Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/CLSR

Computer Law & Security Review

Incentivizing the sharing of healthcare data in the AI Era

Andreas Panagopoulos[1,*], *Timo Minssen*[2], *Katerina Sideri*[3], *Helen Yu*[2], *Marcelo Corrales Compagnucci*[2]

[1] *University of Crete*
[2] *CeBIL, University of Copenhagen*
[3] *Panteion University*

ARTICLE INFO

Keywords:
Artificial intelligence
Data
Healthcare
Collective data management

ABSTRACT

This article contributes to the policy dialogue about how to govern healthcare data in the AI era and how to incentivize patients to share their data. Existing approaches to data-sharing restrict the flow of data. Yet, as healthcare AI technologies rely on data in enhancing their scope, such lack of data hinders the creation of future applications and diminishes the need for data to furnish them. We shift attention to a GDPR based policy that does not restrict data flows and argue that the existing experience in monetizing digitalized copyright material such as music can offer a practical and well tested solution.

1. Introduction

Healthcare data is a very particular type of data. It is not like the clicks, likes, posts and photos people all too often share, which put together an ephemeral and fragmented snapshot of our lives. Healthcare data is considered to be very sensitive, providing information not only about the individual herself but also about her family, parentage and ethnic decent (Forgó *et al.* 2010). It may contain an array of personal information such as physical characteristics, diet, prescriptions, medical reports, laboratory tests, radiographies or genome that captures the frailty of being human in some of its most intimate aspects, making it difficult to legally and ethically persuade people to consent to its use (Fox, 2020; Voigt *et al.* 2020; Middleton *et al.* 2020, 2019).

Yet, such data on its own is inconsequential; a mixture of ink-marks printed on paper, an assortment of pixels randomly spread out on a picture, or a repetitive collection of zeros and ones locked inside a hard drive. It becomes of value only when data-trained algorithms, all too frequently referred to as artificial intelligence (AI), act as a decoder of the information that is bundled up within the data (Rubinfeld and Gal, 2017). Since data-fed AI has the ability to enhance the aptitude of existing applications and increase their scope (Cockburn, Henderson and Stern, 2018) it is important to incentivize patients to share data.

The question is how? After all, it could be argued that it is one's duty to share healthcare data (Cohen *et al.*, 2018). Yet, studies indicate that it is important for people to feel appreciated when sharing such data (Gerke *et al.* 2020) and that they increasingly expect some compensation in exchange for their data (Briscoe *et al.* 2020). Focusing on some form of reciprocity, and employing the law and economics toolbox, the aim of this paper is to identify policies that *will allow agents to claim back some of the value-added their data helped create*, thereby incentivizing the sharing of data with AI firms.

[*] Corresponding author.
E-mail address: andreas.panagopoulos@gmail.com (A. Panagopoulos).

https://doi.org/10.1016/j.clsr.2022.105670

For the average patient claiming back is not straightforward, because the healthcare AI market is controlled by few mega-firms e.g. Google, Microsoft, Apple, etc. The obvious solution is for policies that allow agents to control their data (Fox, 2020) and represent their interests in unison (Madison, Frischmann, and Strandburg, 2009). There exist many collective approaches to governing the control of data. Most are centralized (Health Data Hub),[1] yet some are not (HAT)[2] and may even rely on self-governance (MIDATA).[3] While some are futuristic (SOLID),[4] promise direct payment even in crypto currencies (UBDI),[5] and employ blockchain technology to track data usage (DECODE),[6] others are more traditional (PatientsLikeMe)[7] and may even specialize on specific type of data (Sensotrend).[8] Though many are localized, (Mesinfos)[9] cross-border initiatives (TheGoodData cooperative)[10] exist.

We outline these initiatives and explain that they aim to control the flow of data, inducing its artificial scarcity. This, in principle, should increase the value of what agents can claim back. Yet, this economic reasoning does not apply to AI. As AI is "data-hungry", less data leads to fewer AI applications, limiting the demand for data to train and furnish these applications. By contrast, an expansion in the supply of data increases the strands of interconnected information hidden therein that AI can decipher (Moro Visconti, Larocca and Marconi, 2017) and the subsequent discovery of many more than otherwise novel applications (Prufer and Schottmüller, 2017; Goldfarb and Trefler, 2018) in need of data.

If policies that control the flow of data limit the need for data (and its corresponding value), how about policies that forego such control? As data is intangible and finds multi-territorial uses, absence of control implies that agents do not know how their data has been used and by whom, thereby claiming something back does not seem feasible. Yet, there is experience in performing such a task, albeit in the context of copyright. Consider music: Though its digital nature makes it difficult for creators to identify how their material has been used, they can still claim back some remuneration by setting up an institution known as collective rights management (CRM). This approach, which has been used for over 200 years, does not restrict the flow of music, it is a *laissez-faire* approach. Music is freely played (on the radio, the internet etc.) and then the CRM steps in to claim part of its value-added, to be shared between its members. Building on this experience, we put forth the idea of "collective data management" (CDM), where data is allowed to be freely used and then a representative of the CDM claims back part of the value-added.

As we argue, despite technical challenges that could be addressed with novel methods e.g. blockchain technology (Huberman *et al.* 2020), CDM is not incompatible with current legal context and practice. Nonetheless, CDM faces a shortfall because it has no leverage against firms refusing to offer some remuneration. However, this issue is addressed by the European General Data Protection Regulation (GDPR) that requires explicit consent for the use of healthcare data. Thus, by allowing the CDM's members to control their data-flows (refusing access to non-obliging firms) GDPR essentially becomes a prerequisite for CDM.

2. AI and healthcare

Though AI is a generic term that captures the science of mimicking human intelligence e.g. planning, strategizing and making advanced decisions (Yang, 2017), increasingly AI is defined as the science of teaching a computer to perform human-like tasks. AI learns by training its "rationality" on a given dataset. Though the information included in a dataset is often as unverifiable as the chaos of everyday reality, bundled within it are nuggets of information concealed within an assortment of zeroes and ones. Based on statistical/mathematical reasoning the computer iteratively browses its way through these zeroes and ones, learning to interpret them via a trial and error process. It does so by using weights to relate inputs to respective outputs. Then it measures how close these outputs are to the reality expressed by the dataset. The process is then repeated by adjusting the weights to iteratively narrow down the gap between outputs and reality (Neapolitan and Jiang, 2018).

