Towards Legal Interoperability in International Health Research

by

Adrian Thorogood

A thesis submitted in conformity with the requirements for the degree of Master of Laws
Faculty of Law
University of Toronto

© Copyright by Adrian Thorogood, 2019
Towards Legal Interoperability in International Health Research

Adrian Thorogood

Master of Laws

Faculty of Law
University of Toronto

2019

Abstract

Interoperability is the ability to exchange meaningful data across complex systems. Connectivity is essential for international health research, but can be hindered by incompatible laws and legal tools. First, this paper considers compatibility between intellectual property law regimes and research data licenses. Second, it reviews data privacy law regimes and how treaties, codes of conduct, and standard safeguards can maintain interoperability across borders. Third, it reviews health research regulation, seeking to reconcile the tension between openness and participant/community control over research data. This paper concludes with a critique of two generic strategies for promoting legal interoperability. The first is access governance, where policies and due diligence processes interface between different legal and organizational contexts. The second is algorithmic access, where researchers submit algorithms to run on secure datasets. Ultimately, interoperability is an essential consideration when designing norms and legal tools to support international health research.
Acknowledgments

I dedicate this thesis to my daughter Sophie. I have witnessed a miracle watching you grow this past year.

Above all, I thank my wife, Florence, for her incredible support, which allowed me to pursue this project.

Tremendous thanks also to Professor Ollivier Dyens, Professor Bartha Knoppers, Professor Trudo Lemmens, and Matthew Marinett for their invaluable guidance and feedback on drafts of this manuscript.

Finally, I would also like to acknowledge the generous financial support of the Canadian Open Neuroscience Platform Scholar Award.
# Table of Contents

Acknowledgments........................................................................................................................................ iii

Table of Contents........................................................................................................................................ iv

List of Tables ................................................................................................................................................... viii

List of Figures................................................................................................................................................ ix

Chapter 1 Introduction .................................................................................................................................. 1

Chapter 2 A Theory of Legal Interoperability for Health Research .......................................................... 9

1 A General Theory of Interoperability ........................................................................................................ 11

2 “Legal” Interoperability ............................................................................................................................ 17

3 Healthcare and Interoperability ................................................................................................................ 21

4 Personal Data Portability ........................................................................................................................... 26

5 Health Research Interoperability ............................................................................................................ 28

Chapter 3 Interoperability of Intellectual Property Systems ....................................................................... 38

1 Introduction to IP ....................................................................................................................................... 38

1.1 Copyright ................................................................................................................................................ 39

1.2 *Sui Generis* Database Rights ................................................................................................................ 44

1.3 Standard Licenses ................................................................................................................................... 46

2 Towards Interoperable IP Systems ........................................................................................................... 50

2.1 Harmonization ......................................................................................................................................... 50

2.2 Standardization ....................................................................................................................................... 51

2.3 Other Strategies ....................................................................................................................................... 52

3 IPRs and Research Data: Uncertain Application and Poor Fit? ................................................................ 52

3.1 Case Study: Licensing and the BRCAExchange Public Genetic Variant Database ............................. 53

3.2 Ownership of Research Data ................................................................................................................... 54
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Practical Barriers to Algorithmic Access</td>
<td>180</td>
</tr>
<tr>
<td>4</td>
<td>Legal and Regulatory Barriers to Algorithmic Access</td>
<td>184</td>
</tr>
<tr>
<td>4.1</td>
<td>Copyright and Database Rights</td>
<td>184</td>
</tr>
<tr>
<td>4.2</td>
<td>Data Privacy Law</td>
<td>185</td>
</tr>
<tr>
<td>4.3</td>
<td>Health Research Regulation</td>
<td>187</td>
</tr>
<tr>
<td>4.4</td>
<td>Other Legal Concerns</td>
<td>189</td>
</tr>
<tr>
<td>5</td>
<td>Achieving Federated Analysis</td>
<td>191</td>
</tr>
</tbody>
</table>
List of Tables

Table 2.1 Benefits and Risks of Interoperability

Table 2.2 The FAIR Principles

Table 4.1 BRCAExchange Source Database Licenses

Table 5.1 OECD Fair Information Processing Principles
List of Figures

Figure 2.1 Unintended Consequences of Hyper-territoriality in Cyberspace

Figure 2.2 The Stages of Open Science

Figure 6.1. Spectrum of community engagement
Chapter 1
Introduction

In this thesis, I develop a theory of legal interoperability for international health research. I survey legal barriers to the exchange and re-use of health-related data and identify strategies policy-makers, institutions, and researchers can adopt to foster greater legal interoperability. Interoperability is the ability of complex systems to effectively exchange data and render that data useful. Health research is a global, data-intensive, and collaborative endeavour dependent on the exchange of data between researchers. Interoperability in health research can foster collaboration, improve the quality of results, and accelerate innovation. It is common to conceive of interoperability in terms of technology and data. For data to flow, technical systems including hardware and software need to be compatible and connected. For digital information to be meaningful to both humans and machines, the structure, syntax, and semantics of data must be clear. But interoperability goes beyond technology and data. It is also a function of the ability and willingness of human beings to communicate with, and understand, each other. Background societal institutions, including legal systems, also condition interoperability. These four layers of interoperability – technology, data, human, and institutional – are interdependent. Establishing more interoperability in one layer can spur connectivity and innovation in another. In order to progressively achieve interoperability of international health research systems, no single layer can be overlooked.

The law is commonly viewed as an external constraint on data exchange practices in health research. Intellectual property rights condition relationships between researchers and the flow of information between them through the lens of ownership. Data privacy laws impose constraints as well as positive obligations on organizations that collect, use, and disclose personal data, i.e., information about an identifiable individual. Health research regulations ensure the ethical conduct and scientific validity of research. They are founded on bioethics principles of beneficence, non-maleficence, individual autonomy, and justice. These principles are realized through safeguards, including voluntary and informed consent, oversight by research ethics committees (RECs), and respect for communities. These ethical principles and safeguards generally extend to govern the exchange and re-use of research data derived from human subjects. It is not always clear when these three legal regimes apply to research data. When they
do, it is not always clear they are fit-for-purpose. Regardless, a theory of interoperability highlights that legal restrictions are not only external constraints on health research systems. They are also produced through the interactions of participants, researchers, and institutions. In other words, legal restrictions and incompatibilities can be reduced through the judicious selection of legal tools.

The term “legal interoperability” is commonly used to describe compatibility between the legal systems of different jurisdictions. Interoperability is a point somewhere between the two extremes of legal fragmentation and harmonization. Legal fragmentation is where legal systems are so different that there is no possibility for meaningful interactions or exchange of goods, services, or information between them. Legal harmonization is the opposite extreme where legal systems are rendered virtually identical – a utopian vision that tends to be politically if not also practically infeasible. Legal interoperability is a happy mean between these two extremes. Just as two computers do not need to have the exact same operating system and software to exchange emails, two legal systems do not need an identical constitution and court system to be sufficiently compatible to allow for collaboration and data exchange.

Legal interoperability across jurisdictions is a *sine qua non* for global health research. Legal interoperability across jurisdictions is undermined by incompatible legal variation, hyper-territorial application of laws, and data sovereignty rules. I explore important areas of variation in the areas of intellectual property law, data privacy law, and health research regulation across jurisdictions, with a focus on Canada, the United States, and the European Union. These jurisdictions already support a considerable number of global health research collaborations. Despite significant similarities between their legal and political systems, researchers still face considerable legal barriers to collaboration. I consider strategies for reducing variation that constrains health research.

Another problem for international health research is hyper-territoriality, the tendency towards overlapping claims of state sovereignty over the same data transaction. States may decide to assert jurisdiction based on the location where health research is conducted, as well as where the resulting data are stored, accessed, and re-used. The application of multiple laws to a single data transaction can result in an unintended accumulation of restrictions or policy conflicts, where two sets of conflicting rules apply simultaneously. Hyper-territoriality is not
merely a legal issue: “[j]urisdictional conflict is the realpolitik of the Internet age”. While cyberspace was once a place to escape regulation, we now live in an age of hyper-territoriality, where states vie to assert their sovereignty over data processing activities even in the absence of a clear territorial nexus. In turn, they intrude on the sovereignty of other states. The problem of hyper-territoriality is significant for health research, as data are commonly aggregated from multiple locations and then redistributed to researchers worldwide. This can attract the application of a plurality of legal systems, and raises serious prospects of policy conflict. Finally, states may even attempt to retain sovereignty over data directly by imposing localization rules to physically sequester data within geographic borders. These three challenges – incompatible variation, hyper-territoriality, and data sovereignty – all pose a heightened threat in the health research context, given the increasingly open and global character of scientific communication and collaboration.

There is also persistent uncertainty over the application of intellectual property rights, data privacy law, and health research regulation to data-intensive health research. For example, when are research data and databases sufficiently original to be works protected by copyright? When are bioinformatics data sufficiently rich that they inherently relate to an identifiable individual and pose privacy risks? When does research involving only data derived from humans qualify as research involving human subjects, and merit the safeguards of informed consent, REC review, and community governance? Across these three domains, data scientists face extraordinary uncertainty over what rules apply to their activities, which tends to reinforce a culture of caution. This caution is aggravated where researchers collaborate across borders. International research collaborations will tend to align their governance with the rules of the strictest jurisdiction to ensure compliance and ensure researchers from that jurisdiction can participate.

There are legal tools available that researchers can use to enhance the interoperability of data, within the limits of legal systems. Putting standard licenses on research outputs (including

---

publications, data, and code) can enable research communities to re-use, build on, and share research outputs without fear of infringing intellectual property rights. A combination of carefully worded consents, standard data sharing contracts, and standard privacy and security safeguards can enable sharing that complies with data privacy laws. Forward-looking data sharing plans, incorporated into the language of research consents, can achieve openness while respecting the autonomy of individual participants. Such plans can assuage the concerns of RECs that the interests of participants will be protected across data sharing ecosystems. The design and selection of legal tools by researchers has important consequences on the downstream interoperability of research data. Indeed, legal interoperability can additionally be characterized as the compatibility of legal tools associated with research data, such as licenses, consents, and data sharing agreements.

Unfortunately, researchers are often left to their own devices to navigate applicable laws, narrow and uncertain legal exceptions for research use of data, and cross-border divergences. They are also frequently left to develop and select legal tools that balance openness and control. In light of these pressures, researchers often put little thought into the downstream compatibility of these legal tools. The public interest in health research interoperability is thus largely left to the vicissitudes of private-ordering by researchers. Public-ordering is where law determines the rules of a system, while private-ordering is where individual actors are left to negotiate the rules. In addition, health researchers receive inadequate support from legislators, policy-makers, or institutions for the selection and development of legal tools. This lack of support limits their ability to achieve optimal levels of legal interoperability.

This thesis reviews common legal and regulatory barriers to interoperability in health research, and evaluates the effectiveness of private-ordering strategies to overcome them. Examples of important barriers are drawn from three legal and regulatory domains: intellectual property law, data privacy law, and health research regulation. The consequences of these barriers, and potential strategies to address them, are illustrated using case studies of international health research collaborations. A first case study is the BRCAExchange, a publicly available database that aggregates data about genetic variants in the BRCA1/2 genes and their associated cancer risk. The BRCAExchange is used to illustrate confusion over the application of copyright and database rights to research databases, and to highlight the importance of clear and interoperable research licensing practices.
A second case study is the Human Cell Atlas, a proposed global reference map of gene expression across all human cell types. While all cells in the human body share the same genetic blueprint, it is largely differences in levels of gene expression that account for the incredible diversity between cell types. Genes can be up-regulated or down-regulated to generate different levels of messenger RNA molecules, and in turn different levels of proteins, the molecules that carry out actual functions within a cell. Rich data are needed to understand the links between gene expression, health, and disease. Given that there are around 30 trillion cells in the human body, such an atlas can only be constructed through international collaboration. A major legal challenge for the Human Cell Atlas is the negotiation of divergent data privacy laws.

The final case study I consider is the Canadian Open Neuroscience Platform. This Platform is a catalogue and repository of neuroscience datasets, which include neuroimaging, genetic, and health-related data. Spanning 31 Canadian and international partner institutions, the Platform aims to support a shift towards open science. By improving accessibility to the datasets and software supporting research findings, open science aims to improve the rigor and reproducibility of results, support collaboration, and accelerate discovery and innovation. The experience of the CONP demonstrates the challenges of complying traditional health research regulation principles and safeguards in the context of open, collaborative, and data-intensive research environments. Safeguards including informed consent, REC review and oversight, and community engagement were initially developed to uphold ethical conduct and scientific validity in clinical experiments. Unfortunately, these safeguards to not scale well to fit interconnected research networks.

One persistent myth this thesis aims to dispel is that legal and ethical barriers are fixed, external constraints on health research collaboration and data exchange. Researchers can adopt self-help strategies such as standard licenses, data sharing contracts, and broad consents to mitigate default legal barriers. However, these legal tools must themselves be interoperable to enable the aggregation, exchange, and re-use of associated research data. Policy-makers should therefore be hesitant to rely solely on researchers to achieve interoperability targets, as private-ordering strategies have a number of predictable weaknesses. First, they tend to involve significant transaction costs, such as assessing what activities are (or are not) permitted with research data, as well as negotiating new permissions. Indeed, researchers may need to negotiate access to and re-use of research data with multiple rights-holders and authorities, including the
researchers or research institutions who hold the data (intellectual property rights holders),
individuals (data subjects protected under data privacy laws and human subjects protected under
health research regulation), oversight bodies (e.g., RECs or data access committees), and even
participant communities. Restrictions and transaction costs can compound across legal domains,
all of which are structured by default to require *ex ante* permissions for data re-use. Systems of
*ex ante* permissions do not scale well along with the growing data-intensivity and connectivity of
international health research.

A related issue is the problem of mis-labelling, which arises where researchers lack the
knowledge, resources, or incentives to adopt open and interoperable legal tools. Even assuming a
general willingness to collaborate, mis-labelling can result in unintended downstream
restrictions. For example, researchers often use non-standard licenses when publicly releasing
research outputs (e.g., databases). This can increase uncertainty and friction for downstream
users. Data sharing contracts often include strict privacy and security safeguards that
unintentionally restrict the freedom of accessing researchers to openly share their research
results. Consents and REC approvals pursued at the beginning of research projects often fail to
articulate plans for long-term data management and sharing. Private-ordering tools also have a
tendency to proliferate, leading to unnecessary variation and new incompatibility issues. Even
where effective private-ordering strategies exist, the interoperability of international health
research depends on the willingness of researchers and other stewards of health data to adopt
them. Over-reliance on private-ordering risks creating a gap between interoperability targets and
outcomes.

I canvass a number of solutions to address the weaknesses of private-ordering strategies
in health research. Top-down (regulatory/policy) or bottom-up (private-ordering) initiatives can
be used to clear many of the legal rights associated with research data. Some legislation or
binding policy is necessary to mandate openness and interoperability standards, e.g., sharing of
(publicly funded) research data, as well as adoption of standard research licenses, consent
language, and privacy and security safeguards. Efforts can also be made to raise awareness about
the importance of interoperability among all stakeholders involved in the governance of health
research data. These include regulators, policy-makers, research institutions, hospitals, industry,
researchers, participants, and communities. Education efforts should cover the promise and
inherent risks of interoperability as well as how it can be fostered. Policy awareness can also be
built into technological systems. Explicit, standard, and machine-readable metadata can be persistently attached to research datasets in order to communicate who data can be shared with, what data can be used for, and under what conditions.

Another set of potential solutions focus on institutional supports. The role of technology transfer offices could be adapted to offer help not only with asserting ownership over and extracting rents from data, but also with applying standard, permissive licenses to research outputs that benefit science and society. Research repositories could play a greater role to ensure submitted datasets are accompanied with standard, open-conformant licenses. RECs could ensure that researchers make and respect commitments to publish results openly and to share underlying data. Indeed, data availability can improve the reproducibility and validity of research, and thus its ultimate social value and ethical justifiability. Openness can also align with the altruistic motivations of participants to maximize the scientific and societal impact of their contributions. It is not simply a threat to individual privacy.

In situations where legal restrictions and conditions cannot be cleared from research data, interoperability may still be pursued through innovative access governance systems and access technologies. Access governance encompasses policies, processes, and safeguards. Access policies articulate who can access research data and for what purposes. Access processes typically involve due diligence review by a data access committee to ensure the trustworthiness of requestors and the scientific quality of their proposed project. Trustworthiness is often assessed with reference to institutional affiliations and project-specific REC approvals. Safeguards include data use agreements with provisions on intellectual property, academic credit, privacy and security safeguards, and authorized uses (e.g., the approved project only). Access governance systems act as “interfaces” between the legal context in which data are generated and the legal context in which data are used, providing both certainty and legal interoperability. As the number of access governance systems increases, however, so does the importance of interoperability between them. Standardization, coordination, and streamlining are necessary to avoid long and duplicative review processes and incompatible use conditions.

Access technologies are being develop that change the very nature of data exchange in order to circumvent technical and legal barriers. The exchange of very large research datasets is increasingly technically infeasible. Intellectual property, data privacy law, and health research
regulation present numerous barriers to the international distribution of research data. Techno-solutionists propose that interoperability be pursued at the level of technology so as to reduce tension at the data, human, and institutional layers. Instead of distributing copies of data to researchers, they propose that algorithms be granted access to secure datasets. Algorithmic access means only the results of analyses are released to researchers. Proponents of artificial intelligence and machine learning in healthcare propose a federated analysis strategy where algorithms are trained across a network of secure healthcare (cloud) databases. Patient data never has to leave the hospital’s computing environment, let alone the country. Research algorithms can still be run or trained across large virtual cohorts. A similar strategy is proposed for research on clinically-generated genomic and health-related data. Analysis pipelines can be distributed across computing environments, and the results of local analysis can be aggregated.

While promising, algorithmic access strategies tend to downplay unavoidable human and legal challenges. Dedicated technical standardization efforts are needed to establish interoperable software, computing environments, and network interfaces. Technical standardization efforts rely heavily on collaboration between numerous stakeholders. They raise important governance challenges over who participates in standard-setting processes, and who can access standards. Ultimately, algorithmic access systems will still depend heavily on collaboration. Extensive negotiations are needed to convince hospitals to expose their computer systems to algorithmic access; to invest in computing infrastructure and data scientists; to implement and stay abreast of technical standards for data analysis and exchange; and to maintain high-quality data in standard, structured formats. In short, if it is difficult to convince people to give up control over their data, try convincing them to give up control over their computing systems. Interoperability is never solely a characteristic of technology and data. While technologies like algorithmic access will certainly contribute to health research interoperability, human beings and the laws that govern them will necessarily remain part of the equation.

---


Chapter 2
A Theory of Legal Interoperability for Health Research

**Key Points:**
- Interoperability is the ability of complex systems to exchange meaningful data.
- While data-focused, interoperability is progressively achieved at multiple interconnected layers: technological, data, human, and institutional.
- Viewing complex systems through the lens of interoperability reveals systemic barriers and facilitators to data flows, as well as their risks and benefits.
- Legal interoperability refers to the compatibility of legal systems across jurisdictions, or to the compatibility of legal tools associated with data such as licenses, contracts, or consents.
- Law is typically viewed as a barrier to interoperability. US electronic health record regulations and the European consumer right to portability illustrate law as a facilitator.
- Health research policy is emphasizing greater openness of scientific knowledge, which can only meaningfully be achieved if data is interoperable.
- While health research policy has traditionally focused on interoperability at the data layer, attention is now shifting to its human, legal, and technological components.

The success of health research depends on the value and impact of the data and knowledge it produces. This in turn depends on how often and effectively scientific data and knowledge are exchanged, built on, and re-used. While interoperability is traditionally a technical concept, used in data science and in computer and networking contexts, it can also shed light on the human, organizational, and legal factors influencing the exchange of meaningful data across complex systems. A theory of interoperability helps to reveal why data flows or does not flow across international health research systems. By recognizing interoperability’s barriers and facilitators, its systemic risks and benefits, policy-makers can make informed decisions about when and when not to enhance connectivity.

In this thesis, I aim to convince you that interoperability is also an important legal principle, which can support global collaboration in data-intensive health research environments. Interoperability may be pursued through bottom-up approaches, such as informal or formal
standard-setting. Interoperability can also be articulated top-down as an explicit principle of law making. Legislators seeking to promote collaboration and data exchange between jurisdictions will aim to ensure legal regimes can inter-operate. The importance of interoperability is highlighted even where states seek to undermine it. The threat of non-interoperability may be used to extract concessions from a trading partner on matters such as human rights or privacy protection.

Interoperability may even be a legal mandate. In the US, electronic healthcare record legislation actively promotes interoperability by standardizing privacy and security protections, and by prohibiting data blocking. Data blocking is the intentional, unjustified withholding of patient information requested by other healthcare networks and providers. In Europe, data protection law mandates that organizations collecting personal data from consumers must maintain the data in interoperable form. This protects the consumer’s right of portability, the right to take their data and business elsewhere. Legally mandating interoperability may be necessary where organizations have competitive incentives to maintain the data of patients or consumers in silos.

Similarly, interoperability in health research can be enhanced through bottom-up approaches and top-down approaches. The former include efforts to establish research community standards. The latter may include interoperability requirements as part of funding agency open access and data sharing mandates. For example, given concerns over unclear or incompatible intellectual property licenses attached to research resources, research funding agencies could mandate that data, documents, and software be released under standard licenses permitting re-use, adaptation, and sharing by anyone with limited conditions or restrictions. Privacy and security standards may be established to enable trustworthy exchange between research organizations subject to different data protection laws. Interoperability may even deserve ethical recognition, as interoperability affects the ultimate benefit-risk ratio of research. There are, however, some trends in health research regulation that threaten interoperability. In particular, granting participants, research ethics boards, and patient communities various forms

---

of control over data may lead to an accumulation of restrictions on the re-use, aggregation, and sharing of data. At the very least, awareness of the importance of interoperability should be promoted among all stakeholders involved in research data governance.

This Chapter begins with a general theory of interoperability, which highlights the influence of technological, data, human, and institutional factors on data exchange. I then explore the meaning of legal interoperability. One common characterization of legal interoperability is compatibility between legal regimes across jurisdictions that enable collaboration and data exchange. Another characterization refers to compatibility of legal tools, such as licenses (see Chapter 4). Legal interoperability is increasingly problematic in cyberspace – once viewed as a borderless world but now characterized by hyper-territoriality and multiple, overlapping laws. I then explore the emergence of interoperability as a legislative concept in US electronic health records regulations and European data protection law. These regimes aim to promote the portability of consumers’ personal data between organizations. Finally, I introduce the open access and data sharing movements in health research policy. While openness and interoperability are not synonymous, they are related. The aims of openness to improve reproducibility and accelerate research depends on the quality and interoperability of the scientific resources made available. Interoperability is increasingly recognized as a valuable quality of health research data. But health research policy-makers also need to consider the technological, human, and legal layers of complex systems that determine overall levels of interoperability.

1 A General Theory of Interoperability

John Palfrey and Urs Gasser provide a general theory of interoperability in the context of “complex systems that rely upon a constant exchange of information, most commonly mediated by digital and networked technologies.”\(^5\) They define interoperability as “the ability to transfer and render useful data and other information across systems, applications, or components.”\(^6\) The


\(^6\) Ibid at 3.
focus of such a theory is on how to “manage the unprecedented degree of interconnectivity that has been created between and among people and systems in the digital age”, and how to “maximize the benefits of this unparalleled and growing level of connection and information flow while minimizing its potential risks.”

The theory considers interoperability across four layers:

1. **Technological** - “hardware and code allowing physical connection”, often enabled by an explicitly agreed-upon interface,
2. **Data** – “ability to understand and process what is being transmitted”,
3. **Human** - “ability for humans to understand and act on the data”, and
4. **Institutional** - “ability of societal systems to engage effectively”.

The legal system is one example of an institutional layer of interoperability (or its absence). The general theory goes beyond technical definitions of interoperability, which tend to focus only on technology and data.

The European Commission also has a holistic interoperability framework for public administration. The Commission defines legal interoperability as the ability of organisations to interact towards mutually beneficial goals, involving the sharing of information and knowledge between these organisations, through the business processes they support, by means of the exchange of data between their ICT systems.

Indeed, important components of European and Member State public administrations are implemented through ICT systems, which are referred to as eGovernment systems.

---

9 Palfrey & Gasser, *supra* note 5 at 3.
10 Catherine Doldirina et al, “Legal Approaches for Open Access to Research Data” (2018) at 9 (for example: “Interoperability, as typically used by the computer and information science communities, means the ability of information systems to work with each other because their interfaces are completely understood even when individual components are technically different and managed by different organizations. Interoperability includes technical, syntactic and semantic interoperability. Technical interoperability is usually associated with hardware and software components, systems and platforms that enable machine-to-machine communication. Syntactic interoperability is usually associated with data formats and provides for the exchange of clearly defined classes of data. Semantic interoperability is concerned with ensuring that the precise meaning of exchanged information is understandable by any other application and is the ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results.”).
European Commission also defines four layers of interoperability: technical, semantic, legal, and organizational.\textsuperscript{12} The one distinction is that the European Commission highlights legal interoperability instead of human interoperability. This likely reflects the context of public administration, where data is processed and exchanged to fulfill legal mandates. The Framework characterizes legal interoperability as being “about ensuring that organisations operating under different legal frameworks, policies and strategies are able to work together.”\textsuperscript{13} The European Commission’s definition of organizational interoperability captures some of the human elements highlighted by Palfrey and Gasser. According to the Commission, organizational interoperability “refers to the way in which public administrations align their business processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals.”\textsuperscript{14} The framework encourages European and national law and policy-makers to make interoperability an explicit public policy objective. New policies have interoperability consequences both for the national eGovernment systems in which they are implemented, and across borders. The framework even recommends that special checks be in place to screen legislation for interoperability.\textsuperscript{15}

Interoperability increases interconnectivity in ways that can lead to systemic benefits and risks, which depend on both the means of implementation and the resulting degree (see \textbf{Table 2.1}).

\begin{table}[h!]
\centering
\begin{tabular}{|c|p{\textwidth}|}
\hline
\textbf{Benefits}\textsuperscript{16} & \\
\hline
Innovation & Interconnections enable increased innovation. An example of this is Facebook’s open API which has created a huge market of third party app developers. \\
\hline
Competition & Interoperability lowers barriers to entry and reduce lock-in (allowing switching between technologies or services). An example where this is a problem is lock-in by electronic health record vendors. \\
\hline
\end{tabular}
\end{table}

\textsuperscript{12} \textit{Ibid} at 22.

