-prescription-discount-card.

Intermet 52: https://news.yahoo.com/amazons-empire-surveillance-recent-billion-132800720.html.

Internet 53: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

Kaisler, S.H., Armour, F. J., and Espinosa, A.J. (2016). *Introduction to the big data and analytics: concepts, techniques, methods, and applications minitrack*. Proceedings of the Annual Hawaii International Conference on System Sciences (pp. 1059–1060).

Menvielle, L., Audrain, A.-F., Menvielle, W. (2017). *The digitization of healthcare: new challenges and opportunities*. Polgrave Macmillan.

Miller, D.D. (2019). *The medical AI insurgency: what physicians must know about data to practice with intelligent machines*. Art. no. 62.

Nurse Educ Today, *35*(12), e1-3.

Paolo, G., Bincoletto, G. (2021). *A proactive GDPR-compliant solution for fostering medical scientific research as a secondary use of per-sonal health data*. Trento Law and Technology Research Group Research Paper no. 46.

Stöger, K., Schneeberger, D., Kieseberg, P., Holzinger A. (2021). Legal aspects of data cleansing in medical AI. *Computer Law & Security Review, 42*.

Wetsman, N. (2019). Artificial Intelligence aims to improve cancer screening in Kenya. *Nature Medicine, 25*, 1630–1631.

Zulkarnain, N., Anshari, M. (2016). *Big Data: Concept, Applications, & Challenges*. International Conference on Information Management and Technology (pp. 307–310).