AI systems are being developed to analyze large amounts of healthcare data and understand human conditions, recognize disease patterns, make highly accurate diagnoses and deliver precision health interventions (Agah, 2014). There are various types of AI tools and techniques currently being used in different settings including hospitals, clinical laboratories, and research facilities (Panesar, 2019).

This approach is already employed in products like AICure[11] and Abilify MyCite,[12] in modeling drug syntheses (Segler *et al.*, 2018), identifying the genes responsible for a condition (Leung *et al.*, 2016), interpreting radiological images (Topol, 2019), prescribing drugs to patients based on their medical records (Athreya *et al.* 2017), or even in mundane applications like inspection of health-code violations (Glaeser *et al.* 2020). The COVID pandemic further highlighted its importance with AI been used in preliminary diagnosis (Ai *et al.*, 2020), monitoring and treating patients (Stebbing *et al.* 2020), development of drugs and vaccines (Chen *et al.* 2020), reducing the workload of healthcare professionals (Ting *et al.* 2020), contact tracing and projecting the spread of the virus (Vaishiya *et al.* 2020), or even in offering advance warning of the virus outbreak.[13]

[1] https://www.health-data-hub.fr/.

[2] https://www.hubofallthings.com/.

[3] https://www.midata.coop/en/home/.

[4] https://solidproject.org/.

[5] https://www.ubdi.com/.

[6] https://decodeproject.eu/.

[7] https://www.patientslikeme.com/.

[8] https://www.sensotrend.com/.

[9] http://mesinfos.fing.org/english/.

[10] https://www.thegooddata.org/.

[11] AICure uses predictive algorithms to communicate information regarding the accuracy of patients' medicinal administration to healthcare providers.

[12] Abilify MyCite is a digital pill that helps to deliver and monitor Aripiprazole, an antipsychotic drug.

[13] The BlueDot AI algorithm scoured the internet and gave advance warning of COVID-19 before the outbreak became public.

AI's reliance on data makes it important for data to flow towards AI. The best way to do so would be via organized markets. Yet, the trade of data is faced with transaction costs that exceed the value of the exchange itself (Schwartz, 2003), making the negotiation and licensing of data unprofitable (Burk, 2015). In the absence of markets, agents must bargain some indemnification directly with AI firms. This is not a problem when agents can bargain with firms of equal stature. However, healthcare AI requires cross-competencies that venture into the medical and pharmaceutical sector. Since this is expensive, few firms have succeeded, most notably Google-Verily, Apple, Microsoft, IBM, Amazon and Facebook. These firms have accomplished such boundary spanning through own research and via collaborations. For example, Google is collaborating with Novartis, Sanofi, Otsuka, Pfizer, the US healthcare provider Ascension (Wachter and Casse, 2020) and the UK's NHS. Moreover, a patent search that took place on early January 2022 revealed that Google and Google-Verily hold 244 patent protected inventions in the A61 patent classification that lists healthcare applications.[14] Apple is collaborating with GlaxoSmithKline, Janssen, the Aetna Life Insurance Co, and holds an extensive portfolio of 350 A61 patents. Equally, Microsoft has partnered with Novartis and holds 298 A61 patents.

When the industry is concentrated, agents have to bargain with monopolists, who, in addition, are also monopsonists faced with a fragmented inputs-market. In such markets the prevailing price is minimal and equal to the marginal cost of generating data. From the patients' perspective the obvious solution is to institutionalize the governance of the sharing of data (Madison, Frischmann, and Strandburg 2009), so that agents control with whom and how data is shared, and then bargain as a conglomerate.

3. Controlling the sharing of data

3.1. *Data hubs and ecosystems*

There exist various approaches that aspire to help agents control their data. Depending on the way they operate they can be categorized as hubs, ecosystems or commons. Data hubs and ecosystems represent some of the oldest approaches in controlling and managing data. Their main difference is that hubs specialize (e.g. in healthcare data), while ecosystems try to create a data-complex with a broader purpose in mind. Data ecosystems in particular, are mostly private initiatives through so called Personal Data Management Services (PDMS) and often operate internationally. Their role is in helping individuals collect, store and share their data under their own terms. There are hundreds of PDMS (e.g. Meeco, MyLife Digital, Mydex), most though focus on various financial transactions. One of the best known healthcare PDMS is LunaDNA, a community-owned platform for healthcare research that allows agents to share their data in exchange for ownership shares in the organization. The most popular PDMS is digi.me. Digi.me, in cooperation with the Icelandic Government, created a data ecosystem for the whole of Iceland. It includes data on prescriptions, medications, vaccinations, allergies and medical admissions.

The Hub of All Things (HAT) represents the most futuristic version of PDMS due to its portability across devices. It is a personal server that allows agents to bring their data from the Internet into the HAT, exchange data with various applications and even install private analytics tools. A related technology is the Social Linked Data (SOLID) project, pioneered by MIT and Inrupt Inc. It aims for full data control via a platform that allows agents to control their data, including access control and storage.

Though PDMS can manage data, bearing in mind the multitude of data, this is often delegated to specialized Personal Information Management Services (PIMS). They allow data management in a secure way that is often linked with other data management services. For example, Sensotrend is a Finish PIMS that manages diabetes data collected from healthcare providers. Equally, the US based Universal Basic Data Income (UBDI), which is built on digi.me, matches the data from digi.me with research studies. Its purpose is for individuals to obtain a data-income while protecting their privacy. HAT and SOLID equally allow selected PIMS to access their data.

Data hubs are specialized and operate in a single country. The oldest is the US based PatientsLikeMe, which has more than 600,000 members. It was originally developed for connecting patients with Lou Gehrig's disease, but has expanded to include more diseases. Its aim is to track and share patients' experiences, help patients get a better understanding of their condition and options, and generate the data needed to develop a treatment. Notable Europe-based hubs are Mesinfos and Health Data Hub[15] in France, and Medisanté[16] in Switzerland. Medisanté connects healthcare providers to patients' data for chronic diseases through various connecting devices, a connected care platform and in-country data hosting. Mesinfos involves a consortium of companies that explore the collection, use and sharing of personal data. The Health Data Hub collects data from patients, insurance companies and hospitals to improve the efficiency of healthcare services using AI techniques (Ayoubi and Foray, 2019).