\textsuperscript{13} \textit{Ibid} at 27.

\textsuperscript{14} European Commission, \textit{supra} note 11 rec 3.4.

\textsuperscript{15} \textit{Ibid} rec 27.

\textsuperscript{16} Gasser, \textit{supra} note 8 s 3.
| Autonomy, flexibility, and choice | Interoperability allows for portability, or even allows users to mix and match different technologies to achieve a specific purpose. USB ports allow a variety of competing accessories to be plugged into computers. |
| Access, diversity and openness | Interoperability enables access to data by a wider range of different systems. An example is the single sign-on authentication provided by Facebook or Google which allows users to access various websites with a single password. |

<table>
<thead>
<tr>
<th><strong>Risks</strong>&lt;sup&gt;17&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Decreased security</strong></td>
</tr>
<tr>
<td><strong>Decreased privacy</strong></td>
</tr>
<tr>
<td><strong>Increased Homogeneity</strong></td>
</tr>
<tr>
<td><strong>Decreased reliability</strong></td>
</tr>
<tr>
<td><strong>Decreased accountability</strong></td>
</tr>
</tbody>
</table>

---

<sup>17</sup> Ibid s 4.


<sup>19</sup> Josh Fruhlinger, “What is the Heartbleed bug, how does it work and how was it fixed?”, (13 September 2017), online: *CSO Online* <https://www.csoonline.com/article/3223203/what-is-the-heartbleed-bug-how-does-it-work-and-how-was-it-fixed.html>.

<sup>20</sup> Gasser, *supra* note 8 s 4.4.
where responsibility can be determined, interoperability makes it difficult to hold specific parties accountable, e.g., due the absence of a contractual relationship between all of these parties and the consumer.

A theory of interoperability can provide descriptive value. It encourages organizations to recognize that they are part of complex systems, and to establish interoperability as an explicit concern or organizational goal. It encourages them to consider how a particular decision might impact their ability to collaborate with others. A theory of interoperability reveals both the benefits as well as the risks and costs of “deep interconnection among technologies, data, humans, and institutions.” It therefore frames the design of complex systems in terms of achieving connectivity’s benefits while managing the inherent risks. It highlights the costs of sub-optimal interoperability, in terms of efficiency, innovation, competition, and accessibility. It allows comparisons with other complex systems such as air-traffic control systems, electrical smart grids, or healthcare information technology (IT) systems.

Considering interoperability in other contexts reveals a typology of benefits and risks that may otherwise be overlooked. Examples of benefits include “generativity”, where interoperability at the technology layer (e.g., Internet standards), allows for an explosion of activity at the data layer (e.g., content sharing). Examples of risks include lock-in, where a system struggles to adopt a superior technology or practice because of high switching costs and collective action problems. Levels of interoperability in other fields of human activity provides benchmarks. For example, the drastic increase in financial information exchange contrasts sharply with the limited exchange of data in healthcare. Sectoral comparisons with other sectors also reveal how difficult it can be to define, achieve, and maintain optimal interoperability over time. Optimal levels of interoperability rarely come about as a result of individual decision-making; rather, they come about from a combination of concerted collaboration between individual actors and input from state actors. It can be difficult for law- or policy-makers to mandate interoperability: “regulatory effects are harder to predict as traditional legal instruments

21 Palfrey & Gasser, supra note 5 at 256–7.
22 Ibid at 262.
23 Jonathan Zittrain, The Future of the Internet–and How to Stop It (Yale University Press, 2008) at 75.
trigger side effects, cascading effects and unintended consequences.” Comparisons also reveal a typology of strategies private institutions, collaborations, and states can adopt to direct complex systems towards optimal interoperability.

A theory of interoperability also breaks down conceptual silos between technical and societal issues. Interoperability “is not just about the flow of data or about technology; it involves essential questions of human and institutional interaction as well…” It “depends on how well human beings and institutions can work together.”

Interoperability is also valuable as a normative theory, forcing us to ask “what we want out of all this connectivity.” Recall that seeking interoperability does not equate with seeking complete harmonization of technological, organizational, or legal systems. Interoperability gives collaborating entities flexibility by encouraging them to distinguish between the aspects of their systems that need to be the same to enable data exchange, and those that do not. Seeking interoperability does not preclude protecting desirable levels of diversity, which can allow for innovation through experimentation. For example, radical levels of interoperability may be desirable if the goal is to support disruptive innovation. Generativity is a particularly disruptive form of innovation, commonly associated with Internet technologies. “Generativity is a system’s capacity to produce unanticipated change through unfiltered contributions from broad and varied audiences.” The inputs of generative systems are unfiltered, participatory contributions. The outputs are unanticipated forms of innovation. The path of a generative system, however, is

---


25 Palfrey & Gasser, supra note 5 at 3.

26 Ibid at 256.

27 Ibid at 2.

28 Ibid at 52.

29 Zittrain, supra note 23 at 75.
unpredictable: “Generative technologies need not produce forward progress, if by progress one means something like increasing social welfare. Rather, they foment change”.  

To give an example of a potentially generative activity: requiring research publications to be available open access may well encourage the broad accessibility of scientific knowledge, which can accelerate, coordinate, and improve the overall efficiency of research. Going one step further, however, would be to allow those publications to be available as an input for the development and application of emerging text and data mining approaches. Unanticipated uses of publication libraries are the kind that will lead to surprising discoveries.

A theory of interoperability forces policy-makers to make explicit trade-offs between the inherent benefits and risks of connectivity. There may be legitimate reasons to constrain interoperability, but these will necessarily constrain innovation. Where policy-makers seek to achieve higher levels of innovation, they must be willing to establish significantly higher levels of interoperability and accept the accompanying risks.

2  “Legal” Interoperability

“Legal interoperability, broadly defined, is the process of making legal norms work together across jurisdictions.” It concerns the “process of making legal rules cooperate across jurisdictions.” Excessive fragmentation between legal rules across jurisdictions can discourage commerce and collaboration. Full harmonization of legal rules is infeasible, and even where it is achieved may be undesirable from the perspective of regulatory diversity and competition. Legal interoperability aims for an optimal point between these two extremes. The conditions that the World Trade Organization imposed on China’s intellectual property regime in order to gain

30 Ibid at 84.
32 Gasser, supra note 8 at 25.
accession to the treaty are an example of states seeking greater legal interoperability.\textsuperscript{34} The European Union is explicitly dedicated to improving interoperability between the legal systems of its member states, particularly in areas that affect online commerce (e.g., copyright and contract law).\textsuperscript{35}

While legal interoperability has long been a concern for international trade systems, it has become an even greater concern following the advent of the Internet and greater digital connectivity.\textsuperscript{36} The classic Westphalian system of international cooperation relies on the principles of sovereignty exercised over a geographical territory comprising the state, and non-interference with the sovereignty of other states. The Internet has produced no end of potential for territorial conflicts, given the essential borderless-ness of cyberspace. As Bertrand de la Chapelle and Paul Fehlinger note, “because the Internet is borderless, states are faced with the need to regulate conduct or subject matter in contexts where the territorial nexus is only partial and in some cases uncertain.”\textsuperscript{37} This creates concern about hyper-territoriality, where national jurisdictions attempt to extend the application of their laws to activities more closely connected to other jurisdictions.

A leading example was the \textit{US v Microsoft} Supreme Court case, where US law enforcement demanded access to data held by a US cloud service providers on foreign servers.\textsuperscript{38} Another is the explicit extra-territorial application of the European \textit{General Data Protection Regulation} (GDPR), discussed in Chapter 5, which can impose data protection obligations on companies outside of Europe.\textsuperscript{39} Indeed, “[e]xtraterritorial extension of national jurisdiction is

\begin{itemize}
  \item \textsuperscript{34} World Trade Organization, \textit{Accession of the People’s Republic of China}, WT/L/432 (2001).
  \item \textsuperscript{35} Gasser, \textit{supra} note 8 at 26.
  \item \textsuperscript{36} Chapelle & Fehlinger, \textit{supra} note 1 at 1.
  \item \textsuperscript{37} \textit{Ibid} at 2.
  \item \textsuperscript{38} Sarah Jeong, “The Supreme Court Fight Over Microsoft’s Foreign Servers is Over”, (5 April 2018), online: \textit{The Verge} <https://www.theverge.com/2018/4/5/17203630/us-v-microsoft-scotus-doj-ireland-ruling>. This was followed by the US, \textit{Clarifying Lawful Overseas Use of Data (CLOUD) Act} (HR 4943), 15 cong div V (2018).
  \item \textsuperscript{39} \textit{General Data Protection Regulation}, Europe [\textit{General Data Protection Regulation}].
\end{itemize}
becoming the realpolitik of Internet regulation.”\textsuperscript{40} This creates a paradox of sovereignty as it breaches the principle of non-interference. La Chapelle and Fehlinger propose that a Kantian categorical imperative … should underpin international Internet regulation: Any national policy measure that would be detrimental if generalized around the world should not be adopted in the first place.\textsuperscript{41}

States are also attempting to strengthen digital sovereignty by creating requirements that strengthen the nexus between territory and data, including data localization requirements, strong international intermediary liability regimes, local office requirements, and in some cases full-blown licensing regimes.\textsuperscript{42} The simple threat of exercising digital sovereignty may be enough for states to extract compliance from foreign companies through direct public-private requests (e.g., to remove domains, take down infringing content, or provide law enforcement access to user data).

Traditional tools for international legal cooperation, however, do not seem up to the task of coordinating regulation of the internet. Conflict of laws rules, at least in private international law, provide procedures for determining what law applies in a particular international dispute, but do not resolve underlying differences.\textsuperscript{43} Mutual Legal Assistance Treaties (MLATs) are slow and require “dual incrimination”: a breach of the laws of both countries.\textsuperscript{44} Approaches to achieve interoperability range across a spectrum.\textsuperscript{45} Harmonization is the most heavy-handed approach, which imposes a desired level of cooperation through legal mechanisms. Standardization establishes generally accepted guidelines or practices in a technical, economic, or legal area, strengthened through common, repeated actions. Mutual recognition is a compromise whereby one state accepts another state’s regulation as “good enough”. Cooperation is usually pursued

\textsuperscript{40} Chapelle & Fehlinger, supra note 1 at 3.
\textsuperscript{41} Ibid at 5.
\textsuperscript{42} Ibid at 4.
\textsuperscript{43} Weber, supra note 33 at 6.
\textsuperscript{44} Chapelle & Fehlinger, supra note 1 at 6.
\textsuperscript{45} Weber, supra note 33 at 8.
through networks of regulators who agree on responsibilities to address international challenges. ⁴⁶

There are a number of direct monetary costs of legal fragmentation, and these are only likely to increase as societies become more networked. ⁴⁷ Resulting hyper-territoriality, particularly in cyberspace may also have a number of unintended consequences (see Figure 2.1). Hyper-territoriality may also aggravate inequalities of power and wealth across countries, as more powerful countries extend their influence, potentially limiting upward mobility of the poorer ones.

**Figure 2.1 Unintended Consequences of Hyper-territoriality in Cyberspace** ⁴⁸

Legal interoperability can also refer to compatibility between legal tools associated with data, such as licenses, contracts, and consents. This view of interoperability is elaborated in Chapter 4, which explores the importance of license interoperability to enable the combination, re-use, and sharing of health research datasets. While law or policy may provide some direction,

---


⁴⁸ Chapelle & Fehlinger, *supra* note 1 at 7.
researchers, hospitals, and data repositories often have some freedom to select these legal tools. In this sense, legal interoperability straddles both the institutional and human layers of interoperability.49

3 Healthcare and Interoperability

Interoperability of healthcare data is increasingly being mandated by policy-makers or even by law. This recognizes that patient data can meaningfully inform the care of other patients and can support research to improve the care of future patients. Internationally, the Organisation for Economic Co-operation and Development’s Council on Health Data Governance has developed a framework “to guide health data use and sharing practices, so as to protect privacy, enable efficiencies, promote quality and foster innovative research”.50 The framework recognizes the public health, scientific, and economic value health systems can capture from responsible exchange between healthcare organizations, government agencies, and researchers – both nationally and internationally.51 To facilitate responsible exchange and re-use, the framework recommends efforts to improve the interoperability of healthcare data through

49 Finally, legal interoperability can also encompass legal concepts and terminology as data, data that are exchanged and computed upon. Schartum highlights interoperability challenges relating to “legal semantics” in eGovernment contexts. In order to implement laws and regulations through ICT systems, there is a need for precise and consistent legal definitions. Legislative processes, however, do not traditionally prioritize precise and consistent legal definitions. Such definitions may differ horizontally across laws. Definitions may not be provided, or may only be provided in vague form. This may be done purposefully to allow for flexible application. Definitions may also be continuously elaborated over time through case law. Distinct instances of legal definitions may be derived in different regulatory contexts, though they share a common reference point. Whether this is done intentionally or not, disregard of legal definitions at the level of legislation impacts the ability to implement law into eGovernment systems. “One possible response is to transform questions of defining legal words and phrases from a problem of applying the law to a problem of making the law.” (Dag Wiese Schartum, “Legal Definitions and Semantic Interoperability in Electronic Government” (2015) Yulex 2013 131 at 140.) Schartum suggests that draft legislation could be reviewed for its interoperability impact on administrative and technical systems. Tools could be available to encourage the integration of existing legal definitions into legislation, and to track new definitions added over time. Interoperability is not simply a technical and administrative issue but it begins with the legislative process. In short, “Legislation should always be drafted with [technical] implementation in mind.” (Ibid at 146.) Legal interoperability at the data layer presents a range of interesting questions that are unfortunately beyond the scope of this thesis. I will focus the rest of my discussion on interoperability between legal systems and interoperability between legal tools.


51 Ibid, preamble.
standardization of data elements, formats, and quality control.\textsuperscript{52} It also recommends efforts to improve the interoperability of healthcare data governance, through common policies and processes that determine when data can be accessed and used.\textsuperscript{53}

Interoperability is legally defined and mandated in US healthcare. The US federal government has invested billions of dollars and adopted legislation with the aim of establishing interoperable healthcare information technology (IT) systems.\textsuperscript{54} According to the Office of the National Coordinator for Health Information Technology, systems are interoperable when electronic health information is available and can be securely and efficiently shared, when and where it is needed, to support patient-centered care, enhance health-care quality and efficiency, and advance research and public health.\textsuperscript{55}

The goal is not simple transmission of records, but of clinically meaningful information that “improves patient safety and workflow, enhances value, and enables person-centered care.”\textsuperscript{56} True interoperability, in the view of the US National Academy of Medicine, “is the ability to seamlessly and automatically deliver data when and where it is needed under a trusted network without political, technical, or financial blocking.”\textsuperscript{57} The portability of health information across providers makes it possible to construct a longitudinal patient record, which “enables providers to care for the whole patient rather than a single diagnosis or episode.”\textsuperscript{58}

\textsuperscript{52} Ibid, rec III(2)(i).

\textsuperscript{53} Ibid, rec III(2)(ii). See Chapter 7 on Access Governance for more information.


\textsuperscript{55} The Office of the National Coordinator for Health Information Technology, \textit{Report to Congress on Health Information Blocking} (2015) at 4.

\textsuperscript{56} National Academy of Medicine, \textit{Procuring Interoperability: Achieving High-Quality, Connected, and Person-Centered Care} (2018) at 12–13.

\textsuperscript{57} Ibid at xix.

Important US legislation includes the *Health Information Technology for Economic and Clinical Health (HITECH) Act* (2009)\(^{59}\) and the *21st Century Cures Act* (2016)\(^{60}\). The HITECH Act aims to “drive the rapid adoption of interoperable technologies and services to support the exchange of electronic health information to improve care and efficiency in the US health care system.”\(^{61}\) The *Cures Act* defines an interoperable health IT system as one that

(a) enables the secure exchange of electronic health information with, and use of electronic health information from, other health information technology without special effort on the part of the user; (b) allows for complete access, exchange, and use of all electronically accessible health information for authorized use under applicable State or Federal law; and (c) does not constitute information blocking.\(^{62}\)

The US Department of Health and Human Services (HHS) has recently proposed an updated definition of interoperability in the health ecosystem, which notably encompasses human barriers:

> Interoperability is the ability of different information systems, devices or applications to connect, in a coordinated manner, within and across organizational boundaries to access, exchange and cooperatively use data amongst stakeholders, with the goal of optimizing the health of individuals and populations.\(^{63}\)

Interoperability is defined according to numerous layers: foundational (inter-connectivity requirements), data structure (syntax), data semantics (enabling data to be interpreted and used), and organizational – which “encompasses the technical components as well as clear policy, social and organizational components.”\(^{64}\) The importance of this broader definition was also

---


\(^{61}\) The Office of the National Coordinator for Health Information Technology, *supra* note 55 at 33.

\(^{62}\) *21st Century Cures Act*, 2016 s 4003.


\(^{64}\) *Ibid.*
highlighted in a recent report on health IT procurement: “even if technical interoperability could be achieved, the practice of data blocking and data hoarding limits the flow of information.”

A key concern for US policy-makers is information blocking by healthcare IT developers and exchanges/networks, as well as healthcare providers. This “occurs when persons or entities knowingly and unreasonably interfere with the exchange or use of electronic health information”. Three elements constitute information blocking: interference with exchange, that is made knowingly, and that is objectively unreasonable (in light of other public policy objectives such as privacy and security). A similar but more detailed definition was adopted in the Cures Act. The act also grants authority to investigate information blocking to HHS Office of the Inspector General. Health information developers, exchanges, or networks may be subject to up to $1 million of civil monetary penalties. Information blocking by providers will be reported to an appropriate agency and be subject to as yet undefined “appropriate disincentives”. The Office of the National Coordinator for Health Information Technology is responsible for advancing interoperability and address information blocking, and overseeing a program of voluntary certification of health IT. Despite the existence of certain health data standards, a common view of interoperability is lacking across US healthcare systems. In the absence of a harmonizing or coordinating body across healthcare systems, implementation of these standards is subject to interpretation.

In a 2015 report to Congress, the ONC identified suspect practices indicative of information blocking. These included restrictive contracts, prohibitive prices, non-standard IT implementation, and vendor lock-in. As reflected in the definition of information blocking, the report recognized that not everything qualifies as information blocking. It recommended a comprehensive approach including both targeted actions to deter and remedy information

65 National Academy of Medicine, supra note 56 at 9–10.
66 The Office of the National Coordinator for Health Information Technology, supra note 55.
67 21st Century Cures Act, supra note 60 s 4004.
68 Black, Hulkower & Ramanathan, supra note 58 at 612.
69 The Office of the National Coordinator for Health Information Technology, supra note 55 at 12.
blocking as well as broader strategies to address systemic problems. Systemic problems included when actors may implement technical standards in inconsistent ways; adopt divergent privacy, security, or trust policies that govern how electronic health information is exchanged and used; or engage in other inefficient behaviors that inhibit or reduce opportunities to exchange and use electronic health information to improve care and care delivery.

A proposed rule drafted by the ONC addressed interoperability, information blocking, and health IT certification, and the public identification of “doctors, hospitals, and other healthcare providers who engage in information blocking.” The proposed rule defines “reasonable and necessary activities” that are exceptions and not considered information blocking. They include: 1) preventing harm to a patient or other person; 2) protection of privacy; 3) promotion of security; 4) recovery of reasonable costs; 5) unfeasible requests; 6) licensing of necessary interoperability elements on reasonable and non-discriminatory terms; and 7) temporary interruptions to maintain and improve health IT performance. Each exception is defined by mandatory conditions and qualifiers like “necessary” and “reasonable”.

While the interoperability of health IT systems is a particular concern in the United States, because of the existence of multiple healthcare insurance providers, its importance is also recognized in other jurisdictions. A recent recommendation from the Council of Europe on the protection of health data also highlighted interoperability as key to achieve the benefits of digitization. The recommendation highlighted the importance of “secure, interoperable information systems”, defining interoperability as “the ability to communicate and exchange data

---

70 Ibid at 21.
71 Ibid.
73 The Office of the National Coordinator for Health Information Technology (ONC), Seven Exceptions to the Information Blocking Provision (2019).
75 Council of Europe, Recommendation CM/Rec(2019)2 of the Committee of Ministers to member States on the protection of health-related data (2019).
across different information systems”. Interoperability includes technical and semantic dimensions (i.e., common terminologies between professionals). The benefits of interoperability include better individual care, data portability, public health, and the efficiency and transparency of health systems. The recommendations also emphasize that interoperability must not come at the cost of privacy rights, and that technical standards or certification processes developed to facilitate interoperability should guarantee a high level of security. This example demonstrates the distinction between openness – which typically refers to the removal of legal obligations associated with data – and interoperability – which can involve strict security protections as long as they are roughly similar and compatible as data flows between organizations and countries.

4 Personal Data Portability

Interoperability also finds legal instantiation in data privacy laws, linked to the individual’s right to portability of his or her personal information. The European General Data Protection Regulation (introduced in detail in Chapter 5) includes a robust data subject right to portability, which is a right to receive personal data from a data controller “in a structured, commonly used and machine-readable format” and to “transmit those data to another controller without hindrance”, with some limitations. Where technically feasible, data subjects can also request direct transmission between controllers. The GDPR states that “[d]ata controllers should

76 Ibid s 3.

77 Council of Europe, Explanatory memorandum to Recommendation CM/Rec(2019)2 of the Committee of Ministers to member States on the protection of health-related data (2019) ss 37, 133.

78 Ibid ss 5,9,14.

79 Ibid s 14 (“14.1 Interoperability may help address important needs in the health sector and may provide technical means to facilitate the updating of information or to avoid storage of identical data in multiple databases, and contribute to data portability. 14.2. It is, however, necessary for interoperability to be implemented in full compliance with the principles provided for in this Recommendation, in particular the principles of lawfulness, necessity and proportionality, and for data protection safeguards to be put in place when interoperable systems are used. 14.3. Reference frameworks based on international norms offering a technical structure which facilitates interoperability should guarantee a high level of security while providing for such interoperability. The monitoring of the implementation of such reference frameworks can be carried out through certification schemes.”).

80 General Data Protection Regulation, supra note 39 art 20 (the right only applies where processing is based on consent or contract, and only to personal data provided by the data subject to the data controller).
be encouraged to develop interoperable formats that enable data portability.”

While found in a law focused on privacy, the right to portability has as much to do with competition law, consumer protection, and intellectual property. By giving consumers more control over their data, the right to portability addresses the power imbalance between consumers and companies. It also gives consumers a stronger exit voice – the ability to take their business and personal data elsewhere – which can foster competition between service providers. Portability does not necessarily mean saying good-bye to a controller; it can also support interconnection of services. For technical and competitive reasons, the tendency is for companies to store data in an esoteric form. To address this tendency, the GDPR establishes a standard of interoperability: structured, easy-to-use, and machine-readable.

How this standard is met in terms of a technical implementation remains the challenging interpretive question. The GDPR insists that the obligation is limited to data interoperability, and does not “create an obligation for the controllers to adopt or maintain processing systems which are technically compatible.” But the right does address human elements of interoperability, by insisting it be respected without hindrance. The right of portability promises to increase the overall flow of data between controllers and consumers.

This may be surprising to anyone who considers data protection to be mainly about privacy. According to Palfrey and Gasser, more flow of data inherently means more security and privacy risks. For example, sometimes the personal data of multiple subjects are intermeshed. A balancing decision would need to be made about whether to retain or to share the data. A controller may release the wrong personal data to the wrong person. Hackers may target consumer portals or use identity fraud to extract personal data. The implicit assumption is that these risks are more than offset by consumer and economic benefits.

---

81 Ibid Recital 68.
83 Ibid at 202.
84 General Data Protection Regulation, supra note 39 Recital 68.
The right to portability can also support health research. Data subjects may employ the right to transmit their personal data to or between health researchers (though depending on whether or not the data are “personal”, and on the country, individual access and portability rights may not apply in research contexts). There are already companies that employ access and portability rights on behalf of patients to extract, aggregate, and share healthcare data for research purposes. For example, the company RDMD collects information about patients with rare disease to create a data and recruitment resource for academic and pharmaceutical research. The right to portability provides another example where the law acts as a facilitator, rather than a barrier, to data exchange.

5 Health Research Interoperability

The semantic and technical interoperability of data is naturally a central preoccupation of data science. The popular FAIR principles state that data should be findable, accessible, interoperable, and reusable (Table 2.2).

<table>
<thead>
<tr>
<th>Table 2.2 The FAIR Principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Findability specifies that data should be identified, described and registered or indexed in a clear and unambiguous manner. This implies that datasets are uniquely identifiable using persistent identifiers; that the main data properties are ideally specified using standards; and indexed in a public resource.”</td>
</tr>
<tr>
<td>Accessibility specifies that datasets should be accessible through a step-by-step procedure for data access, preferably by automated means. …</td>
</tr>
<tr>
<td>Interoperability specifies that data as well as metadata are formulated, expressed and organized using common, published standards …</td>
</tr>
<tr>
<td>Reusability specifies the crux of other principles: the characteristics of the data, and their provenance, should be well described, according to established standards.”</td>
</tr>
</tbody>
</table>

88 Ibid.
There is a strong push for greater interoperability in health research. The US National Institutes of Health’s *Strategic Plan for Data Science* highlights the importance of findability, interconnectivity, and interoperability of NIH-funded biomedical data sets and resources.\(^8^9\) The motivations are similar to those articulated for open science: to improve reproducibility of research by enabling validation, refinement, and refutation of results; to facilitate meta-analyses combining data from many studies to increase statistical power; to enable creative re-uses by other researchers; to reduce duplicative research involving unnecessary costs and risks to participants; and to democratize science. Openness can lead to more collaboration across disciplines and across borders, and can even expand who can participate in science.\(^9^0\)

The policy focus in health research has been on enhancing the availability of publications, as well as underlying data and code. Scientific norms have long involved the disclosure of detailed methods and reasoning, as science is a culture of skepticism and scientists have the burden of convincing their peers that findings are valid.\(^9^1\) This culture of skepticism derives from a view that the role of science is to root out error, the sources of which are ubiquitous, and include mistakes, bias, and self-delusion.\(^9^2\) Scientists attempting to establish new findings or understandings must overcome the skepticism of their peers by carefully laying out their methods and rationale.\(^9^3\) Errors may stem from statistical errors, the limits of peer review as a quality control mechanism, or professional incentives to publish or perish. It is doubtful that

---

\(^8^9\) National Institutes of Health, *NIH Strategic Plan for Data Science* (2018) at 2 (“The generation of most biomedical data is ‘highly distributed’, meaning it is carried out by various researchers and research groups. This raises a challenge of interoperability, making data hard to find, use, and combine, and means data requires extensive cleaning.”).


\(^9^2\) *Ibid*.

science is sufficiently self-correcting to remedy these errors. Statistical errors are more relevant where studies have small sample size (and thus smaller statistical power), or test unlikely hypotheses (and scientists tend to seek surprising results). In a number of disciplines including psychology, neuroscience, and pharmaceutical research, there is serious concern that results cannot be computationally reproduced, let alone replicated in other contexts. The resulting lack of generalizability hampers clinical research to translate discoveries into effective drugs and interventions.

The digital era is at once deepening the challenge of computational reproducibility – as computational methods become more complex – while also increasing connectivity between researchers, allowing them to combat it. New information and communication technologies enable greater openness and expectations over openness tend to correspond with what is possible. The openness of publications, for example, is now evaluated across multiple dimensions: being immediately available at the time of publication to readers over the Internet at no cost, under generous reuse rights, with copies available in trusted repositories, and accessibility of appropriate metadata and even the full-text through an application program interface (API) to allow machine-readable discovery and text mining.

Reproducibility in research areas with a heavy computational component, such as bioinformatics, can be improved by making underlying research data available in a scientifically and technically interoperable format. This can enable re-assessment, and in turn validation, refinement, or refutation of findings. Funder and journal policies requiring researchers to release


96 Engineering National Academies of Sciences, Reproducibility and Replicability in Science (2019) at 1 (“We define reproducibility to mean computational reproducibility—obtaining consistent computational results using the same input data, computational steps, methods, and code, and conditions of analysis; and replicability to mean obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.”).

data aim to enhance scientific accountability: simply the fact that the potential results may be reviewed encourages researchers to proceed more carefully. Reproducibility is also enhanced by sharing of the underlying code. There is growing reliance on computing in almost all scientific domains, but this has not been matched by an equivalent increase in the availability of the source code, which renders computing programs “black boxes”.98 Funder and journal policies for sharing code lag behind their policies on sharing data, and this may be aggravated by the proprietary stances of university technology transfer offices.99 A proposal for enhancing the reproducibility of computational research is the sharing of a complete “research compendium”, including the publication, as well as the underlying data and code.100 There are certainly costs to increasing the reproducibility of data, but there are also a range of potential benefits, including reputational gains through exposure and citation, preservation, community building, good citizenship, higher quality outputs, broadening of access to data, and more credibility among peers and the public.101

Another central aim of greater openness is to promote efficient research and innovation. In the pharmaceutical sector particularly, there is increasing evidence of rising costs without corresponding increases in productivity.102 Governments, companies, and researchers also increasingly recognize the potential of sharing data to reduce waste and improve productivity. The basic idea is that there is a lot of unnecessary secrecy in both academic and industry science, which results in duplicative research, and delays where other researchers could have started building on findings. Duplicative research poses unnecessary costs to funders as well as

99 Ibid.
100 Stodden, supra note 91 at 114.
unnecessary risks to human participants. Better reproducibility would also improve efficiency, as researchers could build on existing findings with more confidence and fewer dead ends.

Openness of data also promises to support disruptive forms of science and innovation. It can enable Big Data analytics, text and data mining, machine learning, and artificial intelligence. These methods all require the aggregation and analysis of large quantities of high-quality, interoperable data. Innovations in scientific communication are proceeding in step with innovations in computational analysis. Traditional scientific publications are an increasingly antiquated means of disseminating research results. Reducing the results of a study to an article or a 2-D graph may over-simplify the results and also complicate novel strategies for building on those results.

Innovation promises to come not only from new technologies, but also from opening up science to new communities of researchers. There are general trends in innovation towards greater involvement of consumers in innovation processes. Developers of both digital and physical products are increasingly open to lead user innovations. Likewise, there is increasing enthusiasm to harness citizens – including companies, patients, and amateur scientists – in scientific projects. Indeed, a prominent definition of open data is that it “can be freely used, reused and redistributed by anyone for any purpose”. There is some risk, however, that efforts focused on democratizing research will focus more on inclusion than innovation. Under the right circumstances, openness can support quality, but this is not guaranteed.

The relationship between open science and inequality is starting to receive greater attention. On its face, open scientific knowledge should be a democratizing force. More knowledge in the public domain means less knowledge subject to proprietary exploitation. But there are also concerns that both openness, and the progress towards better health interventions,

104 Open Definition, “The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge”, online: <http://opendefinition.org/>.
will exacerbate inequality. Opening data may disproportionately benefit “only those who possess sufficient resources to allow them to usefully analyze them”.  

This is certainly a risk that openness affects fairness between research groups. Large research groups in Western countries will likely benefit disproportionately because they will have the resources to aggregate and analyze open data. Google’s value-added search service is perhaps the greatest example of a concentration of wealth derived from open data. There is a general concern that economic and technological progress generally is exacerbating inequality. Precision medicine tools and therapies, for example, are likely to be unaffordable even to the average patient in rich countries. To the extent open science is a proxy for such progress, it may attract similar skepticism. A counter-argument, however, is that regulation of science is an inappropriate locus for addressing broad societal inequalities, as restricting scientific freedom and openness is more likely to have a direct impact on innovation, while having a small effect on overall societal equality. Regardless, general concerns about inequality cannot be ignored in the development of open science governance.

Different aims of openness – reproducibility, efficiency, innovation and democratization – are likely achieved at different levels of interoperability. Lower levels of interoperability may allow for reproducibility, while precluding more innovative approaches. Sufficient interoperability to streamline innovation in the biomedical sector may still preclude disruptive innovations from outsiders.

The spectrum of openness in science extends from open access to publications, to the sharing of data and code, and finally to a comprehensive vision of “open science”. The US National Academies of Science, Engineering and Medicine (NASEM) proposes a vision of open science by design “as a set of principles and practices that fosters openness throughout the entire research life-cycle” (see Figure 1.1). Open access to the latest publications and other research outputs provokes new ideas and sparks new collaborations. Downstream openness should be planned during research design and enhanced by standardized data collection and curation. The

---


107 National Academies of Sciences & Medicine, supra note 90 at 21.
reproducibility of research is enhanced during a validation phase. Open dissemination and preservation of research outputs, including results and supporting information such as data and code, reignite the cycle in an efficient manner. OpenscienceASAP, an Austrian advocacy organization, has proposed a principle for open science that scientific knowledge – including software, data, publications, and educational resources – “should be openly shared as early as is practical in the discovery process.”

Such strong calls for openness in science highlight that the optimal level of interoperability for efficient and reliable science may be a radical one. Previously, policy debates over interoperability in science have focused on research results and data. This has proven limited for two reasons. First, “the outputs of research continue to grow in volume and complexity”, so there is a need to find ways to disseminate not only results but also supporting data, code, and other research outputs. Second – according to the general theory outlined above – technological, cultural, and institutional layers have to be attended to, in addition to the data layer.

The NASEM report on open science includes five recommendations. The first three cover appropriate rewards, training, and planning for achieving open science. The fourth recommendation focus on FAIR-conformant research archives. There is an important interplay between the interoperability of data and standards. Datasets and repositories are more interoperable if they use standard metadata vocabularies. Standards can also allow infrastructure and tools to support a greater range of research at lower costs. But there is a tension between standardization and creativity in research. For this reason, a European working group on open science recommends that “standards should be used in a rigorous but parsimonious way – not to


109 National Academies of Sciences & Medicine, supra note 90 at 9.

110 Ibid at 143–44.
restrict expressiveness of research data and keeping control of the ways data are processed and shared.”\textsuperscript{111}

**Figure 2.2 The Stages of Open Science**

The NASEM recommendations also touch on legal interoperability: “Researchers should seek to ensure that their research products are made available according to the FAIR principles and state with specificity any exceptions based on legal and ethical considerations.”\textsuperscript{113} This call for transparency and specificity appears geared towards ensuring ethical and legal issues are not used as cop-outs for data hoarding. It reflects a typical view of law as a fundamental constraint to the interoperability of health research. Law can place restrictions or conditions on the exchange of health data, in order to protect intellectual property, protect individual privacy, or to regulate the quality and ethical conduct of health research. In reality, there are few fundamental legal constraints on data exchange; rather, legal barriers tend to result from the selection of legal tools by the research community. The impact of law on health research interoperability does become more complicated for international collaborations. Laws in some jurisdictions restrict the flow of data across borders. Designing research governance to comply with multiple laws can be challenging where laws differ or even conflict.


\textsuperscript{112} National Academies of Sciences & Medicine, *supra* note 90 Figure S-1.

Legal restrictions, uncertainty, and international divergence are issues that must in the long run be addressed through concerted dialogue between researchers with legislatures. In the meantime, the challenge of enhancing interoperability of health research is largely left to the research community. I refer to the efforts of research communities and individual researchers to establish governance as private-ordering (as opposed to public-ordering established by legislatures). Private-ordering has significant potential to enhance interoperability. This is done through appropriate design and selection of licenses, contracts, safeguards, and consents.

There are, however, common problems where interoperability is left to private-ordering. These include voluntariness, mis-labelling, and negotiation costs. Voluntariness issues arise where researchers are unwilling to adopt the appropriate tools to ensure data are interoperable. Mis-labelling occurs where researchers fail to select the appropriate tools, leading to unintended downstream compatibility issues. Finally, legal regimes tend to require \textit{ex ante} permissions for the exchange and re-use of data. Researchers may need to seek these permissions from a variety of parties, especially where research involves aggregation and re-use of multiple data sources. Negotiating these permissions can involve significant costs and delays.

Chapters 3-6 of this thesis review the state of legal interoperability in health research across the domains of intellectual property, data privacy, and health research regulation. The analysis in each domain covers the following questions: in what ways does law restrict health research interoperability? When is variation between the laws of different countries a barrier to international collaboration? Is there uncertainty over whether or not the law even applies to research in the first place? What tools can individual researchers adopt to overcome legal barriers? Can society rely on the individual actions of researchers – on private-ordering – to achieve optimal levels of health research interoperability?

There are clear opportunities to improve the legal interoperability of health research. Research communities can engage in standardization efforts to improve the interoperability of licenses, consents, contracts, and access governance (see Chapter 7), just as they have done so to standardize the scientific and technical representations of data. Where bottom-up standardization fails, there may be a need for research law- or policy-makers to mandate interoperability for research, much like what has been done for electronic health records and consumer personal data portability. In this way, law is not only a barrier – it can also facilitate interoperability.
Despite these opportunities, a popular instinct is to develop technology to work around the law. A response to persistent technical and legal barriers to data exchange is to adopt computing and networking technologies that enable researchers to submit algorithms to secure databases and receive results without ever accessing the individual-level data. Algorithmic access can also enable analysis and aggregation of results across a distributed network of secure databases. Chapter 8 discusses the promise of such approaches, as well as practical and legal barriers to their implementation.
Chapter 3
Interoperability of Intellectual Property Systems

Key Points:
- Copyrights and *sui generis* database rights apply automatically to some health research outputs such as publications, data, software, and databases.
- Protected resources cannot be legally re-used, adapted, or shared without the owner’s permission. Standard licenses present a solution to remove restrictions and promote interoperability.
- It can be unclear whether or not intellectual property rights apply to research outputs, making it difficult for researchers to select appropriate and compatible licenses.
- International divergence of copyright regimes can affect international research collaborations, though harmonization efforts have reduced divergence.
- Technological developments such as algorithmic access to research data challenge fundamental intellectual property concepts and principles.

Interoperability is essential for collaborative health research and innovation. This section discusses the relationship between interoperability and intellectual property (IP). First, it briefly introduces IP rights relating to scientific knowledge. Scientific knowledge includes publications, underlying data, and software code used in research analyses. The focus of this Chapter is primarily on copyright and database rights. I also introduce standard licenses, which are designed to allow anyone to re-use, adapt, and re-distribute knowledge without the need for case-by-case negotiation with the rights holder. Second, it explores fragmentation of IP systems across jurisdictions, a potential problem for international research collaborations. Third, it discusses fundamental uncertainty over the application of copyrights and database rights in health research contexts. Uncertainty can lead to cautious re-distribution and re-use of health data, as well as incompatible licensing practices. With the increasing complexity of health research – and associated data generation, processing, sharing, and re-use – this uncertainty is bound to grow.

1 Introduction to IP

This section provides an introduction to copyright law as it relates to scientific knowledge and databases, as well as standard licenses as they relate to publications, software,
and data/databases.

1.1 Copyright

Among the various forms of IP, copyright has the most important consequences for the interoperability of health research, “because it applies automatically to most informational outputs of scientific research, including journal articles, less formal research reports, the organization of datasets, and software.”114 In Canada, copyright is outlined by the federal Copyright Act, R.S.C., 1985, c. C-42. Copyright applies to “works” of authorship that are original and fixed in some material form.115 Unlike patents which requires a step of registration, copyright protection applies automatically when the expression is rendered in fixed form. Originality relates to the expression of an idea, and the effort involved in expressing the idea, which must involve skill and judgment.116 A simple way to define an original work is that it was not copied from another work. Fixation means that the work “must be expressed to some extent at least in some material form, capable of identification and having a more or less permanent endurance.”117

Copyright only protects the expression of an idea, not the idea itself.118 This is an extension of a more general proposition that facts and ideas are not included in the subject matter of IP rights, i.e., there is no ownership in information per se.119 Copyright does not apply to ideas or facts themselves. This is a very important consideration in the context of (raw) scientific data, which tend to be factual in nature. The scientific process of processing, analyzing, and

114 National Academies of Sciences & Medicine, supra note 90 at 123.
115 Lesley Ellen Harris, Canadian Copyright Law, 4th ed ed (Hoboken, N.J: Wiley, 2014) at 24 (“Whether a creation is original in the copyright sense is always a factual question and one of degree that ultimately a court must decide.”).
116 Ibid at 25.
117 Wall v Horn Abbot Ltd, 2007 NSSC 197 at 484.
118 Harris, supra note 115 at 22.
synthesizing data into novel expressions, however, can blur the line between facts and original expression. Michael Carroll explains how the creation of publications, annotations, visualizations, figures, charts, and graphs “often involves the kinds of discretionary decisions about expression to which copyright applies”.120 There is a parallel between the uncertain copyright notion of originality, and the uncertain data privacy law notion of identifiability (discussed in Chapter 5). Both uncertainties make it hard to identify and negotiate underlying legal rights and obligations in health research data. A pragmatic response is to admit that

> understanding how copyright applies to the sharing of research data is more work than it is worth unless it is likely or plausible that the creator, owner, or repository in which data resides is likely to seek to limit copying, distribution, or other reuses of data.121

The rights provided by copyright law protect “the economic interests of the creators and owners of literary, dramatic, musical, and artistic works.”122 Copyright law gives the author or copyright owner exclusive rights to do certain things with a work or to allow others to do those things. Copyright includes the right to copy, publish, distribute, be credited for, translate, broadcast, publicly perform or display, and prepare adaptations of an original work or substantial parts of it.123 This bundle of rights can typically be transferred to others, modulated, or waived through a license, usually in exchange for royalties. The duration of these exclusive rights is time-limited (in Canada: the life of the author + 50 years).124 Copyright regimes also provide moral rights to authors, which “protect the integrity of a work and the identification of its creator.”125 Where copyright does attach to scientific knowledge, this clearly presents a barrier to interoperability. Sharing of scientific data typically involves making copies across multiple research laboratories, or even the public display of datasets on the internet (for example in open

---

121 Ibid at Copyright.
122 Harris, supra note 115 at 135.
123 Ibid at 136.
125 Harris, supra note 115 at 135.
Most jurisdictions also recognize that data compilations or databases can qualify for copyright protection if the selection and arrangement of objects in the database meets the originality criterion. In such cases, however, copyright extends only to the “components of the work that are original to the author”, not the objects themselves (which may be unprotected facts or other copyrighted works).\(^{126}\) This adds further uncertainty to the copyright question in health research. Significant effort and perhaps originality are often involved in developing scientific datasets. Indeed, “compilations of datasets—used in meta-analysis, for example—might receive a separate copyright if the selection and arrangement of these involve sufficient discretionary choice.”\(^{127}\) But, “any protection granted to compilations would, in practice, only safeguard against a very limited scope of actions. Copyright law, again, does not extend to the facts contained in the compilation and is limited to the facts’ particular selection or arrangement.”\(^{128}\)

A further challenge arising in health research contexts is determining who owns the copyright. In the United States, the work-made-for-hire rule automatically grants copyright ownership to employers where protected works are created in the course of employment. There is some uncertainty about how this work-made-for-hire rule applies to outputs of research by full-time faculty. Some argue that research outputs are not within the scope of employment or that exceptions implicitly apply in academic contexts, but “[o]n its face, current law does not state any exceptions to the rule.”\(^{129}\) Authors outside the US generally receive copyrights by default where the work is developed within an employment relationship.\(^{130}\) The default both in the US and elsewhere can be modified under an institutional IP policy or employment

\(^{126}\) US:  at 363-64.

\(^{127}\) Carroll, supra note 120 at Box 1.

\(^{128}\) Determann, supra note 119 at 22 (This means that a subsequent compiler will be free to use the facts contained in the prior compilation, as long as the competing work does not feature the same selection and arrangement. To be successful with copyright claims, a plaintiff thus has to prove that the defendant copied more than the merely extracted factual information.”).

\(^{129}\) Carroll, supra note 120 at 9.

\(^{130}\) Ibid.
agreement. Fortunately, many institutions have adopted IP policies that address this uncertainty, often recognizing researchers as the authors (and therefore copyright owners) of journal articles they write, datasets they produce or assemble, and software they create.”

In other countries, employers may still impose an automatic transfer of copyright from the researcher to the employer under an employment contract.

While copyright law protects ownership interests in creative works, it also protects the rights of users of creative works through limitations and exceptions. These can be based on the “status of the user, the type of use, its extent, the type of protected works, or other factors.” According to the Supreme Court of Canada, the Copyright Act is usually presented as a balance between promoting the public interest in the encouragement and dissemination of works of the arts and intellect and obtaining a just reward for the creator… The proper balance among these and other public policy objectives lies not only in recognizing the creator’s rights but in giving due weight to their limited nature.

The Supreme Court urges Canadian courts to “strive to maintain an appropriate balance between these two goals.” American copyright law is predominantly based on utilitarian foundations. Lothar Determann explains that the “immediate effect of [American] copyright law is to secure a fair return for an ‘author’s’ creative labor. But the ultimate aim is, by this incentive, to stimulate artistic creativity for the general public good.”

The Canadian Copyright Act provides exceptions for certain uses that do not infringe copyright, including use for the purpose of research or private study. According to the

---

131 National Academies of Sciences & Medicine, Returning Individual Research Results to Participants: Guidance for a New Research Paradigm (National Academies Press, 2018) at 123.

132 Carroll, supra note 120 at 9.

133 Doldirina et al, supra note 10 at 15.


135 Ibid at 10.

136 Determann, supra note 119 at 18.

137 Copyright Act, R.S.C. 1985, c. C-42, supra note 124 s. 29.
landmark Supreme Court of Canada judgment on users rights, *CCH Canadian Ltd. v. Law Society of Upper Canada*, 2004 SCC 13, the enumerated purposes under fair dealing should be accorded “large and liberal interpretation” so that users’ rights are not unduly constrained. In that case, the court found that the research exception can, but does not necessarily, cover research for commercial purposes (e.g., case research by lawyers).  

Moreover, fair dealing after CCH is no longer seen as a defense to infringement (with the burden of proof on the defendant), but more of a broad exception (with a potential shift of the burden of proof to the plaintiff to show the purpose was not fair dealing).  

In the US, the *Copyright Code* provides that the “fair use of a copyrighted work for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research” does not infringe copyright. One distinction here is that the US approach is an open ended list, though this distinction is less evident since the Supreme Court of Canada ruling took a broad and expansive interpretation of the statutory exceptions in *CCH*. The US Copyright Code includes four factors to be considered. In the context of copying website source code easily accessible through browsers, Jonathan Zittrain argues that “fair use is determined by a fuzzy four-factor test that in practice rests in part on habit and custom, on people’s expectations.” But distinguishing what is a fair use in research contexts is difficult and fraught with uncertainty. One wry commentary is that “[f]air use in

138 Giuseppina D’Agostino, “Healing Fair Dealing-A Comparative Copyright Analysis of Canada’s Fair Dealing to UK Fair Dealing and US Fair Use” (2008) 53 McGill LJ 309 at 12 (“[W]hereas in previous cases and as featured in almost every textbook, fair dealing was conceived of as a defense, CCH construes it more as a “right” and an “integral” part of copyright law.”).  

139 Ibid at 13–14.  


142 Zittrain, supra note 23 at 136.  

143 Stodden, supra note 93 at 115–116.
America simply means the right to hire a lawyer”.144 A new “digital” copyright treaty was established by the World Intellectual Property Organization (WIPO) in 1996, called the WIPO Copyright Treaty, which aimed to deal with the effect of the digital age and new technologies on copyright, including the protection of databases and computer programs, as well as provisions guarding against the removal of rights management information and circumvention of digital locks.145

International protection of IP, especially over the Internet, is a complicated topic. It is difficult to determine the applicable law: is it where the work was uploaded? downloaded? transmitted?146 There is legal uncertainty over whether or not ownership rights continue to apply as copyrighted works travel across borders, given that copyright protections are national. Fortunately for rights holders, automatic protection across borders is the subject of numerous international conventions (e.g., World Intellectual Property Organization, Berne Convention) and treaties (e.g., World Trade Organization, Trade-related Intellectual Property Agreement). The domestic law where the work is used determines the level of protection.147 Many health research collaborations have global scope, and where copyrights are involved, jurisdictional uncertainty is sure to follow.

1.2 Sui Generis Database Rights

In Europe, Mexico, and South Korea, legislatures have provided protections for databases that do not involve an original selection and arrangement of elements, but that required substantial investment. The European Union offers legal protection for databases under the Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996.148 The

145 Harris, supra note 115 at 64.
146 Ibid at 68.
147 Ibid.
Directive seeks to harmonize national laws, encourage investment in database production, and maintain appropriate freedoms for users.\textsuperscript{149} The Directive defines a protected database as “a collection of material (the Contents) arranged in a systematic or methodical way”.\textsuperscript{150} To qualify a database must involve “qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents.”\textsuperscript{151} These rights prevent substantial parts of a protected database from being extracted, re-used, or distributed. They subsist for 15 years, but this term can be continually renewed following additional investment. Substantial investment is left undefined, and it is unclear what kinds of alterations would qualify for renewed protection. There are concerns, however, that such rights of “virtually unlimited duration”, which could reduce competition in value-added services, and negatively impact scientific research.\textsuperscript{152}

Because scientific research often relies on using large amounts of data found in databases, database rights could significantly increase the costs of such research.\textsuperscript{153} Another concern with \textit{sui generis} database rights is that the scope of exceptions or user rights are limited in comparison to copyright. Some exceptions under the Directive were optional for Member States, and the database owner could override many others in contracts.\textsuperscript{154} Contractual overrides weaken user rights, but given the flexibility of contract, they also lead to issues of interoperability as contracts and licenses proliferate.\textsuperscript{155} These rights go beyond international treaties, and so are only protected within the national jurisdictions that provide for their

---


\textsuperscript{151} \textit{Ibid} art 7.


\textsuperscript{153} \textit{Ibid} at 8.

\textsuperscript{154} European Commission, \textit{supra} note 149 s 2.2.2 (potentially reinforced by the Court of Justice of the European Union in Case C-30/14 Ryanair).