3.2. *Commons*

Data commons aim to empower groups of agents to assemble data in terms the group's stakeholders themselves set out (Evans, 2016). Commons build on the idea that the pooling of resources under one framework facilitates a shared purpose for all stakeholders (Frischmann, Madison and Strandburg, 2014). Being mostly localized, they can be found in Switzerland (MIDATA), Amsterdam and Barcelona (DECODE), but there also exist cross-border commons such as the TheGoodData co-operative (Symons, Bass and Alegre, 2017).

TheGoodData is an international effort to collect, pool, and sell members' browsing data (including healthcare data)

[14] A61 is a patent classification that is used to identify technologies related to medical sciences and hygiene. This is the main patent class that pharmaceutical firms and firms producing therapeutics and health related technologies use when classifying their patented technologies.

[15] https://www.health-data-hub.fr/.

[16] https://medisante.ch.

on their terms. MIDATA is a Swiss co-operative that allows healthcare providers to collect a variety of data that is encrypted and stored in a way that permits users to track data and decide on how and with whom to share it. DECODE (Decentralized Citizen-owned Data Ecosystem) is an EU initiative in Amsterdam and Barcelona. These cities allow people to decide which applications, platforms and tools can access their information, with a prerequisite for their data to be used for the common good. For example, in Barcelona participants share their data with the city to be used on projects such as measuring noise levels.

What empowers DECODE is the "digital wallet": an online platform that contains members' data that apps collect. It allows members to share data only with the projects they want. This is done via blockchain technology, which allows transactions to be recorded in a verifiable and permanent way. When agents share their data they place it in this wallet, which ensures that only the required data will be used and that agents define how it is to be used. The promise of this technology is that it takes out the middle man, operating a digital marketplace without intermediaries (Catalini and Gans, 2016).

Overall, commons, ecosystems and hubs share a common purpose, to act as "intellectual property without intellectual property" (Strandburg, Frischmann, and Madison, 2017). Similar to intellectual property (IP) their *raison d'etre* is to endow agents with control over who employs their data and how. They offer the autonomy of data from outside brokers, allowing data to be managed irrespectively of whether it is private property or not. In doing so, they inevitably curtail the flow of data, which does not flow freely from agent to firm. It flows only when agents decides so.

3.3. The future

For the near future the EU is planning governance structures that allow firms access to data. Their aim is to to foster the availability of data and to support responsible and sustainable R&D by increasing trust in data intermediaries and strengthening data sharing across the EU and between sectors. . The anticipated increase in social welfare is envisioned as one incentive for agents to contribute their data. However, as we now explain, other factors may also restrict the flow of data.

This year (2022) the European Open Science Cloud (EOSC) should take its first steps,[17] followed soon by Personal Data Spaces (PDS), such as the European Health Data Space (EHDS) (European Commission 2020, Vayena, 2021), as well as further initiatives, such as the recently adopted EU Data Governance Act[18]. In the following we will focus on the EOSC and the PDS.[19] EOSC aims to create a pool of research data and discoveries (from a variety of disciplines and a wide range of research infrastructures), and a secure environment to access, combine, analyze, store data and share its results.[20] PDS, such as the EHDS, is envisioned as the means of strengthening the control agents have over data in a way that is open to global inflows of data in an accessible way that allows for data analytics and machine learning. To increase the flow of data, agents will be rewarded via increased data access, analytical results, predictive maintenance services, or even license fees.

Both EOSC and PDS intend to establish a common culture of data stewardship to ensure data reuse via an open-by-default research environment supported by FAIR data principles. Though the FAIR principles are a quality standard requiring scientific data to be (F)indable, (A)ccessible, (I)nteroperable and (R)eusable (Wilkinson *et al.* 2016), the "Accessible" and "Interoperable" aspects of FAIR create limitations in how data is accessed and used. Specifically, FAIR recognizes legitimate and necessary reasons for restricting access especially in the health/life sciences. In fact, though data should be accessible it should also be "as open as possible and as closed as necessary", in the sense that IP rights and other means of reusability restrictions are being accepted as long as they remain limited, and the "openness" remains the default position. This inevitably invites questions of ownership rights of data because data usage flourishes at the intersection between the FAIR narrative, trade-secrets and IP rights.

This creates three problems. First, both EOSC and PDS are plagued by disclosure incentives, because without IP or mechanisms to safeguard R&D investments stakeholders may be discouraged from sharing data. Second, even if they share data its ownership may be challenged (Minssen and Pierce, 2018), giving rise to controversies between data "owners" and the entities using the data (Minssen, Rajam and Bogers, 2019). Third, there is a reciprocity issue. Considering that these initiatives are advertised as globally accessible there is no mechanism in place that obliges non-EU entities to grant data access on the same conditions. This opens up the door for unfair competition, which may force agents to restrict access to their data.

Additionally, FAIR faces interoperability problems that may also confine data access. Interoperable data (apart from being accessible and shared) also comprises technical and legal interoperability (Wilkinson *et al.* 2016). This relates to the ability to combine datasets from multiple sources without conflicts among restrictions imposed by data providers (i.e. support of one restriction inherently negating support of another). The fewest restrictions contained in the source datasets will result in the fewest restrictions contained in the combined or derivative datasets. This is not always easy to achieve as FAIR recognizes reasons for restricting access, particularly in healthcare. Examples include FAIR data that contains personal information, limitations concerning informed consent and access agreements, or IP and confidential information (Collins *et al.*, 2018; Graber-Soudry *et al.*, 2020). Overall, a clear-cut pic-

[17] The EOSC was initiated in 2017 when the EU Commission released the EOSC declaration to the scientific community (EU, 2017).

[18] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance (Data Governance Act) COM/2020/767 final.

[19] EOSC and PDS build on US experience, and in particular on the Blue Button Initiative. This health data hub is in a category of its own. It is a first and incomplete step towards data control that was launched in 2010 by the U.S. Department of Veteran Affairs. It endows patients with a copy of their data and allows them to view online and download their own personal health records. Its aim is to improve the quality of patient-clinician interactions with the expectation that these would improve social welfare by contributing to enhanced quality of life, better treatment outcomes and potential reduction in costs.