\textsuperscript{155} \textit{Ibid} s 3.4.3.
protection. In the United States, databases involving substantial investment but not original arrangement and selection receive some protection under misappropriation laws that “safeguard business integrity and fair competition”. Similar unfair competition rules also apply in European countries on top of the sui generis database right, which arguably should have been pre-empted by the Directive. The arguments for such protections are to address a market failure leading to under-investment in database creation because of a risk of free-riding by competitors.

1.3 Standard Licenses

Simple, standard licenses have been developed to facilitate the sharing of software, creative works, and data/databases. The open source software movement promotes openness in two ways. First, it insists on the sharing of source code. Source code is “the human readable form of a programming language [that] contains the complete set of instructions for how a computer processes input data”, and is distinct from the binary (executable) version of a program. Access to the source code allows others to inspect how the code works. Second, open source software licenses are applied to ensure that IP rights in the code do not prevent it from being freely used, adapted, and redistributed. The Open Source Initiative certifies licenses compliant with the Open Source Definition. The definition requires a license that allows for free redistribution (including sale) of the software or derivatives, along with source code, to anyone, in any field of endeavor. The definition was developed to constrain a steady proliferation of licenses, which led to confusion and incompatibility. Some new licenses that restrict what companies can do to commercialize software are gaining popularity, but these licenses

156 Carroll, supra note 120 at Introduction.
157 Determann, supra note 119 at 21.
158 European Commission, supra note 149 s 4.2.3.
159 World Intellectual Property Organization, supra note 152 at 2.
160 Morin et al, supra note 98 at 159.
controversially breach the “open source software tradition of allowing users to do whatever they want with the code.”\(^{162}\)

Standard licenses have been developed to enable the sharing and re-use of creative, educational, and scientific works by a global organization called the Creative Commons, without prior approval from the owner. The Creative Commons licenses are a response to the issue that copyright applies automatically in a protected work. Authors who want to share their works openly have to take explicit steps to do so. A Creative Commons 0 (CC-0) waiver irrevocably places a protected work in the public domain. This waiver can also be used to indicate that scientific knowledge (e.g., factual research data) has no underlying IP rights attached. In countries that do not allow authors to give up copyright entirely, CC-0 also functions as a license posing no constraints on users.\(^{163}\) The Creative Commons also releases a suite of other standard licenses that impose some restrictions or conditions on sharing and re-use. The Creative Commons Attribution 4.0 International license (CC-BY 4.0), for example, imposes a single condition: attribution of the author in a specified manner.

Standard licenses have also been specifically developed for data and databases. Factual data are not protected by copyright, but some processed data and databases can receive IP protection. Given this uncertainty, data licenses often have a dual nature: they function as IP licenses (in the case underlying IP applies) and as contracts (in the case there is no underlying IP). Data licenses also typically address both databases and their contents, as both may be protected separately. The Open Data Commons Public Domain Dedication and License (PDDL) is a public domain dedication similar to the CC-0. It functions both as a waiver of moral rights, database rights, and copyrights (should they exist in the data), and as a license (should there be any IP rights that owners are legally unable to waive). The PDDL was developed by the Open Data Commons to be conformant with the Open Definition, which states that open data is data


\(^{163}\) Carroll, *supra* note 120 at 9.
that can be “freely used, modified, and shared by anyone for any purpose.”\textsuperscript{164} The Linux Foundation has developed two Community Data License Agreements (CDLA), a CDLA-Permissive, and a CDLA-Sharing.\textsuperscript{165} Both allow users to modify and adapt the dataset, but require they provide attribution to data contributors when re-sharing the data or enhanced data. The Sharing version “puts terms in place to ensure that downstream recipients can use and modify [the] data, and are also required to share their changes to the data.”\textsuperscript{166} Neither license imposes obligations on results obtained from “computational use” of the data. In the next Chapter, I discuss the challenges of distinguishing derivative datasets from results, especially in the artificial intelligence context. These licenses apply to both the data and the database (where no underlying IP exists, the protection is contractual). There are also standard licenses that are specifically designed to apply to databases only, and not the data elements they contain, to address sharing and re-use in light of copyright (original databases), and \textit{sui generis} rights (unoriginal databases).\textsuperscript{167}

In summary, there are roughly three categories of standard licenses, regardless of subject matter: permissive, restrictive, and viral. Permissive licenses are those allowing for re-use, modification, and re-distribution of information without restriction (except perhaps an attribution requirement). Examples for media are CC-BY and CC-0, an example for software is the MIT License, and examples for data are the PDDL and the CDLA-Permissive. Some standard licenses impose common conditions (such as attribution) or restrictions (such as no commercial use or no derivatives). Viral (i.e., share-alike or copyleft) licenses are standard licenses that require any derivative works to be shared under the same license. An example for media is the Creative

\begin{footnotesize}
\begin{enumerate}
\item[164] Open Definition, \textit{supra} note 104.
\item[165] Linux Foundation, “Community Data License Agreement - FAQ”, online: \textit{CDLA} <https://cdla.io/faq/>.
\item[166] Linux Foundation, “Community Data License Agreement – Sharing, Version 1.0”, online: \textit{CDLA} <https://cdla.io/sharing-1-0/> (“1.2 ‘Computational Use’ means Your analysis [through the use of computational devices or otherwise] or other interpretation of Data. By way of example and not limitation, ‘Computational Use’ includes the application of any computational analytical technique, the purpose of which is the analysis of any Data in digital form to generate information about Data such as patterns, trends, correlations, inferences, insights and attributes.”).
\item[167] “ODC Open Database License (ODbL) Summary”, (12 September 2009), online: \textit{Open Data Commons} <https://opendatacommons.org/licenses/odbl/summary/>.  
\end{enumerate}
\end{footnotesize}
Commons Attribution-Sharealike 4.0 International license (CC-BY-SA 4.0), used by Wikipedia, an example for software is the GNU General Public License v3.0 (GNU GPL 3.0), and an example for data is the CDLA-Sharing License. Even an attribution requirement like the one contained in a “permissive” CC-BY license is viral, meaning that if the media is used in another’s work, the original attribution is retained. In other words, attribution propagates downstream.\(^{168}\) Viral licenses are alternatively characterized as “open” – because the subject matter and any adaptations always remain open – and as “restrictive” – because they restrict the user’s freedom to license any adaptation in any form.

Concepts of standard licensing and virality have potential application beyond copyright. In health research, some sample and data repositories impose contractual obligations on users to deposit enriched or derived data back in the repository (to be shared under the same conditions). John Wilbanks and Stephen Friend have proposed that instead of designing research consent forms that simply “permit” data sharing, these consent forms could mandate transparency and sharing.\(^{169}\) Many research participants could prefer their data are accessible to many researchers rather than treated as property. Reserving all rights in a copyrighted work is the opposite end of the spectrum. There are a number of intermediate licenses building on CC-BY that are still considered open and require attribution, but that place restrictions on some among the bundle of copyrights, e.g., licenses that reserve the right to distribute derivative works (e.g., CC-BY-ND), that only allow the use of a work for non-commercial purposes (e.g., CC-BY-NC), or a combination of these restrictions. Finally, access to research outputs is often governed by contractual agreements – variously called data use agreements, licenses, or terms of use – even where there are no underlying IP rights. The conflation of IP and contractual licenses for data is discussed in the next Chapter.

\(^{168}\) Stodden, supra note 101 at 39.

2 Towards Interoperable IP Systems

This section reviews approaches that both states and private actors can adopt to improve interoperability of IP systems.

2.1 Harmonization

There has been significant convergence of copyright law across jurisdictions, supported by harmonization through international conventions and treaties. One long standing point of difference was the originality criterion. In common law countries there was historically a divide between creativity and sweat-of-the-brow approaches.\textsuperscript{170} Civil law jurisdictions traditionally used to “search for the mark of the author’s personality in the work”.\textsuperscript{171} Legal regimes have converged towards an originality standard for determining what deserves copyright protection.\textsuperscript{172} It may remain somewhat more difficult to argue that copyright applies to health research data in countries where the originality standard emphasizes originality or the mark of the author, given that research data generation does not typically involve creative choices, except perhaps during later stages of processing, annotation, and display. Another point of traditional divergence was over whether or not authors had moral rights in a work, and whether or not they are able to waive such rights. Disputes over researchers’ moral rights in data seem unlikely. They are more likely to view their data as an unveiling of an objective truth than an extension of their creative self, more likely to view misuse as a breach of scientific rather than personal or creative integrity. Whether copying an unoriginal database infringes IP rights will differ depending on the jurisdiction where the copying takes place.

There is less convergence across jurisdictions with regards to limitations and exceptions in copyright. The US NASEM explain that “[a]ll countries have a list of uses permitted by law, but these lists vary widely, and the identified uses are often specific and narrow. Countries also


\textsuperscript{171} Ibid.

\textsuperscript{172} Ibid.
create their own exceptions to determine whether a use is permitted”. This is partly because international conventions and treaties tend to prioritize establishing minimum standards for protecting the rights of owners, while allowing individual countries to establish greater protections where they desire. The Berne Convention, for example, aims to establish minimum standards in what types of works are protected, the rights protected in those works, and the duration of protection. Trade agreements such as the World Trade Organization, Trade-Related Intellectual Property Agreement (TRIPS) also include minimum standards to facilitate trade in protected works and to combat against international piracy. There has been some concern over a steady ratcheting up or expansion of IP protections in bilateral trade agreements in which countries add tailored clauses usually in the direction of greater protection for owners. As a result, it may be impractical for researchers to rely on limitations and exceptions to re-use copyrighted research outputs, especially internationally.

2.2 Standardization

On the national level, there is some movement towards sector-specific standardization efforts to clarify copyright issues such as the scope of fair dealing, involving creators, users, and rights holders which may then be relied on by courts. Such soft law approaches are somewhat analogous to the codes of conduct approach discussed in Chapter 5 in the context of data protection law, although those codes are also subject to legislative approval. An open question is whether or not such approaches can scale internationally. Ensuring such guidelines complied with multiple, divergent laws, and ensuring appropriate consultation with relevant stakeholders would be challenging. If successful, however, such an approach would ease certainty where protected works are circulated internationally.

173 National Academies of Sciences & Medicine, supra note 90 at 123.
174 Harris, supra note 115 at 60–61.
175 Ibid at 65.
177 D’Agostino, supra note 138 at 53.
2.3 Other Strategies

John Palfrey and Urs Gasser have mapped out a number of generic, top-down national strategies for enhancing interoperability in different areas of human activity, including mandating standards, disclosure of interoperability information (a form of compulsory licensing), transparency rules (i.e., labelling requirements), government procurement, and competition law. In health research, government funding agencies could consider such approaches. Mandating standards for licensing may be one potential route. This approach was recently proposed to the US National Institutes of Health by members of research groups involved in the collection and integration of health research data. The letter advocated for the development of common, mandatory licensing and data reuse plans for publicly funded health research data, and cited the problem that “[t]he current diversity of data use agreements and licenses significantly hampers the ability to reuse and redistribute data in various informatics contexts.”

Transparency rules are another potential route, where researchers are free to select a license, but they must explicitly state what license applies. Given that standard licenses are internationally recognized, it is possible that international policy frameworks could be developed on the topic, and then implemented as national standards. In later sections, I discuss issues with leaving licensing practices up to individual researchers and institutions, and what potential strategies are available to improve the interoperability of licensing even in the absence of top down approaches.

3 IPRs and Research Data: Uncertain Application and Poor Fit?

A rough rule is that copyright law does not apply to research data, which are largely factual, and are unlikely to meet the test for creativity. However, the line between factual information and a creative work is blurred through the scientific process, which proceeds from

---

178 Gasser, supra note 8.

data collection to processing, analysis, and synthesis, and finally to the communication of results through papers, charts, and other increasingly sophisticated visualizations. This section explores uncertainties over the application of copyright and database rights to health research outputs. The example of the BRCAExchange, a publicly available genetic variant database, is used as an illustration. I argue that clarity over the researcher’s intention (to waive or impose restrictions) is more important than clarity over whether or not there are underlying IP rights in research outputs.

3.1 Case Study: Licensing and the BRCAExchange Public Genetic Variant Database

A case study from “open” genetic databases demonstrates how legal interoperability problems can arise even within a network of databases committed to the public distribution of genetic data. The BRCAExchange is a database that aggregates and publicly shares data about genetic variants observed within the BRCA1/2 genes, and clinical evidence linking those variants to breast and ovarian cancer. These data are largely scraped from other databases. “Genetic sequencing labs, academic centers, and companies all work independently to classify BRCA1 and BRCA2 variants as pathogenic or benign using clinical data and other evidence”, but exchange of data between them is often limited.\(^{180}\) Where two clinical laboratories do not have access to the same knowledge about a particular variant, they may give patients with the same variant different test results. This likely means one of the patients is receiving sub-standard care. Legally, the goal of the BRCAExchange is to make high quality BRCA1/2 data publicly accessible with as few legal conditions on re-use, modification, or re-distribution as possible.\(^{181}\) The BRCAExchange does not want to discriminate between researchers, clinicians, commercial tool developers, or patients. All of these different users may find innovative new ways to make use of scientific knowledge. The BRCAExchange is reluctant, however, to apply a permissive, standard license, because many of its source databases, despite being publicly available, are

\(^{180}\) “BRCA Exchange”, online: <https://brcaexchange.org/about/history>.

subject to restrictive or unclear licenses.

### 3.2 Ownership of Research Data

The idea of a single owner of research data presented by copyright does not easily track the complexity of the data trajectory in data-intensive research. Numerous parties or even automated sensors may be involved in capturing data, while numerous other parties may be involved in cleaning and curating data, and finally these data may be shared with a wide community of researchers for analysis.\(^{182}\) It is difficult to track the provenance of data and the extent of upstream contributions, let alone to determine what ownership interests do or should exist in the data. As I discuss in more detail in Chapter 6 on health research regulation, questions of control and ownership are becoming even more complex as research participants and communities receive greater recognition as partners in the governance of research data. Copyright may also be at odds with scientific norms of reproducibility and openness, because by default it prevents the legal copying and verification of results, as well as the creation of derivative works from research data and code.\(^{183}\)

Is there copyright in the genetic variant interpretations shared through the BRCAExchange? Observations linking genetic variation with disease risk are facts, and it is highly unlikely they would be protected by copyright. Is there copyright in the databases of genetic variant interpretations that are sources for the BRCAExchange? It is also unlikely that databases of genetic variant data would qualify as an original selection or arrangement of facts. It is more likely that even an unoriginal genetic database would receive IP protection in jurisdictions with *sui generis* database rights, if substantial investment was involved. However, investment in the curation of individual elements within the database are not taken into account, such as the generation of sequence data and subsequent interpretation. There must be substantial investment in the selection and arrangement of the elements.

\(^{182}\) Stodden, *supra* note 91 at 124.

\(^{183}\) Stodden, *supra* note 93 at 115.
Moreover, if IP rights do apply to the source database, the BRCAExchange is only extracting narrow portions of a database (variants in one gene among thousands). For academic researchers at least, re-use may be covered under the fair use or fair dealing exceptions. One challenging question the BRCAExchange faces in relying on fair use is how it would describe the “purposes” for which it is re-distributing data. The BRCAExchange publicly displays genetic variants primarily for use in clinical interpretation. Would this qualify as only a “research” or “educational” purpose? Arguably the BRCAExchange is enabling both non-commercial and commercial uses by third parties. Would this affect the outcomes of the fair dealing and fair use tests?

The BRCAExchange aggregates publicly available data from numerous countries, and in turn makes the data publicly available everywhere in the world. The first challenge this raises is with interpreting the underlying rights of its source databases – database protections differ across countries. Another set of issues is liability for infringement. Would the BRCAExchange be liable for publicly displaying parts of other databases? Would users of the BRCAExchange be liable if they created a copy of the BRCAExchange database? Moreover, because the BRCAExchange aggregates data from international sources, and also makes data available internationally, it is forced to navigate multiple IP regimes and divergences between them.

In short, it is unlikely that public genetic variant databases are protected by copyright. Presumably, this should simplify licensing as well as re-use, adaptation, and re-distribution. As I discuss in the next Chapter, however, licensing practices continue to be a major problem even for publicly available scientific databases.

### 3.2.1 Fragmentation of Data and Legal Rights

Copyright or *sui generis* database rights may be afforded to research databases, publications, or processed data, depending on the jurisdiction and context. Uncertainty over the existence of underlying IP rights is magnified by the process of data fragmentation, which is a common occurrence in data science environments. Fragmentation occurs in data science where a derivative dataset is created by selecting sub-components of other datasets. The most obvious example of how fragmentation affects ownership is where databases are used for research. As Catherine Doldirina observes,
Even when the selection and arrangement of facts are protected by copyright, the typical software application in extracting data from one or more databases will seldom copy creative elements of the original datasets since the software application will extract its own algorithmically defined selection fit for a particular purpose.\footnote{Doldrina et al, \textit{supra} note 10 at 19.}

Fragmentation may make it hard to assess if substantial copying has occurred under fair use exceptions.

\section*{3.3 The Changing Meaning of Copyright in an Era of Mass Digitization}

Open science communities are concerned with applied challenges of reducing legal restrictions on research data. Less attention is paid to conceptual challenges, such as the meaning or purpose of copyright as it applies to research data shared in digital environments. Maurizio Borghi and Stavroula Karapapa’s analysis of mass digitization of copyright works (e.g., books) for digital libraries or search engines (e.g., Google Books) argue that mass digitization “tells us something new about copyright and its limits”\footnote{Maurizio Borghi, \textit{Copyright and Mass Digitization : a Cross-jurisdictional Perspective} (Oxford : Oxford University Press, 2013) at 18.}:

\begin{quote}
the technological shift occurring with mass digitization impacts on the common copyright perception of the concept of use, on the meaning and function of the work as protected subject matter, on the grounds for entitlement to rights, and on the scope of the exclusivity of such rights.\footnote{\textit{Ibid}.}
\end{quote}

We can see similar changes occurring with how health researchers interact with scientific knowledge.

Borghi and Karapapa note a shift in an era of mass digitization in the very concept of works “as expressions addressed by the author to the public” to that of data that can be computed
upon. Open science rhetoric does sometimes refer to research datasets as a researcher output or creation (usually with the intention of attracting greater recognition for data curation as a valuable scholarly contribution). Visualization tools make it easier for human audiences to interact with databases in an intuitive manner. But there seems to be more excitement around the possibilities of automated data analysis techniques that do not clearly qualify as author-audience communication. These techniques include text and data mining, training machine learning algorithms, and providing knowledgebases for clinical decision support.

Borghi and Karapapa pose a provocative question: “Is there, and should there be, a distinction between the use of works for purposes of experiencing their content and uses on works for purposes of indexing and search, computational analysis, and data mining?” This is also becoming an important concern in health research. Extensive efforts have been made, for example, to shift research publishing towards open access standards to ensure researchers can access research manuscripts. But even where publications are openly available to researchers to read, copyright may limit their ability to carry novel analyses. Text and data-mining approaches often require the temporary copying of substantial amounts of copyrighted material. In short, “[s]ince technology enables uses of works that may be unrelated to their internal purpose and function as embodiments of the author–user coalescence, we need to reconsider the role of copyright as a regulatory system and a driver of innovation.”

Mass digitization also has practical consequences for negotiation and licensing. The copying or computation of works in “bulk” as opposed to units has placed enormous stress on the traditional system of copyright as a system of ex ante permissions: “[c]learing rights for individual uses is not feasible or may be prohibitively costly.” The case of orphan works is illuminating for health research. Orphan works are “still in copyright for which no rightsholder

187 Ibid at 15–16.
188 Ibid at 16.
189 Ibid at 18.
190 Ibid at 2.
has been identified or located after a diligent search.” Health researchers often encounter orphan datasets where rights holders cannot be identified, cannot be found, or are unresponsive to access or licensing requests. This is partly because the systems in place for preserving and labelling datasets are not yet as robust as they are for publications. There is also a potent analogy between the orphan works issue in copyright and the re-consent problem where researchers want to re-use previously collected legacy data for studies not covered in the original consent. Data privacy law and health research regulation may require researchers to seek permission from participants for new uses or for sharing of their data. This presents a practical problem where those participants cannot be recontacted. There are sometimes case-by-case exceptions to consent requirements where consent is impracticable, but even these negotiations with research ethics committees may be prohibitive at scale.

The BRCAExchange demonstrates similar difficulties of mass digitization presented by the Google books situation (albeit on much smaller scale). As the BRCAExchange connects to more and more publicly available or private laboratory databases, it becomes increasingly impractical for users of the BRCAExchange to go back to all of the source databases to negotiate re-use with each of them individually. Moreover, some of the source databases of the BRCAExchange may themselves be data aggregators, with their own source databases. As for the challenge of fragmentation, the BRCAExchange demonstrates how genetic data has the potential to be fragmented and repackaged in myriad (no pun intended) ways. The BRCAExchange draws data specific to two genes from source databases containing genome-wide information. This raises uncertainty as to whether or not the license restrictions applying to genome-wide databases would continue to apply to fragments distributed by the BRCAExchange.

There is significant uncertainty over the application of IP to research outputs. This can discourage researchers from sharing and complicate the selection of clear and explicit research licenses. Naturally, researchers want to determine what rights they have in research outputs

191 Ibid.
192 Ethical Conduct for Research Involving Humans, Canada, Tri-Council Policy Statement (TCPS), 2014 [TCPS2] art 5.5A.
before they decide what rights they want to relinquish. This is the wrong way to go about things. Researchers should instead start by clarifying their intention to share, and any restrictions or conditions they feel must apply. Then a suitable combination of licenses and contracts can be selected, without having to resolve the question of underlying rights.
Chapter 4
Research Data License Interoperability

Key Points:
- Attaching interoperable licenses to **publicly available research outputs** – including data, databases, software, and publications – enables their combination, adaptation, and re-use.
- Licenses are more interoperable where they are standard, transparent, non-restrictive, and do not require case-by-case negotiation with a rights holder.
- Selection of licenses is complicated by unclear application of standard licensing terms to research activities; uncertainty over the existence of underlying IP rights in research data; conflation of IP and contractual licenses; and uncertainty as to whether attribution should be addressed through legal licenses or community norms.
- Across systems, licensing can end up being more restrictive than any single party intended due to license proliferation, which results in unpredicted incompatibilities, and the accumulation of licensing restrictions in derivative products.
- These are illustrated by the efforts of the BRCAExchange database to develop a licensing policy. The BRCAExchange is a publicly available database aggregating scientific knowledge about variants in the BRCA1 and 2 genes and their links to cancer.
- I recommend the following strategies to improve research licensing practices:
  - Researchers: waive or minimize restrictions and conditions, and adopt hybrid licenses for data that act as both IP permissions and contracts.
  - Repositories: promote data submitters to apply standard, open-conformant licenses.
  - Research institutions and technology transfer offices: provide researchers with dedicated support with license selection.
  - Data aggregators: adopt an open data guarantee affirming permissions have been obtained from source databases for re-use, modification, and sharing.

This chapter explores the extent to which researchers can enhance the interoperability of licenses for publicly available research outputs. Licensing practices are linked to the overall
interoperability, or openness, of scientific knowledge. Indeed, the US National Academies of Science, Engineering and Medicine’s *Open Science by Design* report recommends researchers should not only “report all results and supporting information (data, code, articles, etc.)” but they should “use appropriate licenses for sharing research outputs”. Research outputs include data, databases, software, and publications. This Chapter will primarily focus on publicly available health research data and databases. The public availability of research outputs brings IP licensing interoperability into relief, as there are presumably no technical barriers to access or data privacy issues associated with data. Where data are seemingly publicly available but are poorly licensed, some users will find they can “look” i.e., access the data, but they can’t “touch” i.e., re-use, combine, adapt, or re-distribute the data in creative ways.

Three interoperability problems can arise as a result of private-ordering, even where researchers intend to make data widely available. First, negotiation costs are particularly problematic for innovative analyses dependent on aggregation or combination of multiple sources of scientific knowledge. Text and data mining of scientific publications, for example, requires permissions to analyze thousands of publications, owned by numerous publishers, and available under a diversity of proprietary as well as permissive, standard licenses. Second, the selection of licenses remains largely voluntary, even where researchers are required by funders or journals to share results and underlying data. The result is that licenses are often unnecessarily restrictive, inhibiting re-use even where data can be easily accessed. A third problem is that researchers and repositories often apply inappropriate licenses to scientific knowledge. Inappropriate licensing can lead to unintended restrictions on downstream uses and incompatibilities between licenses when researchers attempt to aggregate multiple resources. These three problems – negotiation, voluntariness, and mis-labelling – are mutually reinforcing.

This Chapter proceeds in three sections. First, it reviews definitions and criteria for interoperable research data licenses. Second, it describes the negative consequences of current

---

licensing practices on the interoperability of publicly available health research databases\textsuperscript{195}, using the case study of the BRCAExchange as an illustration. This database aggregates publicly available data about the clinical significance of different mutations within the BRCA 1 and 2 genes, which are linked to breast and ovarian cancer. The database’s effort to adopt a single, permissive license for users has been confounded by the restrictive and incompatible licensing practices of its source databases. Third, it reviews persistent issues complicating licensing in health research contexts. Licensing problems are ironically most evident where health research data are made publicly available. I conclude by proposing several strategies to improve licensing practices in health research contexts.

1 Research License Interoperability Criteria

The Research Data Alliance defines legal interoperability as “the ability to combine data from two or more sources without conflicts among restrictions imposed by data providers … and without having to seek authorization from the data providers on a case-by-case basis.”\textsuperscript{196} The Alliance focuses on the availability of publicly funded research data. More specifically, legal interoperability among multiple datasets from different sources occurs when:

- the legal use conditions are clearly and readily determinable for each of the datasets typically through automated means;
- the legal use conditions imposed on each dataset allow creation and use of combined or derivative products; and
- users may legally access and use each dataset without seeking authorization from data creators on a case-by-case basis assuming that the accumulated conditions of use for each and all of the datasets are met.\textsuperscript{197}

In summary, licenses should be transparent, compatible, and applicable without negotiation. The Research Data Alliance’s criteria for license interoperability tends towards criteria for openness,

\textsuperscript{195} Seth Carbon et al, “An Analysis and Metric of Reusable Data Licensing Practices for Biomedical Resources” (2019) 14:3 PLOS ONE e0213090 (systemic licensing issues are outlined in this recent survey of the licensing practices of public health research databases: many “do not provide [clear] legal permissions for reuse and redistribution”).

\textsuperscript{196} Doldirina et al, supra note 10 at 8.

\textsuperscript{197} Ibid at 8.
because they view even limited licensing restrictions or conditions as barriers to data re-use, integration, and redistribution. Regardless, defining license interoperability criteria brings into focus barriers that only emerge as data exchange scales: uncertainty, policy conflict, and negotiation costs.

Even where researchers seek to share research outputs openly, it can be difficult to select an appropriate license. Seth Carbon et al. eloquently stated the challenge in a recent review of biomedical database licenses:

A lack of licensing rigor and standardization forces data users to manually seek, often repeatedly and from multiple data providers, essential reuse and redistribution permissions. Issues include missing licenses, nonstandard licenses, and license provisions that are restrictive or incompatible. The legal interpretation of, and compliance with, database license and reuse agreements has become a significant burden and expense for many fields in the scientific community, where a complex and lengthy set of legal negotiations may be required for a data integration project to legally and freely redistribute all of its relevant data.\(^\text{198}\)

In short, even where scientific resources are seemingly publicly available, restrictive, ambiguous or missing licenses can limit the ability of data scientists to combine, adapt, re-use, and share them.\(^\text{199}\) Poor licensing practices lead to increased negotiation costs.

The survey referred to above was conducted by the (Re)useable Data Project, which aims to promote the use and re-use of biological and biomedical datasets.\(^\text{200}\) This initiative builds on the Research Data Alliance criteria as well as on the principles that data be FAIR (findable, accessible, interoperable and re-useable).\(^\text{201}\) Indeed, some of the organizations behind the Reusable Data Project have proposed that publicly funded research data should also be FAIR-TLC: traceable, licensed, and connected.\(^\text{202}\) The (Re)useable Data Project rates licensing

\(^{198}\) Carbon et al, supra note 195 at Intro.

\(^{199}\) Monarch Initiative and the TransMed NCATS Data Translator, Response to Request for Information: Metrics to Assess Value of Biomedical Digital Repositories (2016) at “Licensure” (“Not all data resources are free to use, derive, and redistribute, even if they are publicly funded and seemingly publicly available.”).

\(^{200}\) “(Re)usable Data Project”, online: <http://reusabledata.org/>.

\(^{201}\) Wilkinson et al, supra note 87.

\(^{202}\) Monarch Initiative and the TransMed NCATS Data Translator, supra note 199.
practices of publicly available biological databases on a five star scale according to the following criteria: “the findability and type of licensing terms, the scope and completeness of the licensing, the ability to access the data in a reasonable way, restrictions on how the data may be reused, and restrictions on who may reuse the data.”203 In more detail, licenses should be:

1. Documented: Explicit data use terms (ideally formal licenses) should be defined by the resource providers and easy to find.
2. Clear:
   2.1. At a minimum, licenses/data use agreements must be clear and easy to understand. A variety of specific examples of data use/reuse conditions should be included.
   2.2. Licenses should not require negotiation and licenses themselves should be legally redistributable without engaging legal counsel.
3. Minimally restrictive: The licenses and/or data use agreements should explicitly permit downstream data reuse, derivation, and re-dissemination.
4. Standard licenses: … considerations for data are significantly different than those for software and they must be considered separately …
5. Contactable: There should be an appropriate person available for contact with questions about licensure; this person’s contact information should be easy to find.204

These criteria expand on the Research Data Alliance principles of transparency, compatibility, and no negotiation. The importance of a standard license speaks primarily to compatibility but also to transparency, as well-known licenses are more easily understood. The criteria suggest licenses should be comprehensive and explicit with regards to the subject matter covered (e.g., data and code), and the user activities that are forseen. Licenses are also more interoperable where contact information for the rights holder is provided. Even where standard licenses are adopted, negotiation may still be required in some cases.

The (Re)useable Data Project criteria further blur concepts of interoperability and openness. This is natural considering that they focus on databases professing to be publicly available. There may be some legitimate licensing conditions or restrictions placed on publicly available research databases, such as attribution requirements that help to demonstrate the value

203 Carbon et al, supra note 195 at Evaluation Criteria.

204 Monarch Initiative and the TransMed NCATS Data Translator, supra note 199 (“4.1 Standard data license: For data, ideally CC0. 4.2.Standard software license: For software, ideally Apache version 2. Note that software license choices are the subject of much community discussion especially regarding ‘copy-left’ approaches and there are other valid standard options available [such as GPLv2, GPLv3, AGPLv3, etc.]”).
of the database and ensure its sustainability, or share-alike licenses that ensure improvements on
the database also remain publicly available. Some argue share-alike clauses are inappropriately
restrictive for research data. By requiring derivative works to apply the same share-alike license,
they restrict the licensing freedom of downstream users. 205

2 Case Study – Licensing the BRCAExchange
Database

The interoperability barriers and societal costs of accumulating IP restrictions and
conditions are clearly illustrated by the experience of the BRCAExchange public variant
database. The BRCAExchange aggregates genetic variant data from a range of publicly available
data sources. To maximize the interoperability of the data, the BRCAExchange also aims to
provide the data under a single, permissive data license that enables viewing, re-use, adaption, or
re-distribution for any purpose. This goal is stated in the database’s Submission Policy 206 and
aligns with the open data definition. 207 The BRCAExchange hopes its data will be used by
researchers, by clinicians and laboratories for genetic diagnosis, by patients to track and better
understand the health significance of their variants, by pharmaceutical companies in developing
and seeking regulatory approval for precision medicines, and by technology companies to
develop innovative software tools. Applying a standard, permissive license is a commitment to
move from an open science approach to a truly open data approach, to ensure the data are not
only open access but that they are truly in the public domain.

Most source databases of the BRCAExchange are already publicly available, so
aggregation and sharing without licensing restrictions should be trivial. Unfortunately, the
BRCAExchange has encountered a thicket of restrictive or unclear licensing terms associated
with its numerous data sources, described in Table 4.1. This includes restrictions on re-
distribution of data and commercial use, as well as share-alike restrictions (requiring adaptations

Computing in Science Engineering 35 at 37.
206 BRCAExchange, Submission Policy, online: <https://brcaexchange.org/about/dataSubmissionPolicy>.
207 Open Definition, supra note 104.
of the database to be re-distributed under the same viral license). Restrictive, share-alike, and uncertain source database licenses meant the BRCAExchange leadership had to negotiate re-distribution permissions with most source databases. In some cases, it was not possible to reach a new agreement. In others, the source databases were willing to release a limited dataset under the BRCAExchange license. The BRCAExchange still aims to stick with a single, permissive license to facilitate re-use, but this means some data sources will need to be excluded.

The BRCAExchange has so far managed to avoid compromises such as establishing a second, more restrictive licensing tier (e.g., academic use only), or acting as a flow-through entity that avoids licenses altogether and instructs users to negotiate permissions with each source database. The BRCAExchange could act as a conduit to its source databases, allowing each to be governed by its own license. In alignment with the (Re)useable Data Project guidelines for flow-through entities, the BRCAExchange could provide links to the licenses and contact information of source databases, to allow users to seek permission directly (see below). While increasing transparency, this would still leave users of the BRCAExchange with a considerable negotiation burden.

Non-commercial and research-use only restrictions can preclude the integration of data into new, commercial diagnostic tools and targeted therapies. The importance of commercial access to biomedical databases is reflected by the US Food & Drug Administration’s (FDA) recent efforts to encourage the integration of genetic variant databases into regulatory approvals. The FDA now provides official recognition to public genetic variant databases that meet certain quality standards. Recognized databases can be directly relied upon to support regulatory submissions for precision medicines. This streamlines the regulatory process, because drug or device developers do not have to gather all of the evidence about the links between genes, disease risk, and drug response on their own. Indeed, while commercial parties have expressed interest in using the BRCAExchange, they have asked for additional clarity on licensing.


209 Melissa Cline, BRCAExchange Manager, personal communication.
### Table 4.1 BRCAExchange Source Database Licenses

<table>
<thead>
<tr>
<th>Source Database</th>
<th>License characteristics</th>
</tr>
</thead>
</table>
| University of Washington: Database of BRCA1 Saturation Genome Editing Function Scores <https://sge.gs.washington.edu/BRCA1> | - Restrictive (academic-only)  
- Commitment to low-cost licensing for commercial users who follow data sharing best practices. |
- Restrictive member agreement (“privileged”, no publication or distribution w/out permission) |
| The Exome Aggregation Consortium (ExAC): Genome Aggregation Database (gnomAD) <http://exac.broadinstitute.org> | - Open Data Commons, Open Database License (Share-alike v 1.0) |
- Request for attribution in publications and websites.  
- Disclaimer about upstream IP rights. |
| International Genome Sample Resource (IGSR): 1000 Genomes Project <http://www.internationalgenome.org> | - Request for attribution  
- Disclaimer about upstream IP rights. |
- Ambiguous if restricted to academic use. |
| Leiden Open Variation Database (LOVD) <https://www.lovd.nl/3.0/> | - No license.  
- Attribution and sustainability concerns. |

---

210 The Database of BRCA1 Saturation Genome Editing Function Scores was generated using “Saturation Genome Editing (SGE) is a CRISPR/Cas9-based method to functionally test the effects of large numbers of variants in their native genomic context”. BIC was “established to facilitate the detection and characterization of breast cancer susceptibility genes.” ExAC is “a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a variety of large-scale sequencing projects, and to make summary data available for the wider scientific community. The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies.” ClinVar “is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. ClinVar thus facilitates access to and communication about the relationships asserted between human variation and observed health status, and the history of that interpretation. ClinVar processes submissions reporting variants found in patient samples, assertions made regarding their clinical significance, information about the submitter, and other supporting data.” The 1000 Genomes Project “was the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation. Data from the 1000 Genomes Project was quickly made available to the worldwide scientific community through freely accessible public databases”. EVS aims “to discover novel genes and mechanisms contributing to heart, lung and blood disorders by pioneering the application of next-generation sequencing of the protein coding regions of the human genome across diverse, richly-phenotyped populations and to share these datasets and findings with the scientific community to extend and enrich the diagnosis, management and treatment of heart, lung and blood disorders.” LOVD aims to “provide a flexible, freely available tool for Gene-centered collection and display of DNA variants.”
Where licenses permit use of databases for research purposes only, this may also limit valuable clinical uses. By contrast, the BRCAExchange aims to support clinicians and clinical laboratories who interpret the diagnostic significance of genetic variants encountered when testing a patient. Given the expense of commercial laboratory genetic testing, who often use trade secret databases, this publicly available resource is of particular importance in low-resource clinical contexts, such as developing countries.

Avoiding licensing restrictions also allows databases to be used for personal uses. BRCAExchange for example has also developed a mobile application allowing patients to follow the classification of their variant, and receive updates if the interpretation of that variant changes. A patient may receive a genetic test result reporting a variant of unknown clinical significance. The classification of that variant may later be updated to benign or pathogenic as more clinical evidence accumulates. Normally, patients would not be able to receive this information unless they were re-tested. Avoiding use restrictions enhances the interoperability of the BRCAExchange and allows the database to be used to support research, clinical, and personal uses of genetic variant data.

The BRCAExchange’s negotiations with source databases highlighted data producer concerns over attribution, sustainability, and scientific misuse. Even when researchers intended to share data publicly, they remained deeply concerned about the valorization of their data generation, curation, and sharing efforts. Many viewed attribution in publications and in products and services as essential to demonstrate impact, and to attract additional funding. Case-by-case negotiation and legal attribution requirements were perceived as more effective than relying on community citation norms. Some feared standard licenses would not protect against loss of control and unfettered copying. Some adopted default commercial restrictions as they were intent

---

211 BRCAExchange, “BRCA Exchange Mobile App”, online: <https://brcaexchange.org/about/app>.
on negotiating licensing fees from commercial users. These concerns may partially explain why some databases are reluctant to adopt clear, permissive licenses, or to grant permission to the BRCAExchange to do so on their behalf. From a public policy perspective, however, restrictive licensing clearly undermines re-use, while it is unclear that it provides much in the way of benefits for the databases.

The experience of the BRCAExchange echoes challenges reported by other researchers in the literature. Hetionet, another data aggregator, has also reported problems of unclear licensing. Hetionet is a database aiming to aggregate numerous public sources of data that link genes with drug response and disease outcomes. Many of these sources lacked clear licenses, in part because “many researchers don’t understand that simply posting a data set publicly doesn’t mean others can legally republish it.” Moreover, when Hetionet’s creator attempted to negotiate redistribution with the data sources, a number of sources were reluctant to provide authorization, or simply did not respond at all. Licensing may seem like a technicality for publicly available data, especially where data are re-used or re-distributed by academics. But legal barriers accumulate as opportunities for disruptive innovation from outside academia emerge, such as the development of commercially and medically viable software tools.

3 Licensing Pitfalls for Research Data

I now review a number of specific pitfalls that complicate licensing of research data and undermine the interoperability of publicly available health research databases. I focus on peculiar problems that are only noticed through the lens of interoperability, where the consequences of licensing are considered across complex systems rather than within discrete transactions. These pitfalls include: unclear application of standard licensing terms to research activities; uncertainty over the existence of underlying IP rights; conflation between IP and contractual licenses; uncertainty as to whether attribution should be addressed through legal licenses or community

214 https://het.io/.


216 Ibid.
norms; as well as the risks of license proliferation and license accumulation. These challenges suggest licensing of research outputs is non-trivial, and that researchers may need dedicated direction and support from policy-makers, research institutions, and data repositories (see strategies section below).

1. **Unclear application of standard licensing terms to research activities**

Standard licenses permit anyone to copy, re-use, adapt, and re-distribute software, creative works, or data. These permissions may be restricted to non-profit, non-commercial, or (academic) research uses only. Given the various structures of public-private partnerships in health research, however, such terms can be difficult to interpret. Academic research institutions commonly partner with the private sector or seek to commercialize research findings. Many pharmaceutical and technology companies employ data scientists who carry out studies with the aim of publishing in academic journals. Moreover, it can be unclear if these restrictions apply to the data themselves, or also to results discovered from the data, which may include aggregate statistics or even trained artificial intelligence / machine learning (AI/ML) models.217

Standard licenses may also prohibit the distribution of derivative works, or require users to distribute derivative works under a viral, share-alike license. In research contexts, however, it is not always clear what constitutes a derivative work. Does this term only refer to enriched data, or does it also refer to the results of research analyses? Again, results may include statistics or trained machine learning models. An additional twist in AI/ML contexts is that even if a model trained on data is not commercialized or distributed, the outputs of a trained model applied to real-world data (e.g., predictions about patient drug reactions) can hold significant potential commercial value. The Linux Community Data Licensing Agreement license does attempt to distinguish derived data from the results of analyzing data, with share-alike applying only to the former.218 The *Montreal Data License* has recently been proposed as a standard license tailored to AI/ML contexts. The license distinguishes between re-distribution and commercialization of


218 Linux Foundation, *supra* note 166.
data, models (untrained and trained), and model outputs (i.e., predictions). The license also clarifies permissions for a range of specific activities, including research to develop or train models, publishing of models, internal use of model outputs, commercialization of models, and commercialization of model outputs.

New standard licenses may need to be developed for data science or health research contexts. Development will need to balance the need for certainty in a given research area, versus a proliferation of licenses which tends to exacerbate incompatibility. Moreover, as with the open source software movement and the Creative Commons (CC) licenses, there will likely be a need for a research community organization to oversee the development and non-proliferation of standard research licenses.

2. **Uncertainty over the existence of underlying IP rights in research data**

A major reason researchers misapply licenses is the difficulty of determining if there are underlying copyrights in research data and databases. Recall that copyright applies by default if a research output qualifies as an original work. Facts and ideas are not protected, but sometimes researchers make creative choices as they process and annotate data, or organize data into databases. Even where it is unlikely that copyright applies to research data, the absence of a clear license imposes a burden of legal interpretation on users. They may need to seek costly legal advice to determine their rights to re-use, or enter into protracted negotiations with the researcher holding the data. This is unfortunate because the data generator is in the best position to assess if there are underlying IP rights in data, and to state what rights he or she is interested in retaining (if any).

The initial instinct is often to apply a CC license to research data, even though these licenses were developed for creative rather than scientific content. Jonathan Rees *et al.* argue that it is usually inappropriate to apply *any* copyright-focused license, because copyright does not normally apply to research data. The exception is CC-0, which is technically not a license but rather a waiver of rights. The party applying the CC-0 license “overtly, fully, permanently, 

---


irrevocably and unconditionally waives, abandons, and surrenders all of Affirmer’s Copyright and Related Rights and associated claims and causes of action”.  

Because there are some jurisdictions that do not allow copyright to be fully waived or abandoned, the CC-0 includes a fallback clause: a permissive license applies to the extent the waiver fails to apply, which is “royalty-free, non-transferable, non-sublicensable, non-exclusive, irrevocable and unconditional”. Rees et al. argue that CC-0 is the only formal instrument fully appropriate for sharing research data, as it “helps to clarify the legal status of data by signaling that copyright does not apply (or is waived if it does).”

Applying any of the other CC licenses to research databases, however, can lead to legal confusion. Recall that the CC licenses other than CC-0 include some conditions (e.g., attribution), restrictions (e.g., non-commercial use only, no derivatives), or share-alike requirements. If a researcher applying the license does not have any underlying copyright, however, these clauses will have no effect. The CC licenses do not double as free-standing contractual agreements (though some standard data licenses do double as contracts). Researchers who apply CC licenses to research data where they have no underlying copyright have no contractual protection. Moreover, applying a CC license to research data, even if the intent is to share liberally, inaccurately implies copyright is in play, potentially adding to the general confusion in this area.

It may be tempting to apply CC licenses to research data even where the existence of underlying copyrights is uncertain, to provide clarity that even if there are underlying rights they do not restrict re-use, modification, and sharing. Things get more complicated when researchers use CC licenses with conditions, restrictions, or share-alike clauses. Is it appropriate for researchers to use these clauses when they are legally unenforceable, simply to signal expectations? For example, a CC-BY attribution license may be applied to research data simply

221 Creative Commons, “Creative Commons Legal Code”, online: <https://creativecommons.org/publicdomain/zero/1.0/legalcode>.

222 Ibid.

to communicate an expectation: “please cite me”. Arguably, this well-known license is a clear way to communicate such an expectation, even if the legal form is inappropriate and unenforceable.

In response, Jonathan Rees et al. argue that applying a license with restrictions or conditions is disingenuous if the licensor has no intention to take legal action to address infringement. For a single dataset, the decision between a CC-0 waiver and a CC license with conditions, restrictions, or share-alike clauses may not have serious consequences. The real problem, however, is where clauses accumulate over time across multiple datasets, as data are shared, modified, and aggregated. License accumulation, discussed below, can lead to onerous or impractical conditions and excessive restrictions. I recommend that researchers only apply CC licenses to research data when they have good reason to believe they hold underlying IP rights, and where they have an intention to enforce respect for conditions, restrictions, or share-alike requirements. CC-0 remains acceptable, because it waives rights and states the data are in the public domain. Where there is uncertainty over underlying IP in research data, hybrid licenses that function as both IP licenses and as a contracts are a promising strategy discussed below.

Establishing the existence or not of underlying IP rights becomes more complicated as research outputs become more complex and diverse. Outputs may include not only publications, but also data, code, meta-data, as well as charts, graphs, and a range of innovative forms of data visualisations. Licensing a package of diverse but related research outputs need not be difficult if all of the products are subject to the same overall policy. Stodden proposes an umbrella licensing algorithm called the Reproducible Research Standard, which applies to the entire research compendium (publication, data, and code). The Standard clarifies that regardless of the range of underlying rights that apply to the compendium, the only high-level requirement

\[\text{\footnotesize\textsuperscript{224}}\] \textit{Ibid} (“Using a legal instrument, such as a copyright notice, in order to compel behavior is useful only if one is prepared to take on the high cost of litigation. … a pseudo-legal attribution requirement would be, in the most charitable interpretation, a clumsy way of expressing a community norm and, in the least charitable interpretation, an attempt to coerce behavior through empty legal threat.”).

\[\text{\footnotesize\textsuperscript{225}}\] Carroll, \textit{supra} note 120.
imposed is for attribution.\(^{226}\)

3. **IP licensing v.s. contractual data use agreements**

Another pitfall is confusion between permissive licenses, which grant permissions to anyone to re-use protected works without prior approval, and data use agreements – which are contracts that apply to users even in the absence of underlying IP rights. Confusingly, these widely-used agreements are also sometimes referred to as licenses.\(^{227}\) Researchers often fail to distinguish between IP licenses and data use agreements, though there are important differences. This conflation stems from the uncertainty over the existence of underlying rights in the research data. It leads to three overlapping problems: a) unexpected weakness of contractual terms vis-à-vis parties not subject to the agreement; b) excessive strength of contracts against a user who is a party to the agreement; and c) mistaken application of a standard copyright license to research data that is not a protected work, leading to gaps in protection.

*a) Contracts – unexpectedly weak protection vis-à-vis third parties:* Contracts only bind the parties subject to the agreement. This is much narrower protection than copyright, which grants the owner rights against all third parties.\(^{228}\) Copyrights apply automatically, while contracts are only valid where they are entered into voluntarily. In web environments where research data are often exchanged, there can be some uncertainty over the validity of click- and browse-wrap agreements as it is harder to prove the user was provided with, read, and expressly agreed to the terms. Validity usually comes down to whether or not the presentation of the terms constituted sufficient notice.\(^{229}\) In terms of enforcement, copyright owners can seek an injunction in court against anyone infringing their rights, and may also be able to recover statutory damages and costs (e.g., legal representation).\(^{230}\) Enforcing data use agreements, by contrast, “is much more

\(^{226}\) Stodden, *supra* note 205 at 39.

\(^{227}\) Doldirina et al, *supra* note 10 at 22.

\(^{228}\) *Ibid* at 22.


\(^{230}\) National Academies of Sciences & Medicine, *supra* note 90 at 125.
difficult because the author of the terms has to prove that the use has caused measureable economic damages.”\textsuperscript{231} Moreover, especially where data are made publicly available over the web, data use agreements may be practically difficult to enforce because “it is hard to attribute any leaked copy to any particular … signer.”\textsuperscript{232}

\textit{b) Contracts} – \textit{illegitimately strong restrictions vis-à-vis contracting users:} Contracts governing research data use can often be more restrictive and non-standard than IP licenses. Licensing of IP is flexible but somewhat constrained by the bundle of rights provided by copyright law, by limitations and exceptions such as fair use and fair dealing, and by the existence of communities promoting widely recognized, standard licenses. The freedom to contract, by contrast, is theoretically only limited by the creativity of the contracting parties and public order limitations. There is also less awareness or community efforts around standardizing data use agreements in the biomedical community.\textsuperscript{233} Contracts that lock up research data may confound the aims of IP laws that treat data as in the public domain\textsuperscript{234} and funding agency requirements that researchers make their data available to the research community for broad re-use.

\textit{c) Misapplication of content-specific licenses} – \textit{unexpected gaps in protection:} Another problem is where researchers apply copyright licenses with conditions or restrictions when there are no underlying IP rights. For example, a CC-BY license grants permission for re-use and sharing of a protected work without prior permission, on the condition of attribution. This license, however, is often applied to research data where there is no underlying IP. Because the license only applies to copyright-protected materials (it is not a contractual data use agreement), users would be under no legal obligation to provide attribution, defeating the purpose of applying the license. Licenses designed specifically for data and databases tend to also apply as contracts, in case there are no underlying rights.

\textsuperscript{231} Ibid.

\textsuperscript{232} Rees et al, \textit{supra} note 223.


\textsuperscript{234} National Academies of Sciences & Medicine, \textit{supra} note 90 at 125.
The best way to deal with uncertainty over underlying rights in research data and to ensure comprehensive agreements is through hybrid data licenses that function as both IP licenses and contracts (see licensing strategies below).

4. Attribution: inappropriate application of legal licenses to enforce community norms

The case of attribution raises the issue of whether or not researchers should articulate community norms in legal form. A central element of research data licensing debates is whether or not to legally require attribution through a license, or to leave attribution to community norms. The Research Data Alliance is highly skeptical of licenses that legally require attribution, preferring the CC-0 public domain waiver to the CC-BY attribution license. They argue that attribution should not be treated as a legal requirement, but rather as an ethical norm. In their view, even simple legal conditions like attribution can rapidly accumulate as data are combined, modified, and re-distributed. Jonathan Rees et al. agree:

Legal attribution requirements are inferior to community norms - attribution requirements can lead to “stacks” of attributions whose management is challenging and error-prone. … the way to solve this is by participating in initiatives that work on better tools to manage attribution of integrated and recombined datasets, rather than making scientists believe they may be violating the law, or worse yet, that they cannot recombine the data in the way they need to because of technical or legal limitations.235

Victoria Stodden, by contrast, argues that research outputs should include a viral attribution clause, so researchers can receive appropriate credit for their work.236 An attribution requirement, she argues, “largely mirrors how scientific work is typically cited and built upon, with the difference that the attribution process is formalized in a legal license, as opposed to academic citation.”237

This technical debate about the interoperability of legal attribution clauses seems overblown. In my view, what matters is that norms are clearly communicated. In collaborative environments, Yochai Benkler explains that “articulating social mores can go a long way into getting everyone to align their behavior without the need for rewards, punishment, or monitoring.