[20] https://www.egi.eu/about/newsletters/what-is-the-european-open-science-cloud/.

Table 1 – Approaches to the governance and control of healthcare data.

Method	Type of initiative	Specialized	Area of operation
PDMS			
Meeco	Private	All types of data	International
MyLife digital	Private	Mostly healthcare data	International
Mydex	Private	All types of data	International
HAT	Private	All types of data	International
SOLID	Private	All types of data	International
digi.me	Private	All types of data	International
LunaDNA	Private	Healthcare	International
PIMS			
UBDI	Private	All types of data	International
Sensotrend	Private	Diabetes data	Finland
HUBS			
PatientsLikeMe	Private	Healthcare data	USA
Medisanté	Private	Chronic disease data	Switzerland
Mesinfos	Private	Hospital data	France
Health Data Hub	Private	Healthcare data	France
CENTRALLY PLANNED			
EOSC	EU initiative	Research data	EU
PDS	EU initiative	All types of data	EU
COMMONS			
MIDATA	Private	Healthcare data	Switzerland
DECODE	EU initiative	Various types of data	Amsterdam, Barcelona
TheGoodData	Private	Mainly internet data	International

ture of all governance structures and their main characteristics is given in Table 1.

4. When is data in demand?

The assumption the previous section rests on is that data is in demand and thereby of value, in which case data subjects can indeed hope for some remuneration for the use of their data. Yet, it is not clear why this is so. Data displays decreasing returns to scale (Bessen, 2018) i.e. an increase in the data employed in training AI leads to a less than proportional increase in its “rationality”. Thereby, the utility of additional data stagnates with use. Consequently, when firms already have the needed data in stock, there may be little need for additional data. To put this into perspective, the system described by Esteva *et al.* (2017) for classifying forms of cancer was trained to offer accurate classifications by using just a few hundred images per cancer type (e.g. it classified skin melanoma after being trained on about 1,000 images) despite having available a library of 129,450 clinical images of 2,032 different diseases.

The key for data being in demand is the availability of datasets of considerable *volume and variability* for AI to apply its trade on, accompanied by *non-rivalry* i.e. the ability to reuse data[21] without exhausting its capacity on one single application (Jones and Tonetti, 2019). Think of a dataset that includes a limited amount of data that focuses only on patients’ symptoms. Due to data’s non-rivalrous nature this dataset can be used by AI to uncover hidden commonalities that venture beyond a single application. However, due to the limited breadth of the data, these applications will inevitably revolve around diagnosis and prescription. Though these applications require data for their training, their limited number implies that decreasing returns to scale will inevitably curb the need for data. This is not so if data’s volume and variability increases.

If this dataset is merged with data on patients’ characteristics, despite the fact that each dataset involves its own veracity, the conglomerate contains more fragments of useful information than the sum of its parts. These allow additional heretofore hidden multi-dimensional meanings to resurface (Moro Visconti, Larocca and Marconi, 2017), bringing new applications to the foreground such as: mental health and psychological tracking, diet and fitness tracking, disease management guidance, wellness recommendations, history and records, predictive impact modeling tools, automated patient query support (Yella *et al.* 2018). A further merger with clinical data and genome data will help establish survival rates and likely responses to treatment (Yu *et al.* 2017) as well as genomic screening that allows for genome wide association studies that can reveal novel genotype-phenotype associations (Frey, 2019) to be used in drug design, the recognition of drug targets and drug screening (Wei and Denny, 2015).

To rephrase, when a patient’s data is fuzzed with other data it becomes interconnected with different types of information, and as a conglomerate of increased volume and variability it conveys to AI systems a multitude of knowledge about many independent uses. This expands the scope of AI -its market applications (Goldfarb and Trefler, 2018), allowing exploration of new markets (Prufer and Schottmüller, 2017). It is

[21] Consider Google-Verily, which uses eye images in identifying diabetic retinopathy. These images can be redeployed in other areas Google is active on e.g. in combating presbyopia, the detection of diabetes trough monitoring glucose in tears, or in detecting cardiovascular problems through the optical analysis of the eye’s blood vessels.

this increase in applications that drives the demand for training data, which will stagnate only when AI runs out of uses (Farboodi and Veldkamp, 2019), and only at that point decreasing returns scale set in.

By the same token, if the data of some individuals is detached from the whole we do not only forego information regarding their medical conditions, we do without all their interconnections as well. Thereby, this drop in the supply of data (and its volume and variability) is amplified, diminishing the demand for data. However, while the need for governance of data uses is increasing (O'Doherty *et al.*, 2021) most governance approaches curb data flows. Accordingly, we propose to shift attention to methods that avoid doing so.

5. Managing data

5.1. *Collective rights management*

Pierre-Augustin Caron de Beaumarchais was a French polymath with a background as a watchmaker, patentee, playwright, musician, diplomat, spy, publisher, satirist and financier. Worried about the use, reuse, and bundling of his plays (*Le Barbier de Séville* and *La folle journée, le Mariage de Figaro* spring to mind), in 1777 he founded La Société des Auteurs Dramatiques, the first communal organizations to protect authors' rights. The good writers who were members of this society did not care about how their literary ideas were used. Their single concern was to claim back some of the value the fruit of their genius had created for others. For this purpose, the society was given the right to monitor the use of their ideas and represent its members as a single authority. Furthermore, the members were well aware of the limitations of monitoring. They understood that it is impossible to account for all occasions of usage and were content with a remuneration that broadly and in average terms captured how their ideas were put to use.

The solution put forth by Beaumarchais is collective by nature, but it does not focus on aggregating the literary production of its members, so as to bargain with whoever is interested in using, reusing, or bundling it. Its intention is not to limit how the material is employed, instead its aim is to monitor average usage and then bargain back a fair share (Brousseau and Bessy, 2005). The name of this approach, *Collective Rights Management* (CRM), is indicative of its purpose to act as the manager of the society's members.