---

235 Rees et al, supra note 223.
236 Stodden, supra note 205 at 39.
237 Ibid.
and control.” Academic researchers are expected by their peers to describe the scientific provenance of their data and ideas, so serious failures to provide attribution are unlikely. Those imposing attribution licenses rarely seek to enforce these requirements through community means, let alone through the courts. Simple, standard licenses function perfectly well as communication tools. However, contractual forms that leave no interpretive wiggle room for diverse use contexts, or that emphasize a threat of punishment are inappropriate. Such threats can sow distrust, and may paradoxically discourage collaborative behavior.

5. **License proliferation**

A theory of interoperability forces one to think about the consequences of individual licensing choices across complex systems. The (Re)useable Data Project found in its survey of publicly available biological database licenses that the largest single type of licenses were custom licenses, suggesting that resource providers either felt that a standard license did not meet their needs or that they were not knowledgeable about standard licenses.

Misapplication of licenses is aggravated by license proliferation, where there are simply too many licenses that it becomes difficult to assess compatibility or to predict it downstream. In standard licensing communities, “there is a general consensus that new licenses should not be created unless really necessary.” A proliferation of licenses reduces clarity and familiarity, raising the demand on users to legally interpret licensure. Moreover, license proliferation can lead to a greater chance of policy conflict. Incompatibility is already a problem between common forms of open content and software licenses. The OpenMinTeD project has developed a Licence Compatibility Matrix to promote legal interoperability in the context of Text and Data Mining.

---


239 Ibid.

240 Carbon et al, *supra* note 195 at Discussion.


242 Ibid.
research, though the Matrix has broader applications for digital research resources.\textsuperscript{243} Data use agreements are even more susceptible to proliferation as researchers are free to pick and choose contractual terms.

6. \textit{License accumulation or “stacking”}

License restrictions and conditions have a tendency to accumulate in health research data. The issue of license accumulation can be analogized to the accumulation of a chemical like DDT in the food chain, benign at lower levels but seriously affecting species at the top of the food chain. With research, the issue is partly due to a lack of clear “exchange” of data ownership as data are shared between researchers and repositories. Consider the following terms of use on molecular data applicable to databases hosted by the US National Centre for Biotechnology Information (NCBI):

NCBI itself places no restrictions on the use or distribution of the data contained therein. Nor do we accept data when the submitter has requested restrictions on reuse or redistribution. However, some submitters of the original data (or the country of origin of such data) may claim patent, copyright, or other intellectual property rights in all or a portion of the data (that has been submitted). NCBI is not in a position to assess the validity of such claims and since there is no transfer of rights from submitters to NCBI, NCBI has no rights to transfer to a third party. Therefore, NCBI cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in the molecular databases.\textsuperscript{244}

A similar example is the European Bioinformatics Institute terms of use, which states in relation to the databases it hosts that

\[ \text{the original data may be subject to rights claimed by third parties, including but not limited to, patent, copyright, other intellectual property rights, biodiversity-related access and benefit-sharing rights.} \textsuperscript{245} \]

\textsuperscript{243} “Consulting on Licences for TDM”, online: OpenMinTeD <http://openminted.eu/omtd-services/tdm-consulting-on-licenses/> (“The matrix encodes in a clear way the compatibility among popular licences and terms of use of data resources, software and web services, i.e. aims to assess the possibility that resources licensed under specific licences can indeed be combined and their combination feasibly result in a derivative work. It also assesses whether, considering specific clauses [e.g. Non Commercial Use or Share Alike], there is likely to be a conflict between the licensing terms considered.”).

\textsuperscript{244} “Policies and Disclaimers - NCBI”, online: <https://www.ncbi.nlm.nih.gov/home/about/policies/>.

\textsuperscript{245} European Bioinformatics Institute, “Terms of Use for EMBL-EBI Services”, online: <https://www.ebi.ac.uk/about/terms-of-use>.
Even though the terms of use encourage anyone to download and re-use molecular data, Michael Carroll argues the resulting legal uncertainty “interferes with the productive reuse of research data.”

Problems of license application are aggravated further by the challenges of tracking licensure, monitoring use, and enforcing legal rights in the case of misuse. The resulting problem is that licensing information is lost in digital environments, leading to a lack of clarity. Perhaps the one thing worse than an accumulation of licenses is an accumulation of doubt over what licenses apply.

The problem of accumulation is compounded in situations where datasets are derived from other datasets that were not labelled or are improperly labelled. Above I discussed the problem of applying a content-specific license to research data that does not qualify as a protected work, which may lead to an unexpected lack of protection. Content-specific licenses have another related weakness: they often apply only to the author’s original contribution, and not to other works that may be incorporated. In other words, application of a CC license is not evidence that the author has “cleared” pre-existing rights that third parties may hold in the data. Quite the opposite. Just like the NCBI terms of use cannot guarantee that the rights of data contributors to NCBI databases have been cleared, a CC license provides no guarantee that the author has cleared rights in upstream works. The CC-0 license, for example, disclaims responsibility for clearing the rights of third parties.

The Creative Commons recommends addressing this problem with comprehensive and transparent licensing practices. Licenses can include a caveat that they apply “except where

246 Carroll, supra note 120 Introduction.

247 Doldirina et al, supra note 10 at 19 (“Version tracking of datasets is relatively widespread and encrypted hidden identifiers may be carried by datasets, but there is no widely deployed technology currently available to comprehensively track extractions from datasets used in subsequent derived datasets. There is thus no way at this time to automatically carry forward the license and contract legal conditions nor the applicable national laws that might be germane to those extractions. For these pragmatic reasons, widespread legal enforcement of attribution and other licensing provisions as applied to contributions to derivative datasets is typically very difficult in the current technological environment.”).

248 Creative Commons, supra note 221.
otherwise noted”, and content subject to different restrictions can be clearly identified.\(^{249}\) One of the initiatives contributing to the Reusable Data Project has framed this issue as a problem for “flow-through entities” aggregating or providing links to datasets:

L2: Transparent about flowthrough implications

If others’ data is redistributed, clarity about the licensing implications of the redistribution is critically important. Concrete metrics:

- Documentation about which source resources/data, if any, come with flowthrough implications.
- Links to the original licenses/data use terms of all redistributed content. It is currently commonplace that such terms do not exist; in such cases, it should be clearly stated that license/terms could not be found.
- If specific authorization has been obtained for redistribution.\(^{250}\)

While this seems like a simple solution, it requires significant effort on the part of researchers or repositories sharing scientific knowledge. It also exposes them to liability risks because researchers or repositories essentially make representations about the existing rights in data. This is probably why so few research repositories are willing to affirmatively state their data are free of IP rights. So how can the research community improve the legal interoperability of publicly available research outputs?

### 3.1 Improving the Interoperability of Research Data Licenses

Researchers can adopt a number of strategies to improve the interoperability of licenses associated with publicly available research data. But researchers cannot be expected to resolve these challenges alone. They need guidance and support with licensing from their institutions and technology transfer offices, policy-makers, and repositories.

1) **Researchers: Adopt hybrid licenses for research data**

Uncertainty over underlying rights and the resulting confusion between IP licenses and contracts can be avoided by using hybrid licenses that function as both IP licenses and contracts.

---

\(^{249}\) Creative Commons UK, “Frequently Asked Questions on Creative Commons & Open Access” (2017), online: <https://zenodo.org/record/841086#.WYWwTWYpLtE4> at 2.

\(^{250}\) Monarch Initiative and the TransMed NCATS Data Translator, *supra* note 199.
An example of data licenses that take this approach are the Linux Foundation Community Data License Agreement licenses. To the extent the data involves copyright, express permissions are granted and conditions or restrictions are imposed. Where there is no copyright, the license has a fallback to outlining the same permissions, restrictions, and conditions in a contract. While the bilateral nature of a contract provides weaker protection, this is still the best protection that can be obtained for data without underlying IP. Indeed, “[t]he Agreement is intended to ensure that all parties give and receive uniform, predictable rights in Data” regardless of data type or jurisdiction.251

2) Researchers: Remove legal restrictions and conditions

The simplest way to avoid licensing interoperability issues is to reduce or eliminate IP restrictions or conditions associated with research data. Guideline 3 of the Research Data Alliance Principles emphasizes responsibilities and strategies to reduce or eliminate default legal constraints on access to and reuse of research data generated from government or publicly-funded collections.252 If researchers commit (to funders and participants) to make the products of their research publicly available, they should also explicitly commit to do so under a standard, open-conformant license. Ideally, this commitment should be made at the time the research project is proposed, so it is clear to funders and participants exactly how open and interoperable the outputs will be. Otherwise, there is a risk that researchers make data technically, but not legally, available. This can lead to a “look, but don’t touch” scenario for some prospective data users. Committing to clear licensing at the beginning of a research project will also force researchers to think carefully about the overall data trajectory. This is analogous to the principle of ethical horizon, which “refers to the perspective of investigators at the time of data creation who look to future uses of the data, their own and all subsequent investigators.”253 Similarly, licensing considerations relating to downstream data sharing can be taken into account at the

251 Linux Foundation, supra note 165.
time the research protocol is designed. This strategy is important for all research outputs, as it is often hard to predict when these outputs will become valuable inputs for future research. This is demonstrated by text and data mining, which uses traditional research outputs such as publications as the raw materials. Copyrights in publications, even open access publications, remain an important barrier to such activities.\textsuperscript{254}

3) Policy-makers and Repositories: Encourage use of open-conformant licenses

Urs Gasser describes a number of private or community approaches to addressing interoperability including standards and open standards.\textsuperscript{255} In the case of IP, research communities can attempt to agree on standard licenses and data use agreements. They can develop community templates through consensus-based discussion. The problem with a bottom-up approach like this is that adoption still depends on researcher voluntariness, and appropriate implementation still depends on a level of legal competence. Simply providing a template as a web resource does not mean researchers will be aware of it, will bother to adopt it, or will adopt it properly.

Data repositories can play a role by guiding, or even mandating, researchers contributing resources to share openly. According to Michael Carroll, they can require “depositors to grant permission to downstream users or to give up any intellectual property rights they may have in the data.”\textsuperscript{256} Peter Desmet argues that repositories should require a transparent statement of rights (in machine-readable form) and “require CC0, and shift the discussion about ethical data use (including attribution) to norms rather than ill-suited legal tools.”\textsuperscript{257} This can reduce legal uncertainty for users and encourage re-use. However, even the major government-funded repositories in the United States and Europe discussed above are reluctant to impose such strict requirements on data contributors. It is not clear if this out of concern for discouraging voluntary

\textsuperscript{254} Richard Eckart de Castilho et al, “A Legal Perspective on Training Models for Natural Language Processing”, (2018), online: <http://eprints.gla.ac.uk/159231/>.
\textsuperscript{255} Gasser, supra note 8.
\textsuperscript{256} Carroll, supra note 120 at Introduction.
\textsuperscript{257} Peter Desmet, “Analyzing the Licenses of all 11,000+ GBIF Registered Datasets”, online: <http://peterdesmet.com/posts/analyzing-gbif-data-licenses.html>.
contributions, or for liability resulting from encouraging infringement. Where repositories rely on voluntary submissions, and where they have limited budgets, it is unlikely they will be able to impose such requirements.

An alternative would be for repositories to adopt an open-by-default approach. Contributors would be allowed to retain rights in data exceptionally, as long as the nature and scope of the rights are clearly stated. Additionally, they could be asked to specifically justify their reasons for restricting use. Requiring clear (and detailed) justification exposes contributors to the judgment of the research community, and has a helpful side-effect of generating data about “the precise nature of legal barriers to disclosure and their appropriateness…” Presumably some of the reticence to explicitly license research outputs is fear of criticism for restrictions or conditions they impose.

A step further, towards a more top-down approach, would be for policy-makers (e.g., funding agencies or even legislators) to impose a standard license or a limited menu of standard licenses for the outputs of publicly funded research. Indeed, one of the projects behind the (Re)useable Data Project called on the US National Institutes of Health in a community letter to apply a standard license to biomedical research data. A problem in the health research domain, however, is that while licensing interoperability problems are recognized, there is still significant complexity, confusion, and lack of consensus over what standard terms should apply. Licensing initially appears to be a technical issue, but the persistent confusion surveyed here suggests confusion extends to the policy level: What boundaries should be placed on openness to incentivize and ensure the sustainability of contribution efforts? Or to ensure a fair distribution of the benefits of openness? Or to preserve the openness of scientific knowledge over time?

258 Stodden, supra note 91 at 116 (Stodden proposes as part of her “Principle of Scientific Licensing” for reproducible computational research, which extends to the research compendium [including publication, data, and code] that “[l]egal encumbrances to the dissemination, sharing, use, and reuse of scientific research compendia should be minimized, and require a strong and compelling rationale before their application.”).

259 Ibid.

260 Haendel et al, supra note 179.

261 Carbon et al, supra note 195 at Introduction.
Ultimately, there is an important link between standard licensing, which initially seems like a technical issue, and the sustainability and valorization of datasets, which are key policy issues for researchers curating and sharing data. Often a major concern for databases is appropriate attribution. They may attempt to address this through legal licenses either because they do not have faith in community norms, or because they misapprehend the nature and consequences of legal licenses. On one hand, effective attribution is arguably essential for sustainability and the ability of the database to attract funds for continued data generation and curation. On the other hand, it could be argued that the true value of the database is in how easily and how often it is “incorporated into other resources, extending its value, length of use, and audience reach.” Until better policies, use-tracking tools, and metrics are developed for health research databases, the licensing interoperability issues explored here are likely to persist.

4) **Data Aggregators: Adopt open data guarantees**

A novel strategy would be for researchers and repositories to provide some form of legal assurance their data are in the public domain. This would be a contractual analog to the CC licenses, but would also apply where there is no underlying IP. The goal of the CC licenses is simple: to affirmatively state that a protected work can be re-used and shared without prior permission from the copyright owner. It is clear such a tool is needed given the automatic restrictions copyright places on the re-use, modification, and sharing of original works of authorship. For research data, uncertainty over underlying IP rights and poor licensing practices erode certainty, especially as data are aggregated.

I propose that data aggregators make efforts to clear third party rights in data, and provide assurances to users that this has been done. This strategy is hinted at by the NCBI terms of use (see Table 4.1). The terms state that there are no underlying IP rights in molecular data, and expressly states that the NCBI does not impose additional contractual restrictions on molecular data. The NCBI does not, however, provide assurances that rights have been cleared in third party data, as it does not require data submitters to affirm they retain no IP rights. Even popular, standard licenses (e.g., Creative Commons) undermine certainty by disclaiming provider responsibility that the data are free of third party copyrights.

---

262 *Ibid* at Discussion.
The Linux Community Data License Agreement licenses include a valuable clause where the Data Provider represents that reasonable care has been exercised to ensure the data are free of third party IP constraints or privacy and confidentiality obligations. This provides additional certainty to users of repositories that the repository license is accurate and comprehensive:

4.1 Each Data Provider represents that the Data Provider has exercised reasonable care, to assure that: (a) the Data it Publishes was created or generated by it or was obtained from others with the right to Publish the Data under this Agreement; and (b) Publication of such Data does not violate any privacy or confidentiality obligation undertaken by the Data Provider.263

Data aggregators like the BRCAExchange seeking to make such representations will bear a greater burden of the legal analysis and negotiation. But this is a worthwhile value-added service, as it will save numerous data users from having to go through a similar process.

5) Research Institutions and Technology Transfer Offices: Provide licensing support

Institutions need to support researchers with license selection. A potential candidate for delivering such support in universities is the technology transfer office (TTO). TTOs are responsible for establishing the business model of a university.264 Traditional TTOs focus on commercializing the knowledge generated by academic departments through patenting and launching start-ups.265 Universities and TTOs are shifting towards a wider conception of technology transfer to encompass social and economic well-being.266 Some adopt a “smart bazaar” approach aiming to openly disseminate science and educational courses, preferring disclosure rather than patenting of inventions, and non-exclusive licensing.267 This trend could extend to more collaborative and open approaches to licensing.

263 Linux Foundation, supra note 166.
267 Baglieri, Baldi & Tucci, supra note 264.
Challenges would remain. Researchers may be unwilling to work through these offices if there are concerns over delays or ceding of control over their research outputs. A shift by some TTOs towards more open business models may increase the already significant diversity of licensing policies across Canadian universities, adding to frustration for investors and industry. Such a change would require a substantial shift in culture, training, and metrics for success.

It is also unclear how open approaches will affect the international competitiveness of Canada’s research sector. Concern over US dominance and the transfer of Canadian publicly funded research to other countries supports calls for stronger commercialization practices in universities. Indeed, some technologies developed at Canadian universities are virtually being given away to Chinese and US companies. Huawei for example has invested $50 million with a network of 13 Canadian universities to create a steady pipeline of IP in telecommunications technologies fully transferred to the company. Canada’s innovation strategy is questioned when public investment simply strengthens foreign patent portfolios, and Canadians have to pay for technologies a second time in the form of royalties. More open approaches are thought to accelerate innovation generally, but there is no guarantee this acceleration will directly benefit Canada. On the one hand, there may be an opportunity for Canada to distinguish itself based on open forms of innovation. On the other hand, the best-resourced countries and companies may be the first to capitalize on open resources. This debate cannot be settled without first experimenting with open data and permissive licensing approaches like that of the BRCAExchange.

## 4 Conclusion and Recommendations

This Chapter began by introducing the Research Data Alliance’s criteria for research

---

268 Rasmussen, supra note 266 s 2.3.
269 Bubela & Caulfield, supra note 266.
license interoperability. Publicly funded research data should be made available under standard licenses that enable combination and re-use of data without case-by-case negotiation with providers. Similarly, the (Re)useable Data Project highlights the important of clear, comprehensive, standard, and non-restrictive licenses for publicly available biological and biomedical datasets.

Health research outputs are increasingly made publicly available, but responsibility for licensing these outputs is still largely left to individual researchers. Reliance on private-ordering continues to result in inconsistent licensing practices, limiting the utility of available data. Even where databases are publicly available and not likely subject to IP protection, they are rarely accompanied by a permissive, standard license. The efforts of the BRCAExchange to aggregate publicly available genetic variant data under a single, permissive license continue to be hampered by the diverse and ambiguous licenses of its source databases, resulting in high negotiation costs for data users.

Researchers can enhance the interoperability of licenses by selecting standard and permissive licenses. For research data, hybrid licenses that function as both IP permissions and contracts are advisable. Policy-makers, repositories, and technology transfer offices should support researchers in selecting appropriate licenses. Finally, data aggregators can enhance interoperability by assuring users that third party rights have been cleared from research data.
Chapter 5
Interoperability and Data Privacy Law

Key Points:
- Data privacy law can apply restrictively to health research, especially international collaborations.
- Uncertainty over identifiability, responsibility, and having a legal basis to conduct research involving identifiable information complicates interoperability.
- Sector-specific codes of conduct can act as legal “interfaces” between data privacy regimes, facilitating compliance and interoperability across jurisdictions.
- The Human Cell Atlas, an international health research collaboration aiming to map how gene expression varies across all human cells, illustrates how data privacy law tends to act as a barrier to health research interoperability.
- Researchers can facilitate data exchange by adopting interoperable consents, data sharing contracts, and privacy and security standards.

Data privacy law aims to protect the privacy of individuals from excessive intrusion by public and private entities engaged in automated processing of personal data i.e., data about an identifiable individual. I employ the term data privacy law to encompass European data protection laws, and Canadian personal information protection laws. Data privacy law typically imposes restrictions and obligations on entities managing databases containing personal data, and is therefore an important regulatory framework for health research. Clinical informatics and bioinformatics (e.g., genomic, neuroimaging) research involving human participants often face compliance challenges. Rich bioinformatics data tends to be unique to the individual, and thus potentially identifiable. They may also be linked with, or may directly contain sensitive information about an individual’s health status. The central privacy concern is that if an unauthorized party gains access to an individual’s identifiable health information, the individual may be harmed through discrimination or stigmatization.

This Chapter begins with a brief introduction to data privacy law. It then discusses the problem of insufficient interoperability between the data privacy regimes of different jurisdictions, which can frustrate the flow of data across borders, particularly for international research collaborations. It is not only variation between laws that presents a problem, but also
hyper-territoriality – where multiple states assert jurisdiction over the same data processing activity – and data localization such as restrictions on cross-border transfers. I also discuss a problem of “fit” between data privacy law and health research. Data privacy law is broadly concerned with commercial, government, and healthcare processing of personal data to make decisions that directly affect the rights and interests of individuals. Health research, by contrast, focuses on producing generalizable knowledge, not on influencing the rights and interests of individuals (with the exception of returning individual findings of clinical relevance to participants). This problem of fit is reflected by persistent uncertainty in health research contexts over identifiability, controllership, and the legal basis for processing. Challenges with data privacy and international research collaboration are illustrated by a case study of the Human Cell Atlas.

I then review strategies for increasing interoperability of data privacy rules across jurisdictions, including harmonization through international treaties and trade agreements, as well as standardization through sector-specific, legally-binding codes of conduct. Ultimately, I conclude that such efforts are unlikely to progress in the current political climate. Even if they do, they are unlikely to provide additional interoperability for research. Even rigorous European data privacy law, though generally aiming to promote harmonization and friendly to research, leave much of the research-specific details to Member State law.

Legal interoperability is not only a matter of compatibility between legal regimes, but also a matter of interoperability between the legal tools applied by researchers, such as consents to broad research purposes (and/or cross-border transfers), data sharing agreements, and privacy and security safeguards. The freedom of researchers (and participants) to establish high levels of interoperability under data privacy law is contested and more limited under Europe’s collectivist regime. Beyond these limits, the central weaknesses of private-ordering approaches to establish interoperable outcomes – negotiation, voluntariness, and mis-labelling – are evidenced in the data privacy context.

1 Introduction to Data Privacy Law

Concern over the increased levels of automated processing of personal data (data about an identifiable individual) by public and private sector organizations led to the development of
influential, international guidelines on privacy by the Organization for Economic and Cultural Development (OECD). These guidelines outlined the following foundational principles:

<table>
<thead>
<tr>
<th>Table 5.1 OECD Fair Information Processing Principles</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collection Limitation Principle:</strong> There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.</td>
</tr>
<tr>
<td><strong>Data Quality Principle:</strong> Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.</td>
</tr>
<tr>
<td><strong>Purpose Specification Principle:</strong> The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.</td>
</tr>
<tr>
<td><strong>Use Limitation Principle:</strong> Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with [the Purpose Specification Principle] except: (a) with the consent of the data subject; or (b) by the authority of law.</td>
</tr>
<tr>
<td><strong>Security Safeguards Principle:</strong> Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorised access, destruction, use, modification or disclosure of data.</td>
</tr>
<tr>
<td><strong>Openness Principle:</strong> There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.</td>
</tr>
<tr>
<td><strong>Individual Participation Principle:</strong> An individual should have the right: a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; b) to have communicated to him, data relating to him within a reasonable time; at a charge, if any, that is not excessive; in a reasonable manner; and in a form that is readily intelligible to him; c) to be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial; and d) to challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended.</td>
</tr>
<tr>
<td><strong>Accountability Principle:</strong> A data controller should be accountable for complying with measures which give effect to the principles stated above.”</td>
</tr>
</tbody>
</table>

---


Data privacy law can be briefly summarized in terms of consent, confidentiality, and individual rights: personal data is not to be processed (i.e., collected, used, or disclosed) without the consent of the individual or another legal basis, and personal data shall be kept confidential and protected by appropriate security safeguards. Individuals also have certain legal rights that must be respected by entities holding their personal data, including rights to access a copy of their personal data, to portability (see Chapter 2), to rectify errors, and to receive a description of how their data have been processed. This simplistic summary could extend to include emerging principles and obligations including privacy-by-design. Privacy by design emphasizes embedding privacy as a priority in all aspects of organizational governance. Processes should be in place to protect privacy, including education and training of technical staff, privacy impact assessments, identifying personnel responsible for privacy, and breach handling/reporting.\textsuperscript{274} This expanded compliance burden increases the challenge of ensuring interoperability across organizations and countries.

The right to data protection is related to but distinct from the right to privacy. Privacy is a broader concept, including the privacy of spaces, persons, and information.\textsuperscript{275} According to Edward Dove, privacy “embodies a broad range of rights and values, such as the right to be let alone, intimacy, seclusion, and personhood.”\textsuperscript{276} Data protection focuses on informational privacy. It is often applied comprehensively across all sectors, whereas the right to privacy is typically considered as a matter of public law in the context of law enforcement. Data protection involves positive obligations to protect personal data and “requires a more proactive approach”,

\begin{footnotes}
\footnotetext{274}{Woodrow Hartzog, \textit{Privacy’s Blueprint: The Battle to Control the Design of New Technologies} (Harvard University Press, 2018) at 12.}
\footnotetext{275}{\textit{R v Dyment}, [1988] 2 SCR 417 (C) (available on http://canlii.ca/t/1fte6) para 19.}
\end{footnotes}
whereas privacy is more classically a negative right.\textsuperscript{277} Data protection is arguably also less context-specific so therefore easier to enforce.\textsuperscript{278}

The European Union \textit{General Data Protection Regulation} (GDPR) is a prominent data privacy law with direct application across European Member States. The GDPR is an omnibus law, meaning that it covers data processing across all public and private entities, with limited exceptions and some extended sector-specific rules.\textsuperscript{279} The default rules of the GDPR are largely designed to constrain excessive processing of personal data by governments and powerful commercial companies. This privacy-by-default stance raises concerns about unintended consequences not only for digital innovation, but also for scientific collaboration.