Specifically, CRM refers to the licensing of copyright material by for-profit organizations commonly known as *collecting societies*, which act on behalf of rights owners. These societies offer a mechanism used in copyright, where managing individual rights within a complex industry with many stakeholders may not be realistic for an individual. Within the context of CRM rights owners transfer to the society rights to: sell licenses, collect and distribute royalties and enforce owners' rights. Since rights owners differ drastically in their needs and characteristics (e.g. they can be musicians, authors, performers, composers, writers, record labels, etc.) collecting societies have become specialized. For example, there exist artist rights groups that license and collect royalties for the reproduction of paintings, or even collectives that collect royalties for copies from magazines and scholarly journals.[22]

In the EU CRM is governed by the Collective Rights Management Directive that aims at ensuring that right-holders have a say in the management of their rights and at improving the functioning and accountability of collecting societies. As copyright content (especially music) is mostly digitalized, the Directive intends to facilitate the multi-territorial licensing of authors' rights for online use. Overall, its objective is to ensure collecting societies act in the rightsholders' best interest through common standards of governance, financial management, and transparency.[23] Drawing on CRM's functionality and its experience in managing online copyrighted content we put forth the idea of *collective data management*.

The aim of collective data management (CDM), unlike the data governance approaches of section 3, is not to control how data is compiled, shared and used, which eventually lessens the scope of data. CDM should, in fact, allow firms to amass the data, process it, employ it and profit from it. However, it should be collectively entrusted by its members to: a) monitor data usage and b) bargain back some *ex post* remuneration.

Before focusing on the technical issues surrounding CDM and its practical ability to perform these two functions, we need to consider how CDM fits within our existing legal framework. After all, for any policy to be successful it needs to accord with current norms and practices and differ from existing legal practice as little as possible. As CDM builds on the experience of CRM, its use does not openly deviate from current norms and practice. Nevertheless there are legal challenges.

Key to implementing a CDM framework is to ensure that transparency and accountability is built into the governance infrastructure to foster public trust and confidence (Lucena, 2015). Because of its personal nature, access to the use of healthcare data creates a social relationship that needs to be acknowledged. As this relationship entails responsibility and awareness of how data use aligns with societal values (Baker and Karasti, 2018), firms, healthcare providers and patients must be encouraged to be mutually responsive to each other to reduce the risk of opposition and increase effective uptake (Yu, 2016). Thereby, the technical infrastructure of CDM must respect data governance processes designed to achieve transparency, integrity, security and accountability. Fortunately, at the EU level the Collective Rights Management Directive already provides a blueprint on how to set such governance standards, facilitating the needed transparency and accountability.

5.2. *Monitoring and bargaining*

As long as CDM functions within a context that we are not unaccustomed to, our focus should be on the practical side of

[22] The best known collecting society is the US based Broadcast Music Inc. (BMI), which represents songwriters and composers. In 2020 BMI distributed and administered $1.311 billion in royalties to its members.

[23] Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective management of copyright and related rights and multi-territorial licensing of rights in musical works for online use in the internal market Text with EEA relevance *OJ L 84, 20.3.2014, p. 72–98.*

CDM, the technical challenges it faces, and on the existence of *tested* solutions that can help CDM monitor data usage and then bargain back some indemnification. Monitoring involves gaining an understanding of how data found use within firms. Due to the digital nature of data, this can only be accomplished if data is automatically shared between the proprietary IT systems of different agents in a way that permits the CDM to balance between the needs for internal confidentiality and data sharing (Tsarsitalidis *et al.* 2021). This implies that monitoring should allow an understanding of which specific part of the information is shared, for how long, under which conditions, circumstances and filters, and with what authorization/authentication mechanism. Though such monitoring seems technically challenging, data tracking of this kind is already employed by some of the methods outlined in section 3.

Consider DECODE (or Digi.me), which uses distributed ledger[24] technology (often referred to as blockchain technology) to track data. This technology creates an audit log that permits the immutable attribution of data usage because it allows transactions between two parties to be safely recorded in a verifiable and permanent way. To provide an example, in DECODE agents can choose to share their data for a common cause by placing it in a digital data-bank that is shared, and synchronized, across multiple sites through a peer-to-peer system so that it does not require centralized data storage, in a way that allows for data tracking. Such technology provides a secure authorization and authentication mechanism that allows DECODE to identify the data that is shared, the end user and the duration of usage. Plus, DECODE is also in a position to know the identity of the agent sharing her data despite any anonymization that may take place prior to firm usage.

Tracking of this type has its limitations, in the sense that it is still hard to gain an understanding of how the data of a particular data-subject found use (and the value added it helped create). Nevertheless, someone skilled in this art, when equipped with such information, should be in a position to gain an appreciation of data usage in its plurality and the value added it created as a conglomerate. To rephrase, an expert who understand the production process and has at hand accurate information about the inputs of production should be able to gain a good judgment of the respective outputs of production and their average value. This is not a novel way of gaining an understanding of the output that technologies create. Economists regularly focus on inputs, e.g. patents, R&D expenditure, specialized personnel etc., to gain a good understanding of how certain high-tech sectors are progressing (Greenhalgh and Rogers, 2010, Ch. 3).

Having an irrefutable log of data usage and a representative perception of the value added it generated, the CDM would proceed with its second function, to bargain with AI firms, ascertaining the members' interests, distributing a fair share to each. This share does not have to be pecuniary. One can envision an array of solutions ranging from better and more novel treatments, personalized services, to cheaper drugs etc that can then be distributed to members. In cases where IP rights may become relevant, an appropriate governance through CDM could make such rights part of the solution to increase legal interoperability (Graber-Soudry *et al.*, 2020).

Yet, to bargain on an equal level one needs a way to enforce her position against the likes of Google, Microsoft, Apple etc. The communal approaches outlined in section 3 were all built from the ground up as methods to restrict access. To use their data you had to abide with their rules. CDM on the other hand is built on the idea of unrestricted access to data. As such it lacks a "stick" to enforce its will against firms that, having already used the data, refuse to remunerate. For CRM this is not a problem as the content it focuses on is copyright protected, unlike data, which has no property rights.

This is not an unsolvable issue as long as patients have some control over their data and can exclude gratuitous usage, in which case all the CRM has to do is instruct its members to avoid "donating" their data to non-abiding firms. Such provisions are already encompassed in the GDPR. Specifically, the GDPR is a centralized EU approach in allowing individuals some control over their data. It requires firms to institute a compliance program to address legal duties,[25] and applies to "personal data" that is "processed" by a firm, where "processing" includes nearly anything a firm would do to data (Price *et al.*, 2019). Since the GDPR considers healthcare data as sensitive data, thus a special category of data, to process it firms need explicit consent (Art. 9 of the GDPR). Plus, GDPR follows an extraterritoriality approach, and non-EU firms may fall within scope of the GDPR (Art. 3 of the GDPR) if they offer goods or services to individuals in the EU, even if they have no establishment in the EU and are not performing any data processing activities within the EU (EDPB Guidelines 3/2018).[26] Thereby, the existence of GDPR has the capacity to endow CDM with the needed leverage in its bargaining negotiations, and, conversely, the absence of GDPR style provisions negate the ability of CDM to meet its ends.

GDPR is a *sine qua non* for CDM for one more reason. It offers guarantees that when data is rendered anonymous it should indeed no longer be identifiable. Specifically, depending on its attributes, data is defined as anonymized or pseudo-anonymized. The GDPR defines the process of pseudo-anonymizing data as "*the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and*

[24] A distributed ledger is a replicated, shared, and synchronized digital data that is geographically spread across multiple sites in a way that it does not require a central administrator or centralized data storage. It only needs a peer-to-peer network that allows each site to replicate update and save an identical copy of the ledger.

[25] These duties can include: an obligation to obtain information legally (Art. 21); an obligation to collect and process only as much data as is necessary for a pre-articulated purpose (Art. 1(c); obligations to affirmatively notify individuals when their data have been received -even from a third party (Art. 14); obligations to perform risk assessments and impact assessments and implement risk-mitigation measures (Arts. 24, 35-36; Rec. 76, 77); and obligations to design new technology with privacy and other rights protections in mind (Art. 25).

[26] Guidelines 3/2018 on the territorial scope of the GDPR (Article 3) Version 2.1 12 November 2019.

organizational measures to ensure non-attribution to an identified or identifiable individual" (Article 4(3b)).[27]

The problem is that the additional information required to ensure non-attribution is increasingly becoming fluid. Though individuals can be directly identified from unique personal characteristics such as their name or telephone number, in practice they can also be indirectly identified from personal attributes that place them within a specific subset of the population, for example the fact that they have a particular health condition, a specific job title, or even their postcode. New techniques that allow for data triangulation and data-fusion (Rocher, Hendrickx and De Montjoye, 2019) are now used to combine seemingly unrelated information to reconstruct a person's identity, in which case pseudo-anonymized data runs the risk of privacy breach. Since healthcare data includes sensitive information (Forgó *et al.* 2010) and far-reaching conclusions can be drawn by combining it (Corrales Compagnucci *et al.* 2019) such possible lack of anonymization has raised concern among patients. Studies showed that many patients believed that there is risk of such breach (Langarizadeh *et al.* 2018), making it hard to persuade patients to consent to the use of their data (Fox, 2020; Voigt *et al.* 2020; Midleton *et al.* 2020).

Anonymized data on the other hand is data that does not run the risk of reverse-identification. The GDPR defines anonymized data as "*information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*" (Recital 26 of the GDPR),[28] and only when data fulfils this definition it is exempt from GDPR provisions. As anonymized data protects one's identity it can address privacy concerns, promoting the sharing of data that AI relies on.

The discussion has stayed away from how CDM should address data uses by publicly funded research institutions and for research reasons in general. Research is no longer restricted to a philosophical inquiry that involves tinkering with scientific experiments for the purpose of knowledge creation. It now also is a profits generation activity because these institutions are actively encouraged to license the fruits of their research (Panagopoulos and Sideri, 2021). However, as society understands such activity as part of the welfare enhancing ideas-generation/dissemination process, in fostering welfare policymakers have created so-called research exemptions.

For example, in aiding scientific research (Meszaros and Ho, 2021), Article 89 of the GDPR understands such data uses as "privileged" in the sense that they are exempt from the provisions of Articles 15, 16, 18 and 21 of the GDPR, which respectively define the data subject's rights of access, rights to rectification of inaccurate data, as well as rights to restriction of processing and rights to object. Furthermore, Article 5(3)(a) of the Information Society Directive[29] allows exceptions for teaching or scientific research, and Article 3 of the Digital Single Market Directive[30] similarly allows exceptions for text and data mining, including the use of data bases, for the purposes of scientific research.

There is an additional (practical) reason dictating the exclusion of scientific research from CDM related actions: decreasing returns to scale. Unlike firms that invest in data of volume and variability to furnish multitudes of applications, research institutions employ data for the single purpose of answering one or few scientific questions. However, as explained in section 4, when AI is used to develop few applications the explanatory capacity of additional data stagnates with use (Bessen, 2018), because the mean accuracy of prediction increases with the number of training data but at a decreasing rate (Varian, 2018), limiting the demand for data and its value. Thereby, the research institution and CDM have little to bargain on.

6. Conclusions

This paper contributes to the ongoing policy dialogue about how to best govern data in the AI era (Micheli *et al.* 2018) and, specifically, on how to incentivize the sharing of healthcare data despite concerns about misuse (Fox, 2020). The main message the paper confers is that (since AI relies on data) policy makers should not lose sight of the supply of data. However, as current methods of incentivizing data-sharing inevitably limit its supply, diminishing the scope of AI, we focus in ways to offer incentives without restricting data-flows.

Collective data management is not an out of the blue novelty. It is not an approach to monetizing digitalized content in a multi-territorial fashion whose practical side is yet to be discovered, needs experimentation, and requires substantial changes in our legal context, mentality and practice. It is a GDPR based solution that builds on the wealth of globally accumulated experience in numerous copyright related contexts. While technical challenges might remain to address the necessary trade-offs and to design a secure system protecting individual rights, these could be addressed with new technical methods such as blockchain technology (Kostick *et al*., 2022, Huberman and Hogg, 2021; Corrales Compagnucci *et al*. 2019).

Declaration of Competing Interest

The authors declare no conflict of interest.

Data Availability

No data was used for the research described in the article.