Two trends can be identified in European law: 1) continued strengthening of data protection regimes, and 2) concerted effort towards greater consistency of rules across Member States.\textsuperscript{280} This reflects an important quality of legal interoperability. The stronger the requirements associated with data, the greater the need for compatibility requirements between countries or organizations. The GDPR also contains important data localization rules, which restrict transfers of personal data outside of Europe. There are some legal avenues still available, including adequacy and appropriate safeguards. Transfers can be made to foreign jurisdictions where their laws have been deemed by the European Commission to offer an adequate level of protection (art 45), or where appropriate safeguards are in place such as regulator-approved contracts (art 46,47).

In the US, privacy protections are largely sector-specific, with some notable state-level data privacy laws (e.g., California). A well-known example of federal sector-specific regulation is the US \textit{Health Insurance Portability and Accountability Act} (HIPAA) and its Privacy Rule,

\textsuperscript{278} \textit{Ibid} Abstract.
\textsuperscript{279} Dove, \textit{supra} note 276 at 1014.
which applies to the healthcare sector.\(^\text{281}\) Sector-specific laws tend to only apply to certain types of entities (e.g., HIPAA’s “covered entities” and their “business associates”) and certain types of data (e.g., HIPAA’s “protected health information”).

Canada has personal information protection laws at both the federal and provincial levels that, while a patchwork, offer comprehensive coverage of the public, private, and healthcare sectors.\(^\text{282}\) These laws govern the collection, use, and disclosure of personal information, or in the case of health-sector laws, personal health information. The federal private sector law, the *Personal Information Protection and Electronic Documents Act* (PIPEDA), establishes a national baseline. Provinces may adopt private sector laws that displace PIPEDA if the law is deemed to be “substantially similar” to PIPEDA by the Governor in Council (the Governor General of Canada acting on the advice of Cabinet).\(^\text{283}\) PIPEDA also applies to transfers across provincial or national borders.

PIPEDA has formally been deemed adequate to protect the transfer of European’s data to Canada by the European Union. There is some uncertainty as to whether or not PIPEDA’s privacy safeguards are sufficient to maintain its adequacy status under the GDPR.\(^\text{284}\) There is also some uncertainty as to the adequacy status of provincial private sector laws, despite their substantially similar status.\(^\text{285}\) With regards to international data transfers in the private sector, Canadian law largely adopts an accountability model, where the Canadian data custodian holding


\(^{282}\) Adrian Thorogood et al, “Protecting the Privacy of Canadians’ Health Information in the Cloud” (2016) 14:1 Canadian Journal of Law and Technology 173.


personal data must ensure appropriate contractual or other safeguards are in place.\textsuperscript{286} However, the federal privacy regulator, the Office of the Privacy Commissioner of Canada has recently issued draft guidance moving towards a default individual consent requirement for international transfers of personal information (see “Data Localization” critique below).\textsuperscript{287}

The influential European approach to data protection has raised a conceptual muddle and corresponding scholarly debate as to whether data protection rights are (or should be) property rights or personality rights.\textsuperscript{288} Property rights are a bundle of exclusive rights, a form of sovereignty, over a resource. Personality rights protect an individual’s being and identity, and “aim to guarantee and preserve that individual’s sense of their own existence.”\textsuperscript{289} The former view sees personal data as an external resource, the latter as something intrinsic to an individual’s identity. European conventions, data protection legislation, and jurisprudence do not explicitly categorize data protection rights as one or the other.\textsuperscript{290} The strengthening of individual rights (e.g. to be forgotten, access, portability) under the GDPR suggest they are property rights. It is also clear that personal data have economic value and exchange hands in transactions. This would mean that even if data protection law does not establish default entitlements, it has no choice but to regulate ownership. If data protection rights related only to one’s personality, then they would only confer the right to control personal data in certain circumstances.

It appears data protection rights are evolving from classical negative rights focused on the liability of data processors to personality rights. While personal data may be alienable, personal data rights would not be, because they are linked to individual dignity and identity.\textsuperscript{291} Henry


\textsuperscript{288} Pearce, supra note 280.

\textsuperscript{289} Ibid at 193.

\textsuperscript{290} Ibid at 199.

\textsuperscript{291} Ibid at 203–04.
Pearce argues that European data protection law offers a quasi-property right, providing limited exclusionary rights in particular contexts or relationships. Data protection rights are indeed framed as rights emerging from a relationship between data controllers and data subjects. A primary motivation of data protection law has long been to protect “individuals from events capable of harming their wellbeing that were thought to stem from their relationships with private and public-sector organisations.”

Depending on whether a jurisdiction leans towards property or personality conceptions, data protection law may be understood differently: does data protection law provide a guarantee of exclusivity in personal data, or something less? Do such rights trump other interests, or are they subject to balancing? Can data protection rights be waived?

In health research, data need to flow between researchers, institutions, and countries to enable greater collaboration, transparency, and reproducibility, efficiency, and innovation. Data privacy law places limits or conditions on such flows, and is typically viewed by the research community as contributing to sub-optimal levels of data exchange. The concept of legal interoperability is key to resolving the apparent tension between data sharing policy and data privacy law. Where the data privacy laws of different jurisdictions are interoperable, law-makers, institutions, and citizens should be able to trust the flow of data across borders. Where the privacy and security safeguards offered by different institutions are interoperable, data should be able to flow between them with relative ease. Stronger data protection rules in some regions, however, may challenge the willingness or ability of other countries and their organizations to harmonize their laws or governance upwards. And stronger data protection rules may also need to be accompanied by stronger countervailing interoperability requirements. This is why the GDPR is an attempt to both strengthen European rules and to harmonize them to enable cross-border flows.

Ibid at 208.
2 Case Study of the Human Cell Atlas: Cellular Gene Expression Data Gets Personal

The Human Cell Atlas provides a prominent and current example of the challenges of global health research collaboration in an era of strengthening data privacy law. The Human Cell Atlas is an international collaboration aiming to map how gene expression differs across cell types – the fundamental units of life. The differentiation and specialization of human cells into tissues and organs is largely a function of their gene expression (as every cell has roughly the same genome). New single-cell genomic and RNA sequencing technologies, combined with high resolution tissue imaging technologies, enable the study of gene expression. Molecular characterization of hundreds of thousands of cells can be done simultaneously (massively parallel sequencing). These massive quantities of data are parsed by powerful computational algorithms. The HCA is also envisaged as an “open resource” that will dramatically accelerate discovery about biology and insights into disease diagnosis and treatment. Data will be made openly available on the HCA’s Data Coordination Platform “to the maximal extent allowed by law and ethics”.

The focus on sequencing tissue types distinguishes scientifically, legally, and ethically from previous genomic sequencing and data sharing initiatives. There are four major sources for tissue: tissue biopsy from research participants, left-over tissue from clinical procedures, deceased transplant organ donors, and postmortem examinations (i.e., rapid, “warm” autopsy). These last two sources complicate assessment of legal compliance, as data privacy law regimes and health research regulation do not typically extend to deceased participants. Moreover, designing consent models addressing research and data sharing in surgical, organ-transplant, or peri-mortem contexts raises complex challenges for tissue biobanks. Who consents to research, sample, and data sharing (the individual, family member, or proxy)? How much information can

---

293 “Human Cell Atlas”, online: <https://www.humancellatlas.org/>.


295 Ibid at 6.

296 Regev et al, supra note 294 Appendix I.
meaningfully be provided in these sensitive consenting contexts? Examples of organ transplant and post-mortem biobanks include the Genotype Tissue Expression project (GTex) and the Cambridge Biorepository for Translational Medicine.\textsuperscript{297} GTex has successfully enrolled donors for post-mortem tissue sampling and full open-access sharing of resulting data.\textsuperscript{298} These biobanks face data privacy concerns and impose contractual obligations on accessing researchers which can complicate subsequent data sharing.

### 3 Data Privacy Law: Uncertainty and Lack of Fit for Health Research

This section discusses the awkward fit of data privacy law, namely the European variety, to health research. The awkward fit results in persistent compliance uncertainty for researchers. First, it is hard to determine in health research contexts if data are identifiable, pseudonymized (regulated as personal data under the GDPR), or anonymous. Second of all, it can be hard to determine who is responsible for data privacy compliance across international research collaborations. Third, while research purposes have a privileged status under the GDPR, Member States are also granted extensive flexibility in their regulation. Rules can differ significantly across Member States. The resulting legal uncertainty contributes to a culture of caution in health research. When in doubt about identifiability, for example, researchers and their institutions tend to treat the data as identifiable.

#### 3.1 Scope and Identifiability

The concept of identifiability determines what falls within the scope of data privacy laws. Identifiability generally means “the potential of data to enable identification of a person.”\textsuperscript{299} Health researchers can use and share non-identifiable data to avoid the application of data

\textsuperscript{297} “Cambridge Biorepository for Translational Medicine”, online: <https://www.cbtm.group.cam.ac.uk/>.


privacy laws. Unfortunately, it is difficult to determine when data are non-identifiable. When in doubt, researchers tend to cautiously treat data as identifiable. They then have to limit the purposes of processing, and keep data confidential and secure. Uncertainty and international variation begins with the definition of identifiability and related terms, aptly described by Bartha Knoppers and Mark Phillips as the discombobulation of de-identification.300 It is uncertain, for example, how easy or practicable identifying an individual must be for data to be personal.301 Across Canada, some personal information protection laws do not define personal information, while others define identifiability with reference to standards such as “readily ascertainable” or “reasonably foreseeable”.302 The EU GDPR states that “account should be taken of all the means reasonably likely to be used…”303 The standard of “reasonably likely” refers to both difficulty and probability, considering objective factors including costs, time, and available technology.304

If identifying an individual from data is only likely in combination with some auxiliary information, are they personal? Many data privacy laws, including the GDPR, explicitly include indirectly identifiable information in the definition of personal data. This complicates identifiability assessment as one must consider what kinds of auxiliary information are potentially available to allow for indirect identification.305 This leads into a second question of perspective: who is the legally relevant agent of identification? Is it just the data controller? Or could it be any person, including those using illegal means to access the data, such as hackers?306

Identifiability is also a moving target over time. As more and more personal data are made public and computing power and statistical techniques improve, “there is an increasing

301 Bygrave, supra note 299 at 130.
302 Thorogood et al, supra note 282 at 201.
305 Bygrave, supra note 299 at 130.
306 Ibid at 133.
likelihood that the information could be aggregated, cross-referenced, and linked in order to re-identify previously de-identified records”.

The scope for health researchers to confidently avoid the scope of data privacy laws through de-identification narrows quickly as they generate or link richer individual-level datasets, share data with numerous collaborators, or make the data available over long periods of time. These legal uncertainties are difficult to resolve in health research contexts.

An ideal solution would be legislative or judicial clarity on these issues. Some legislators and courts have attempted to constrain the broad definition of personal data. Some laws only extend protection to data that can affect a person’s legal, economic, or social standing; or that can affect an individual’s personal, family, business, or professional life. Another way courts have constrained the scope is with regard to the user’s intent to identify. As health researchers primarily aim to produce generalizable knowledge, they do not intend to make decisions to directly affect individuals and are generally unlikely to do so. A legal safe harbor with reduced compliance burdens for research is another potential option, as long as certain steps are taken to reduce identifiability and to safeguard the resulting data.

Identifiability also tends to increase unpredictably as researchers link multiple datasets. Individual-level linkage of data occurs when different datatypes about the same individual are connected, such as genomic data and electronic health-record data. The resulting connected dataset may present a higher risk of re-identification than the source datasets. Datasets available for linkage may be unknown or may not even exist at the time the identifiability of a dataset is evaluated, complicating risk assessment. In the other direction, researchers may only be interested in analysing sub-sets or fragments of an otherwise identifiable dataset. How to determine if a sub-set is identifiable? Consider an example from genetics. An individual’s whole


308 Bygrave, supra note 299 at 136.

309 Ibid.

310 Stodden, supra note 91 at 123.
genome may be considered identifiable. But what about a single genetic variant? A common variant alone is shared by a large fraction of the population, while a rare variant is usually defined as being seen in less than one percent of the population, and may be unique to an individual or family. What about the combination of 5, 10, 50, or 100 common variants? Statistically, even a limited combination of common variants will constitute a unique individual genetic bar-code.

Other privacy terms are even more unclear. Consider the terms de-identified and anonymized. Do these terms describe a process applied to data or the nature of the data? Used as verbs, these terms generally refer to the process of removing identifiers. Do the terms also imply a specific success state? In other words, does de-identification mean the process of removing identifiers so that the resulting data are no longer identifiable? One often encounters compound terms such as “genuinely de-identified” or “completely anonymized”. Use of such adjectives emphasizes, perhaps redundantly, a success state. De-identified and anonymized are also used as adjectives, legally synonymous to anonymous or non-identifiable. Canadian health research ethics guidelines distinguish between directly identifiable (e.g., name), indirectly identifiable (reasonably expected to identify an individual in combination with other indirect identifiers), coded (direct identifiers are separated from the information and replaced with a code), anonymized (direct identifiers are irrevocably stripped), and anonymous information (the dataset never contained identifiers).

Note the definitions of coded and anonymized only refer to stripping direct identifiers, suggesting the definitions are more focused on a process than a success state.

Perhaps the greatest confusion is over pseudonymized data, a new category of personal data introduced and regulated by the GDPR, albeit with a reduced compliance burden. Pseudonymized is related to the term coded. Coding can be defined as a process where identifiers are separated from research data and replaced with a unique code. Coding can also refer to a

311 Phillips & Knoppers, supra note 300 at 1103.
312 TCPS2, supra note 192 ch 5A.
313 General Data Protection Regulation, supra note 39 art 4(5).
success state where data “can only be re-identified with access to a deliberately crafted re-
identification mechanism.”314 According to Miranda Mourby et al., the GDPR definition appears
to encompass both the technical measure applied and the nature of the resulting data:
“pseudonymisation requires not just a process but an ultimate ‘success state’, in which the data
cannot be attributed to an individual without the use of additional information” i.e., the key-
code.315

Things become complicated when pseudonymized data are shared with a third party – a
common occurrence in collaborative health research. For example, a hospital may provide a
researcher access to pseudonymized health data. Is the researcher subject to the GDPR?
Assuming a success state was achieved, the data are non-identifiable relative to the researcher.
But the hospital still deliberately retains the code. According to Mourby et al., where personal
data are pseudonymized within an organization, it is still possible that the data may be
considered anonymized vis-a-vis an external researcher.316 The relevant factor is the relationship
between the organization and the researcher. Given the relationship, is it reasonably likely the
researcher could re-identify the individual by gaining access to the re-identification mechanism?
This interpretation is supported by the US-EU Safe Harbor Agreement, which exempted from its
rules US companies receiving pseudonymized medical research data, even where the key
enabling re-identification was still held by a person or organization in Europe.317 While this
agreement was under the former Directive, and while the Safe Harbor was struck down by the
European Union Court of Justice (CJEU), it suggests it is reasonable to consider pseudonymized
data as anonymous in the hands of researchers without access to the key-code.

314 Phillips & Knoppers, supra note 300 at 1103 [emphasis added].
315 Miranda Mourby et al, “Are ‘Pseudonymised’ Data Always Personal Data? Implications of the GDPR for
316 Ibid.
Parliament and of the Council on the adequacy of the protection provided by the safe harbour privacy principles
and related frequently asked questions issued by the US Department of Commerce 2000, OJ L 215, 2582000
and of the Council on the adequacy of the protection provided by the safe harbour privacy principles and related
frequently asked questions issued by the US Department of Commerce] at 520; See also Bygrave, supra note 299 at
133.
This interpretation is also supported by reference to a judgment of the CJEU in *Breyer v Bundesrepublik Deutschland*. This case concerned partial data (an IP address) rather than pseudonymized data. A public authority had access to an IP address but could not associate that IP address with an individual without additional information held by an internet service provider (ISP). However, the public authority could obtain that additional information from the ISP through legal means, so the IP address could be considered identifiable relative to the public authority.

A similar issue arises for encrypted personal data. Such data may be meaningless to a third party who does not hold the encryption key. An important aspect of determining if the encrypted data are personal data is the relationship between the party holding the meaningless encrypted data only, and the party holding the encryption key. It is common sense that researchers should not have a greater compliance burden simply because a third party holds a key-code. The wider the distribution of data, however, the greater the need to protect the key-code.

The Human Cell Atlas was originally proposed as a publicly accessible portal for sharing gene expression data. With the coming into force of the GDPR, the question arose: are gene expression data identifiable? Gene expression data are often expressed in terms of counts or quantities of the number of RNA molecules being expressed by a particular gene in a particular cell at a particular time. Counts alone would not appear to be identifiable. Particular experiment, however, may collect metadata describing participants’ clinical characteristics or genetic make-up. Whether an individual is healthy or sick, and whether an individual has particular genetic variants or not, provide important context for understanding gene expression. A privacy-protective alternative is to map gene expression to a reference genome rather than an individual genome, though this may reduce some relevant scientific information.

---

318 *Case C-582/14 Patrick Breyer v Bundesrepublik Deutschland [2016] ECLI: EU: C: 2016:779.*

A second challenge is that gene expression data are generated from raw RNA sequencing reads. Human Cell Atlas researchers also want to share raw RNA sequence data to improve the quality of new RNA sequencing platforms and analysis algorithms. But raw RNA sequence data can be used to infer the individual’s genetic sequence data, and is therefore potentially identifiable. Overall, the identifiability of RNA sequencing experiment data is highly dependent on context – the underlying technology used, the metadata and genetic reference map employed, and how much data (quantifications only or also raw sequence data) are shared. Such uncertainty has delayed the launch of the Human Cell Atlas, which is now looking to develop a GDPR compliance programme for its data portal, and a managed access tier. Additional uncertainty arises where data pseudonymized at a local research site are submitted to the HCA and then released to the community. Are these data still considered pseudonymized vis-à-vis the portal and accessing researchers? If so, what are their compliance obligations? As an international database, the HCA must also contend with different national laws and interpretations. This results in a quandary for designing consortium governance. More permissive sharing will facilitate use but potentially exclude contributors subject to strict interpretations. More restrictive sharing will facilitate contribution from various sites, but will limit access and use.

### 3.2 Who is the Researcher Responsible for Data Privacy?

The exchange of data between researchers does not fit neatly in the European data protection categories of “controller” and “processor”. Controllers are entities that determine the purposes and means of processing personal data. Processors are entities that process personal data on behalf of a controllers. Two entities can be joint controllers where they both have a role in determining purposes and means of processing. European data protection law applies directly to both controllers and processors, though processors are subject to narrower requirements. This structure does not align well with the way health researchers collaborate. Researchers receiving data may be leading their own research project, or may be collaborating on equal footing with the researchers who collected the data.

The interoperability challenge is not simply to determine what laws and obligations apply, but also to determine who – across complex networks of repositories and researchers – is accountable for fulfilling those obligations. Essentially, this is a coordination issue. A failure to coordinate can result in gaps in protection, or costly duplication. Consider, for example, the Data
Coordination Platform established as part of the HCA. The Platform is designed to ingest information from gene expression experiments, to clean and process that information, and to allow researchers around the world to access and visualize that data.\textsuperscript{320} In part in the spirit of international collaboration, in part for technical reasons, the Data Coordination Platform was established at four different host institutions: The European Bioinformatics Institute, the Chan Zuckerberg Initiative, the University of California Santa Cruz, and the Broad Institute.

Under the GDPR, these four institutes are probably considered “joint controllers” of the personal data collected from participants in gene expression studies and made available on the Platform.\textsuperscript{321} This is because they “jointly determine the purpose and means of processing”.\textsuperscript{322} Joint controllers must determine their respective responsibilities in a specific agreement, and any controller can be held liable by a data subject.\textsuperscript{323} This may not be a natural set-up, however, as the recipient researcher may have limited ability to interact with the data subject. Determining the division of responsibilities is not straightforward, and even a properly tailored agreement carries the risk that it may not be fully respected by the other party.

An important design question for international health research repositories is whether they should be structured as central repositories or as conduits to a distributed network of datasets “controlled” by individual researchers or research institutions. Legal control can remain distributed even where the data are technically centralized in a cloud-based repository. It is getting harder and harder for international repositories – both legally and practically – to accept responsibility for personal data that is collectively exploited by a community of researchers. In this sense, the HCA’s joint controllership model is an exception. International collaborations with limited organizational expertise or infrastructure may not be able to take on such a role.

\textsuperscript{320} Regev et al, supra note 294 at 42.


\textsuperscript{323} Ibid at 106, 246.
Where centralization is precluded, the aggregate of local compliance costs is likely to be higher, as each researcher sharing data has controller obligations. Negotiation costs will increase exponentially as users seek to access multiple datasets.

The issue of responsibility must be considered not only for the repository hosting the data, but also for the researchers around the world accessing the repository. From an IP perspective, the Atlas would like contributing researchers to waive all rights in their data except for attribution through standard licenses. It may be unclear if the researcher or institution has rights to waive. From a data protection perspective, researchers seeking to access data may need to agree to data protection obligations. Can researchers themselves enter into an online contract, or must their institutions sign data access agreements? From an interoperability perspective, it is clearly better if the researchers had autonomy to enter into agreements, without having to additionally negotiate with their institutions.

A final issue is the unclear relationship between researchers and their institutions. The GDPR categories focus on organizations – namely governments and commercial companies. It assumes organizations establish the purposes of processing, ensure appropriate security safeguards, and ensure appropriate contractual protections are in place where data are exchanged with other institutions. In health research, however, it is often the individual investigators who design research projects, and who exchange research data with each other. There is a motivational schism between the entity determining the purposes of processing, and the entity ensuring this is done in compliance with regulations. This imperfect employment relationship can lead to tensions between the institution and the researcher. The most obvious example of this issue is with data transfer agreements, which can include security and data protection obligations. There is often an insistence by data repositories that institutions and not investigators sign these agreements, leading to negotiations and delays. This compounds with IP licensing issues, discussed in Chapter 4, stemming from uncertainty over who owns and has authority to permissively license research outputs: researchers or their institutions?

### 3.3 Exceptional Treatment of Research Purposes

The GDPR does give scientific research purposes special status. Research is not defined but is to be interpreted broadly to include privately funded research (rec 159). Research is an
activity aiming at producing generalizable results, not to make decisions about individuals (with the exception of incidental findings). It is the results of research that may be used to inform individual patient care. Research purposes have exceptional status in the GDPR. Recital 55 states that further processing for statistical or scientific research purposes should be considered compatible with the original purpose. This must be read in conjunction with Recital 146 and Article 89, which require further processing for scientific research to be subject to appropriate technical and organizational safeguards, which an emphasis on the data minimization principle and anonymization or pseudonymization.

Under the GDPR, organizations generally need a legal basis to process personal information, such as consent (art 6). Processing of special categories of data, including genetic and health data, is prohibited by default (art 9(1)). Two exceptions are the explicit consent of the individual (stricter than consent as a legal basis for processing), and where processing is necessary for research purposes (arts 9(2)(a),(j)). The processing must also be proportionate and subject to safeguards, where provided by Member State law.

The GDPR does recognize that research consent cannot always be fully specified in advance. This recognition aligns with the trend in international and national health research guidelines to recognize broad consent, which is consent to research not fully specified in advance but subject to ongoing governance (see Chapter 6). There is an ongoing interpretive debate about how broad such a consent may be under the GDPR: must the consent be limited to certain areas of health research? Some confusion has resulted because the legislator adopted the word “specified” purpose for special categories of data (arguably including genetic data), but adopted the word “specific” purpose for personal data. The Article 29 Working Party interpretation is

324 General Data Protection Regulation, supra note 39 rec 33 (“It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose.”).


326 Dove, supra note 276 at 1022.
that broad forms of consent are not permitted for special categories of data. Should individuals be offered choices or preferences about the scope of the interoperability of their data? On the one hand, the GDPR is concerned about data grabs, so it places restrictions on unnecessary “tying” of the consent to data processing to the choice to receive a particular good or service. On the other hand, giving participants full control over their data amounts to giving them a “right to participate” even if they reject legitimate conditions of openness.

The GDPR also reduces compliance obligations of organizations who process data for research purposes (with safeguards in place), such as limiting the individual rights of data subjects to access, object, restrict processing, or rectification (art 89). While the GDPR is a regulation with direct application aiming to create a digital single market across Europe, Member States are given flexibility with regards to research provisions. For example, the reduced compliance burden for research vis-à-vis individual rights is up to the discretion of the Member State.

Arguably, the challenge for research is not the GDPR, but the potential for restrictive or incompatible national research provisions. Ireland, for example, adopted a health research regulation under its Data Protection Act 2018 that states all data processed in health research can only be carried out with the data subject’s explicit consent. There is an exception for new research of public importance outweighing the interest of getting the individual’s consent. But this public interest basis does not apply for previously collected health research data and samples. Any research without a GDPR compliant consent must technically grind to a halt when the regulations came into force. By basically equating the research exception to the public interest exception, the Irish approach arguably renders the separate GDPR research exception useless.