Acknowledgments

The authors would like to acknowledge the helpful comments and discussions with Kacper Szkalej, Dirk Czarnitzki, Rein-

[27] https://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm.

[28] https://www.privacy-regulation.eu/en/recital-26-GDPR.htm.

[29] DIRECTIVE 2001/29/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

[30] DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

hilde Veugelers, Gideon Markman, Peter Gianiodis, Glen Cohen, Sarah Gerke, Vincenzo Denicolo, Nicolas Vonortas, Alan Marco, In-Uck Park, Robert Merges, Jeffrey Roberts, Christos Colympiris, Andre Spithoven. Panagopoulos is grateful to the seminar participants at TECHNIS and the faculty at the Department of Management, Strategy and Innovation at KU Leuven for their help during a very interesting sabbatical year. MCC's, TM's and HY's research for this contribution and the payment of the open access fee was supported by a Novo Nordisk Foundation grant for a scientifically independent Collaborative Research Programme in Biomedical Innovation Law (Grant agreement number NNF17SA0027784). "Sideri is grateful for the financial assistance of the EU's Marie Skłodowska-Curie grant agreement Biased AI DLV-794604." All errors and omissions remain our own.

REFERENCES

Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, … Xia L. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology 2020.

Athreya AP, et al. Data-driven longitudinal modeling and prediction of symptom dynamics in major depressive disorder: Integrating factor graphs and learning methods. 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); 2017. p. 1–9.

Ayoubi C, Foray D. Machine Learning in Healthcare: New models of innovation, productivity and the future of the Baumol disease. EPFL working paper 2019.

Baker KS, Karasti H. Data care and its politics: Designing for local collective data management as a neglected thing. Proceedings of the 15th Participatory Design Conference: Full Papers-Volume 1; 2018. p. 1–12.

Bessen J. AI and Jobs: The role of demand (No. w24235). National Bureau of Economic Research; 2018.

Briscoe F, Ajunwa I, Gaddis A, McCormick J. Evolving Public Views on the Value of One's DNA and Expectations for Genomic Database Governance: Results from a National Survey. PLOS ONE 2020;15(3) March 11,. doi:10.1371/journal.pone.0229044.

Brousseau E, Bessy C. Public and Private Institutions in the Governance of Intellectual Property Rights. In: Andersen B, editor. Intellectual Property Rights:Innovation, Governance and the Institutional Environment. Edward Elgar Publishers; 2005.

Burk DL. Patents as Data Aggregators in Personalized Medicine. BUJ Sci. & Tech. L. 2015;21:233.

Catalini C, Gans JS. Some simple economics of the blockchain (No. w22952). National Bureau of Economic Research; 2016.

Chen S, Yang J, Yang W, Wang C, Bärnighausen T. COVID-19 control in China during mass population movements at New Year. The Lancet 2020;395(10226):764–6.

Cockburn IM, Henderson R, Stern S. The Impact of Artificial Intelligence on Innovation (No. w24449). National Bureau of Economic Research 2018.

Cohen IG, et al. Is There a Duty to Share Healthcare Data?. In: Cohen Glenn, et al, editors. in Big Data, Health Law, and Bioethics. Cambridge University Press; 2018. p. 209–22.

Collins, S., Genova, F., Harrower, N., Hodson, S., Jones, S., Laaksonen, L., … & Wittenburg, P. (2018). Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data.

Corrales Compagnucci Marcelo, Meszaros Janos, Minssen Timo, Arasilango Arasaratnam, Ous Talal, Rajarajan Muttukrishnan. Homomorphic Encryption: The 'Holy Grail' for Big Data Analytics & Legal Compliance in the Pharmaceutical and Healthcare Sector? European Pharmaceutical Law Review (EPLR) 2019;3(4):144–55 Available at SSRN: https://ssrn.com/abstract=3488291 or http://dx.doi.org/10.2139/ssrn.3488291.

Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective management of copyright and related rights and multi-territorial licensing of rights in musical works for online use in the internal market Text with EEA relevance OJ L 84, 20.3. 2014, p. 72–98.

European Commission (2020), European Commission, Press release (2020). Commission and Germany's Presidency of the Council of the EU underline importance of the European Health Data Space (accessible at https://ec.europa.eu/commission/presscorner/detail/en/IP_20_2049).

Evans BJ. Barbarians at the gate: consumer-driven health data commons and the transformation of citizen science. American journal of law & medicine 2016;42(4):651–85.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. nature 2017;542(7639):115–18.

Farboodi M, Veldkamp L. A Growth Model of the Data Economy. Working Paper, Columbia Business School; 2019 New York, June 20.

Forgó N, Kollek R, Arning M, Kruegel T, Petersen I. Ethical and Legal Requirements of Transnational Genetic Research. Hart Publishing; 2010.

Fox K. The Illusion of Inclusion—The "All of Us" Research Program and Indigenous Peoples' DNA. New England Journal of Medicine 2020;383(5):411–13.

Frey LJ. Artificial Intelligence and Integrated Genotype–Phenotype Identification. Genes 2019;10(1):18.

Frischmann BM, Madison MJ, Strandburg KJ. Governing Knowledge Commons. Oxford University Press; 2014.

Gerke S, Minssen T, Cohen IG. Ethical and Legal Challenges of Artificial Intelligence-Driven Health Care?. In: Bohr Adam, Memarzadeh Kaveh, editors. in Artificial Intelligence in Healthcare. Elsevier; 2020.

Glaeser E, Hills A, Kominers S, Luca M. How Does Compliance Affect the Returns to Algorithms? Evidence from Boston's Restaurant Inspectors. Harvard Business School Working Paper 2020.

Goldfarb A, Trefler D. AI and international trade (No. w24254). National Bureau of Economic Research; 2018.

Graber-Soudry O, Minssen T, Nilsson D, Corrales M, Wested J & Illien B (2020) Legal Interoperability and the FAIR Data Principles (Version 1.0). Zenodo. doi:10.5281/zenodo.4471312.

Greenhalgh C, Rogers M. Innovation, intellectual property, and economic growth. Princeton University Press; 2010.

Huberman BA, Hogg T. Privacy and data balkanization: circumventing the barriers. AI and Ethics 2021:1–7.

Jones CI, Tonetti C. Nonrivalry and the Economics of Data (No. w26260). National Bureau of Economic Research; 2019.

Kostick-Quenet K, McGuire AL, Minssen T, Mandl KD, Kohane I, Cohen G, Gasser U. How NFTs could transform health information exchange. Science 2022;375(6580):500–2.

Strandburg Katherine J, Frischmann Brett M, Madison Michael J. The Knowledge Commons Framework in Frischmann. In: Strandburg BM, Madison MJ, editors. Governing Medical Knowledge Commons. Cambridge University Press; 2017 ch. 1.

Langarizadeh Mostafa, Orooji Azam, Sheikhtaheri Abbas, Hayn D. Effectiveness of Anonymization Methods in Preserving Patients' Privacy: A Systematic Literature Review. eHealth 2018:80–7.

Leung MK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: a review of computational problems and data sets. Proceedings of the IEEE 2016;104(1): 176–197.

Lucena C. Collective Rights and Digital Content: The Legal Framework for Competition, Transparency and Multi-territorial Licensing of the New European Directive on Collective Rights Management. Springer; 2015.

Madison MJ, Frischmann BM, Strandburg KJ. Constructing commons in the cultural environment. Cornell L. Rev. 2009;95:657.

Meszaros Janos, Ho Chih-hsing. AI research and data protection: Can the same rules apply for commercial and academic research under the GDPR? Computer Law & Security Review 2021;41.

Micheli M, Blakemore M, Ponti M, Scholten H, Craglia M. The Governance of Data in a Digitally Transformed European Society. Second Workshop of the DigiTranScope Project; 2018.

Middleton A, Milne R, Howard H, Niemiec E, Robarts L, Critchley C, … Smith J. Members of the public in the USA, UK, Canada and Australia expressing genetic exceptionalism say they are more willing to donate genomic data. European Journal of Human Genetics 2020;28(4):424–34.

Middleton A, Milne R, Thorogood A, Kleiderman E, Niemiec E, Prainsack B, … Vears D. Attitudes of publics who are unwilling to donate DNA data for research. European journal of medical genetics 2019;62(5):316–23.

Minssen, T., & Pierce, J. (2018). Big Data and Intellectual Property Rights in the Health and Life Sciences.

Minssen T, Rajam N, Bogers M. Clinical trial data transparency and GDPR compliance: Implications for data sharing and open innovation. Science and Public Policy; 2019.

Moro Visconti, R., Larocca, A., & Marconi, M. (2017). Big Data-Driven value chains and digital platforms: from Value Co-Creation to Monetization. Available at SSRN 2903799.

Neapolitan RE, Jiang X. Artificial intelligence: With an introduction to machine learning. CRC Press; 2018.

O'Doherty KC, Shabani M, Dove ES, et al. Toward better governance of human genomic data. Nat Genet 2021;53:2–8. doi:10.1038/s41588-020-00742-6.

Panagopoulos Andreas, Sideri Katerina. Prospect patents and CRISPR; rivalry and ethical licensing in a semi-commons environment. Journal of Law and the Biosciences 2021;8(2):lsab031.

Panesar A. Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes. Coventry: Apress; 2019.

Price II WN, Kaminski ME, Minssen T, Spector-Bagdady K. Shadow health records meet new data privacy laws. Science 2019b;363(6426):448–50.

Prufer J, Schottmüller C. Competing with big data. Tilburg Law School Legal Studies Research Paper Series No. 06/2017 2017.

Rocher L, Hendrickx JM, De Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications 2019;10(1):1–9.

Rubinfeld DL, Gal MS. Access barriers to big data. Ariz. L. Rev. 2017;59:339.

Schwartz Paul M. Property, privacy, and personal data. Harv. L. Rev. 2003;117:2056.

Segler MH, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018;555(7698):604.

Stebbing J, Phelan A, Griffin I, Tucker C, Oechsle O, Smith D, Richardson P. COVID-19: combining antiviral and anti-inflammatory treatments. The Lancet Infectious Diseases 2020;20(4):400–2.

Symons T, Bass T, Alegre PB. Me, my data and I: The future of the personal data economy. European Union, Horizon; 2017. p. 2020.

Tsarsitalidis S, Corrales Compagnucci M, Kousiouris G, Dahi A. Feeding Smart Contract Legal Requirements with Semantic and Event Detection Logic Structures from Modern Service Oriented Supply Chains. Smart Contracts: Technological, Business & Legal Perspectives. London: Hart Publishing; 2021. p. 145–60.

Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. Nature Medicine 2020;26(4):459–61.

Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56.

Vaishya R, Javaid M, Khan IH, Haleem A. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 2020.

Varian H. Artificial intelligence, economics, and industrial organization. The Economics of Artificial Intelligence: An Agenda. University of Chicago Press; 2018.

Vayena E. Value from health data: European opportunity to catalyse progress in digital health. The Lancet 2021;397(10275):652–3. doi:10.1016/S0140-6736(21)00203-8.

Voigt TH, Holtz V, Niemiec E, Howard HC, Middleton A, Prainsack B. Willingness to donate genomic and other medical data: results from Germany. European Journal of Human Genetics 2020:1–10.

Yang G. Office Operating Problem Scoring System Based on AI. In: Hui YANG, editor. Artificial Intelligence: Science and Technology, Proceedings of the 2016 International Conference (AIST 2016), Shanghai, China; 2017. p. 21.

Yella JK, Yaddanapudi S, Wang Y, Jegga AG. Changing trends in computational drug repositioning. Pharmaceuticals 2018;11(2):57.

Yu Kun-Hsing, Berry Gerald J, Rubin Daniel L, Ré Christopher, Altman Russ B, Snyder Michael. Association of omics features with histopathology patterns in lung adenocarcinoma. Cell systems 2017;5(6):620–7.

Yu H. Redefining responsible research and innovation for the advancement of biobanking and biomedical research. Journal of Law and the Biosciences 2016;3(3):611–35.

Wachter RM, Cassel CK. Sharing Health Care Data With Digital Giants: Overcoming Obstacles and Reaping Benefits While Protecting Patients. JAMA 2020;323(6):507–8.

Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome medicine 2015;7(1):41.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, … Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 2016;3(1):1–9.