---

327 European Commission - Article 29 Working Party, supra note 325 at 18–19.
4 Interoperability Between Data Privacy Laws

The purposes of data privacy law include both ensuring appropriate levels of privacy protection for individual data subjects, and establishing optimal levels of interoperability within and between jurisdictions – in terms of the flow of personal data – in the public, commercial, and health sectors. The GDPR for example promotes the cross-border flow of data between EU Member States, though it does restrict flows outside the EU unless adequate and demonstrated protections are in place. Three challenges of legal interoperability between legal systems stymie international health research. First, variation across jurisdictions is problematic because it leads to uncertainty. When in doubt about jurisdiction or rules because research data have been combined or transferred across borders, researchers are likely to respect the rules of jurisdictions with the strictest potentially applicable law.

A second related issue is hyper-territorial application of data privacy laws, where two countries assert jurisdiction over the same data processing activity, which can result in conflicting requirements. A third problem is data sovereignty, where states adopt strategies like data localization requirements to ensure data remains subject to local laws. Data sovereignty directly restricts the global transfers essential for health research collaboration. Concerns over transborder flows were highlighted in the original OECD data protection guidelines, which encouraged its Member countries to consider the implications of their laws on other countries, ensure the uninterrupted and secure transborder flow of data, to generally restrain from restricting transborder flows, especially disproportionate restrictions.329 I turn to a review the three problems of legal variation, hyper-territoriality, and data localization, followed by a review of strategies that can be employed to enhance cross-border legal interoperability.

4.1 Variation between jurisdictions

There is both uncertainty and disagreement between jurisdictions over the aims of data privacy law. While safeguarding privacy is a clear aim, data privacy laws may also explicitly or implicitly protect other individual interests in personal integrity and autonomy, as well as the

329 Organisation for Economic Co-operation and Development (OECD), supra note 272 arts 15-18.
interests of controllers (e.g., companies and government entities) or even nations (as a form of economic protectionism). There is fundamental disagreement across the Atlantic over the nature and role of data privacy law, and the balance it should strike between important societal interests such as promoting innovation and efficiency or ensuring public safety through government surveillance, and protecting individual privacy. Ian Marsden and Christopher Brown distinguish between two general regulatory approaches, those based on human rights, and those based on economic efficiency (e.g., competitiveness). Both of these approaches can be identified at the heart of data privacy law, though different jurisdictions put different emphasis on one or the other.

In the European Union, privacy and data protection rights are found in the European Constitution (Charter), supranational treaties (Convention), and in numerous national constitutions. They are described in the language and structure of human rights. Article 8 of the European Convention on Human Rights, an important supra-national treaty, protects the right to privacy, a qualified right against interference by a public authority. Article 8 of the EU Charter of Fundamental Human Rights, part of the European constitution, recognizes data protection as a fundamental right: “[Personal] data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law”. It also recognizes individual rights to access and rectification.

In terms of economic efficiency, data protection law can be described as having two aims. The first is to ensure the flows of personal data across borders necessary for the exchange of goods and services. The second is to maintain consumer confidence (particularly in e-

---

330 Bygrave, supra note 299 at 118–124.

331 Ian Brown & Christopher T Marsden, Regulating Code: Good Governance and Better Regulation in the Information Age (MIT Press, 2013) at 12.


commerce where there are concerns about misuse of personal data and fraud). The US approach which depends heavily on industry self-regulation and consumer privacy, emphasizes the importance of efficiency goals. In other words, the EU treats data protection as a human right – part of a project to construct a European identity – whereas the US approaches data privacy within a “marketplace discourse”, protecting consumers in a data-driven market economy. The European approach is more collectivist, placing important limits on forms of private-ordering – contract and consent – relating to personal data, as well as limits on “tying” the collection of personal data not necessary to offering a good or service.

The fundamental schism between US and European regulatory approaches has become more apparent with the adoption of the EU GDPR. This difference in philosophy leads to a different default treatment of data processing. According to Edward Dove, in “the US … the starting presumption in law is that processing personal data is lawful unless it is expressly forbidden, in Europe, processing personal data is prohibited unless there is a lawful basis that permits it.” European-style data protection law is clearly concerned with encouraging the appropriate flow and use of personal data. From a political perspective, data protection law is calibrated to constrain the excessive interoperability of commercial tech giants and US government surveillance. The Canadian approach is somewhere between the two, with consent generally required for processing of personal information, but without the emphasis on explicit consent, the prohibition on processing special categories of data, and the serious regulatory fines.

Industry stakeholders in the United States have criticized the European approach as protectionist, and highlight the dangers of restrictive data protection rules for innovation and “consumer welfare” in digital services. Alan McQuinn and Daniel Castro highlight that a number of values are at stake in addition to consumer privacy, including free speech,
productivity, competitiveness, and innovation. Consumer welfare includes privacy, but also involves lower prices and the development of new services and products. Again, consumer confidence that their privacy will be respected affects their ability to engage with these new services.

Arguing for innovation, strict data protection rules (such as general opt-in consent requirements, purpose specification, the right to deletion, data minimization, limits on data retention, and barriers to international transfers) mean less data are available to develop or improve services. Positive obligations like privacy by design significantly increase costs of compliance and impact R&D budgets, costs that are passed on to the consumer. Privacy by design emphasizes embedding privacy as a priority in all aspects of organizational governance. It involves organizational processes geared to protecting privacy, including education and training of technical staff, privacy impact assessments, identifying personnel responsible for privacy, audits, and breach handling/reporting.

GDPR compliance is an expensive proposition. It may limit competition between small to medium sized business and established multinationals who can afford the additional compliance costs. Likewise, the compliance burden of the GDPR may disfavor academic health research collaborations with limited funding against industry ones. Data protection laws can be barriers to trade in (digital) goods and services. Clauses enabling cross-border flows of personal data are increasingly found in trade agreements, which have long been forums for addressing international interoperability between legal systems (discussed below).

4.2 Extra-territoriality and Jurisdictional Conflicts

Extra-territorial application is a particular concern with data privacy law, because the territorial-nexus connecting a jurisdiction to a data processing activity is often unclear, especially on the Internet. The geographic location of the firm or entity controlling the data, the location of the individual to whom the data relates, and the location where the data are actually processed can all differ. A legal presence or significant level of business activity in a jurisdiction will
usually be enough to attract regulation.\textsuperscript{339} There is some case law in Canada suggesting that Canadian personal information laws apply where foreign websites or companies collect personal information about Canadians\textsuperscript{340}, and where harm or inconvenience is experienced in Canada.\textsuperscript{341}

The territorial scope of the GDPR “follows the data that it protects and therefore has direct bearing on the activities of organizations based in countries around the world.”\textsuperscript{342} The GDPR applies to data controllers and data processors based in the EU, regardless of where the processing takes place. It also applies to controllers and processors not based in the EU that offer goods and services to Europeans, or that monitor their behavior taking place in Europe.\textsuperscript{343} Extra-territorial application can lead to situations where multiple data privacy laws simultaneously apply. This is particularly a problem for research databases comprised of data derived from multiple jurisdictions, such as the Human Cell Atlas.

4.3 Data Sovereignty

An additional strategy to address global privacy is through data sovereignty laws. These laws seek to create facts on the ground that preclude the need to seek extra-territorial application. The GDPR continues the European approach to data sovereignty of restricting by default cross-border transfers of personal data outside the European Union. Two important bases for transferring data to third countries (outside the EU) include an adequacy decision – where the European Commission determines that a third country offers comparable privacy protections (art 45) – and the adoption of appropriate safeguards to govern the transfer (art 46-47). Consent is recognized as a derogation from specific cross-border transfer restrictions. The informed consent

\textsuperscript{339} Ibid at 50.


\textsuperscript{341} Zuckerman c. Target Corporation, 2015 QCCA 1809; St-Arnaud v. Facebook Inc., 2011 QCCS 1506. See also Ibid s. 8.17.

\textsuperscript{342} Dove, supra note 276 at 1014.

\textsuperscript{343} General Data Protection Regulation, supra note 39 art 3.
standard is that the individual must be informed of “possible risks of such transfers … due to the absence of an adequacy decision and appropriate safeguards” (art 49(1)).

Safeguard measures are typically subject to oversight by European supervisory authorities. Requirements for foreign controllers and processors to incorporate a local subsidiary, or to have local “representatives” may also be considered a light form of data sovereignty.\(^{344}\) Two Canadian provinces – Nova Scotia and British Columbia – have adopted data localization rules under their public sector personal information laws.\(^{345}\) Under these laws, personal information cannot be transferred outside Canada except with the explicit consent of the individual or under other narrow exceptions.

An alternative is to treat cross-border transfers as a matter of private-ordering. Canada’s federal private sector law PIPEDA does not explicitly address international transfers. Canada’s federal privacy regulator, the Office of the Privacy Commissioner of Canada (OPC), has held a policy position that international transfers are governed like any other transfer between organizations, and are subject to the accountability principle. This principle holds Canadian organizations accountable for the decision to make transfers, and for ensuring appropriate contractual or other safeguards are in place when transfers are made. The OPC recently released a draft of a revised Policy on Transborder Dataflows for consultation.\(^{346}\) This consultation emphasizes a general obligation of commercial organizations to seek individual consent when transferring personal information outside of Canada.

The OPC’s rationale is that “cross-border data flows are not only matters decided by states (trade agreements and laws) and organizations (commercial agreements); individuals ought to and do, under PIPEDA, have a say in whether their personal information will be disclosed

\(^{344}\) Chapelle & Fehlinger, *supra* note 1.


\(^{346}\) Canada, *supra* note 287.
outside Canada.” This is because individuals “reasonably expect to be notified if their information was to be disclosed outside of Canada and be subject to the legal regime of another country.” The policy is nuanced by recognition that there are exceptions to consent requirements, and that consent may be implied depending on the sensitivity of the data and reasonable expectations. Setting aside the legitimacy and binding quality of regulators’ interpretations of the law, it is arguable that a general consent-by-default requirement for cross-border transfers amounts to a data localization requirement, which could have ramifications under Canada’s trade agreements.

Essentially, the OPC position can be thought of as a labelling requirement: “this service involves cross-border transfers”. The likely result of this guidance being finalized is simply that privacy policies will become more detailed vis-a-vis cross-border transfers. Uncertainty remains over the standard of informed consent for international transfers. How detailed should the consent be about the destination of transfers? What risks of transferring personal information to third countries will be highlighted? There is a small market risk to interoperability that Canadians will vote with their feet and only use services that keep data in Canada. Introducing a new consent requirement is also transitionally problematic for previously collected data, especially where there is not an ongoing relationship with the individual. Ultimately, a labelling requirement – as long as the standard for informed consent is clear – will probably be largely inconsequential on business practices, and may even improve interoperability by removing uncertainty over individual consent. PIPEDA is also a private-sector statute applicable to commercial organisations, so many healthcare and research organizations would be unaffected, unless perhaps if they rely on commercial cloud service providers.

Application to the HCA

In summary, legal fragmentation between jurisdictions, extra-territorial application of laws, and data sovereignty strategies all pose serious challenges to the interoperability of global health research. This can be particularly true for open science approaches, like the efforts of the

347 Ibid.
348 Ibid.
HCA to establish a global reference map of gene expression across different human cell types. The rich bioinformatics data collected in gene expression studies, especially where coupled with rich phenotypic data, is unique to the individual, and potentially re-identifiable. Under the GDPR, such data may be considered personal data. Moreover, open science collaborations like the HCA aggregate research data from many countries, and then re-distribute the data to the international research community over the public Internet, or over secure, web-accessible platforms. The risk is that the laws of numerous jurisdictions may apply simultaneously to the activities of the HCA Data Portal as well as to researchers accessing HCA data.

Extra-territorial application of data privacy laws may be justified based on the location of the participants, the location of the repository or repositories, or the location of researchers accessing and using data. In the case of the HCA, the database will likely be mirrored for technical reasons across numerous organizations and countries. Extra-territorial application can lead to an accumulation of positive obligations on the HCA joint controllers, such as privacy by design requirements to carry out privacy impact assessments or to hire a data protection officer. Legal and operational costs accumulate as the number of jurisdictions imposing distinct positive obligations increases.

GDPR data localization requirements present an important issue for the HCA when looking to distribute European data to its multiple, mirrored repositories, and to the international research community. Prospectively, European researchers contributing to the HCA may not be able to overcome these barriers through consents, contracts, and safeguards. Retrospectively, researchers would need to voluntarily address all of these challenges of international transfer at the time of research design and recruitment. Where this is not done, or not done properly, researchers may need to seek approvals from privacy authorities, research ethics committees, and/or individual participants to release data.

Where it is possible to aggregate data in a central repository, both the repository and researchers accessing data from it may need to comply with the laws of multiple jurisdictions, which may vary and potentially conflict. In response, the HCA must either establish standard limits on access and use to correspond to the strictest jurisdiction, or establish tracking systems
to ensure dataset-specific limitations are respected. A technological work-around is to design research networks in a manner enabling individual-level data to be retained and protected locally, while still enabling analyses of virtual cohorts. These strategies are, however, still under development and may involve significant technical costs and limits on what users can achieve. Technological approaches, and their legal ramifications, are discussed in Chapter 8.

5 Strategies to Improve Interoperability between Data Privacy Regimes

This section reviews three strategies to promote the interoperability of data privacy regimes, including harmonization through international treaties, unilateral legislative approaches like adequacy requirements, and developing sector-specific data privacy codes of conduct.

5.1 International treaties

International treaties can be used to harmonize data privacy rules from the top down. Harmonization is “the legal model for institutionalizing a desired cooperation by confining actors and policies into the normative corset of [international] rights and obligations.” In 2013, the International Conference of Data Protection and Privacy Commissioners released a Resolution on Anchoring Data Protection in International Law. The Resolution argues for extending the human right to privacy under Article 17 of the UN International Covenant on Civil and Political Rights (ICCPR) (1966), to include a binding international human right to data protection.

---

349 As an example of such a system see J Patrick Woolley et al, “Responsible Sharing of Biomedical Data and Biospecimens via the ‘Automatable Discovery and Access Matrix’ (ADA-M)” (2018) 3:1 NPJ Genomic Medicine 17.

350 GA4GH, supra note 3.

351 Weber, supra note 33 at 8.

352 International Conference of Data Protection and Privacy Commissioners, Resolution on Anchoring Data Protection and the Protection of Privacy in International Law (2013).

Resolution finds support in *General Comment No. 16* on the ICCPR, which highlights the importance of regulating collection and storage of personal information by public and private bodies, as well as the importance of safeguards against unauthorized access.\(^\text{354}\)

An alternative to a UN treaty is non-member accession to the Council of Europe’s Convention 108, but this comes with strong regional bias.\(^\text{355}\) There is already widespread international agreement on OECD-type minimum baseline data protection, with flexibility for states to adopt higher levels of protection at their discretion. The US remains a notable outlier, surprising considering its international regulatory influence in other areas, but unsurprising giving its lead in the data-driven economy.\(^\text{356}\) Given fundamental ideological and cultural disagreement over how to regulate data privacy between the US and EU, further harmonization is unlikely in the near future.\(^\text{357}\) Significant efforts have been made *within* the European Union to harmonize data protection law across Member States – catalyzed by the 1995 Data Protection Directive\(^\text{358}\) – both to protect the fundamental rights of citizens and to promote the free flow of data between Member States. The GDPR takes this one step further as a regulation, which has direct application without having to first be transposed into national law.\(^\text{359}\) There does seem to be at least an initial willingness of private entities worldwide to comply with the GDPR, and of legislatures outside the EU to consider convergence with the GDPR.\(^\text{360}\) According to Lee Bygrave, the challenge of international harmonization is even greater today with a more diverse

---


\(^{355}\) Bygrave, *supra* note 299 at 205–6.

\(^{356}\) *Ibid.*

\(^{357}\) *Ibid.*


\(^{359}\) *General Data Protection Regulation, supra* note 39.

\(^{360}\) Canada, for example: House of Commons of Canada, *supra* note 284.
international community, more diverse stakeholders that must be engaged, and the lack of a clear international venue where a deal can be successfully brokered.\textsuperscript{361}

In contrast to an international treaty on data privacy, international trade agreements increasingly seek to remove or prevent barriers to the free flow of personal data relating to the exchange of goods and services. From the perspective of international trade in (digital) goods and services, barriers to the cross border flow of data are seen as protectionist. The United States insisted in the recent negotiations of the USMCA trade agreement that Canada agree not to impose barriers on the free flow of data, with a particular focus on electronic commerce and the financial sector (where Canadian banks seek to adopt US cloud service providers).\textsuperscript{362} Exceptions were provided for sensitive government data, which must reside on servers located in Canada. The free flow of data raises concerns for Canadian regulators, such as the Office of the Superintendent of Financial Institutions, about access to data held on foreign servers for regulatory purposes.

A dynamic balance between human rights (privacy) oriented international instruments and efficiency oriented trade agreements could theoretically be a formula for optimizing interoperability. States are not, however, currently enthusiastic about delegating even more sovereignty to international organizations, especially on matters that affect the competitiveness of their knowledge economies.\textsuperscript{363} International trade negotiations have also shifted from multi-to bi-lateral character, favoring more powerful nations who can extract tailored concessions from weaker trade partners and leading to greater fragmentation.\textsuperscript{364}

In addition to political will, there are important political limits to legal harmonization. The law on the books can be drafted in a very similar manner, but differences in how the law is applied in practice may be hard to identify and even harder to harmonize. This issue was

\textsuperscript{361} Bygrave, supra note 299 at 205–6.


reflected in the *Schrems* case that led to the invalidation of the US-EU Safe Harbour agreement, over concerns about mass surveillance practices by US law enforcement.\(^{365}\) Even a harmonization project such as the GDPR has important limitations. Despite being a Regulation, there is still significant scope for Member law derogations, especially with regards to scientific research. As Edward Dove argues, “it remains to be seen whether the new rules will be implemented across Europe in a harmonized way that delivers the clarity and certainty it promises, for researchers and research participants alike”.\(^{366}\)

### 5.2 Unilateral Legislative Approaches: Adequacy Ultimatums

The European GDPR continues to adopt an adequacy approach, which encourages other countries to adopt European-style data protection regimes in exchange for lowering barriers to cross-border transfers, and in turn greater access to European markets. Such approaches will only be successful if other countries have strong economic incentives to access European markets, and do not have major sovereignty-based reservations about harmonizing their laws towards European standards. Many countries will follow Europe’s lead in establishing strong data privacy protections, in order to retain access to its large market. Unilateral approaches, however, can breach what La Chappelle refers to as the Golden Rule of internet regulation: any “national policy measure that would be detrimental if generalized around the world should not be adopted in the first place.”\(^{367}\) Other countries may have legitimate reasons to prefer more business friendly approaches. In practice, adequacy approaches are slow, bureaucratic, expensive, and subject to uncertainty and realpolitik.

A more restrained approach favored in Canada is organizational accountability. Canada’s federal commercial sector law PIPEDA holds a custodian of personal information accountable for data transferred to third parties or foreign jurisdictions.\(^{368}\) The law is not designed to apply

---

\(^{365}\) *Maximilian Schrems v Data Protection Commissioner*, 2015 ECJ, Case C-362/14.

\(^{366}\) Dove, *supra* note 276 at 1014.

\(^{367}\) Chapelle & Fehlinger, *supra* note 1 at 5.

\(^{368}\) *Personal Information Protection and Electronic Documents Act, SC 2000, c 5, supra* note 286 Principle 4.1.
directly to foreign companies receiving these transfers. This approach encourages the custodian to ensure appropriate contractual or other safeguards are applied to data, or to not transfer data if there are concerns.

5.3 Mutual Recognition and Cooperation

One way to reduce intrusion on the sovereignty of other countries is through mutual assistance treaties. The US-EU Privacy Shield is an example, albeit a relatively asymmetric one in the EU’s favor, where US regulators agree to enforce European privacy protections against US companies. Despite differences in legal culture, the Privacy Shield may form the basis for transatlantic rapprochement through “coercion, persuasion, and acculturation”. At the very least, legislators drafting data privacy laws should be aware of the importance of interoperability between jurisdictions.

5.4 Codes of Conduct

The GDPR provides for sector-specific codes of conduct, approved by data protection authorities and monitored by an industry body. The codes must be approved by the European Commission, after review by the European Data Protection Board, a body composed of the heads of national Data Protection Authorities, which aims to ensure consistent application of the GDPR across Member States. Codes achieve flexibility by allowing rules to be tailored to the data processing practices of a particular industry. Moreover, rules tailored by industry consortia,


370 Schwartz & Peifer, supra note 335 at 121.

371 Chapter 2 discusses European Union’s legal interoperability framework, which acts as a guide prompting national lawmakers to consider compatibility of rules across borders.

372 General Data Protection Regulation, supra note 39 art 40.
rather than legislatures, can be adapted more rapidly in response to business or technological change (though the amendments would need regulatory approval).373

Codes include technically detailed rules that provide certainty for companies, as well as confidence for regulators and data subjects. Such detailed rules could address pseudonymization (the separation of identifying information from other data of interest to reduce risks to privacy while retaining a link to the identifiers with a unique code), data minimization (removing identifiable or sensitive information not necessary to achieve organizational goals), security standards including data breach reporting, consent models (opt-in v.s. opt-out), informed consent standards, the transparency of information provided to the public and to data subjects about processing, processes for fulfilling data subject rights (e.g., access and portability), and safeguards for cross-border transfer.374

Codes of conduct are voluntary accountability tools: they provide detailed rules against which the compliance of organizations can be demonstrated or challenged.375 Added accountability can improve the trust and confidence of data subjects. The hope is these codes offer cost-effective co-regulatory regimes involving sectoral monitoring bodies, reducing demands on government regulators.376 Practically, these bodies may be better resourced and more expert in regulating their sector than overburdened government regulatory authorities. But there is a need for safeguards. Monitoring bodies are accredited and must satisfy criteria including independence; freedom from conflicts of interest; expertise; sufficient policies, procedures, and resources; and transparency.377

373 Ibid art 40(5).
374 Ibid art 40(2).
375 European Data Protection Board, Guidelines 1/2019 on Codes of Conduct and Monitoring Bodies under Regulation 2016/679, Text at 6 (other tools include data protection impact assessments and accreditation).
376 Ibid at 8.
377 European Data Protection Board, supra note 375 s 12.
In terms of interoperability, transnational codes of conduct can act as a kind of interface between legal systems. This is important even given the status of the GDPR as a regulation. Harmonization gaps can still arise in how Member States apply the regulation. In technology contexts, interfaces are a common way to ensure interoperability between different systems. Application programming interfaces (APIs), for example, are a common means of allowing different systems to automatically communicate. The systems can be quite different as long as they agree to exchange data according to the structure of the API. A code of conduct could provide an interface that helps to bridge harmonization gaps between the laws of different jurisdictions. Industry could design the code to substantively comply with the laws of multiple jurisdictions. Regulatory bodies can then formally approve the code as compliant with each regime.

It would seem, however, that codes of conduct depend on the existence of substantially harmonized laws. Codes can provide detailed interpretations of legal standards, but they cannot re-write the law. They cannot resolve extensive fragmentation or fundamental policy conflicts between legal regimes. It is also likely that a transnational code of conduct would have to be designed to comply with the rules of the strictest jurisdiction, decreasing its attractiveness. Codes of conduct are also yet another layer of regulation, which could be seen as increasing the regulatory burden on organizations. New forms of fragmentation may also emerge across sectors. An entity trying to work across two sectors may need to comply with two distinct codes of conduct, which may not be compatible. And given the greater level of technical detail found in codes of conduct, policy conflicts may be more likely between codes than between laws.

Codes of conduct in the general sense are a form of standardization. Standards are specifications, guidelines, or requirements, usually developed through consensus of diverse stakeholders (e.g., industry, governments, and interest groups). Standardization is the process of developing and implementing a set of standards with the intention of increasing quality or the efficiency of a particular activity. They encourage similar data processing practices to be

\[^{378}\text{Ibid} \text{ at 4.}\]
adopted across organizations. In turn, they may encourage data exchange across organizations, as there is more confidence that data protection standards will be respected. What distinguishes GDPR codes of conduct from other, informal standards initiatives is that they are formally approved by legal authorities, and have legal effect. An interesting challenge for all standards initiatives is the governance: who participates in developing the standard. The GDPR lays out some formal criteria concerning the governance of codes of conduct. The “code owner” who develops and maintains the code must demonstrate it is an effective representative body of the applicable sector. Relevant stakeholders, including data subjects, should be consulted.\textsuperscript{380}

There is an effort underway to develop a code of conduct for health research by the Biobanking and Biomolecular Resources Research Infrastructure-European Research Infrastructure Consortium (BBMRI-ERIC).\textsuperscript{381} The effort includes a wide range of stakeholders from industry, patient groups, and health research institutions aiming to develop the code and submit it to the European Commission for approval. A particular challenge for developing a health research code is that despite the coming into force of the GDPR, there is still significant fragmentation of laws across the EU, especially relating to research data. This is a result of a number of factors: different speeds of implementing the GDPR into national rules, the fact the GDPR allows for various derogations in national law relating to research, and the existence of data-type and sector-specific (e.g., biobanking) national laws in some countries but not others.\textsuperscript{382} There are also political challenges, considering that cultural perspectives on data privacy differ across countries. The diverse organizations and stakeholder representatives the GDPR requires be involved in developing a code of conduct may also put emphasis on different values.

Moreover, it is unclear if codes of conduct will work for diverse sectors like health research. Consider the difference in data management practices between a straightforward

\begin{flushleft}
\textsuperscript{380} General Data Protection Regulation, supra note 39 Rec 99.


\textsuperscript{382} Anne Cambon-Thomsen, Emmanuelle Rial-Sebbag & Bartha Maria Knoppers, “Trends in Ethical and Legal Frameworks for the Use of Human Biobanks” (2007) 30:2 European Respiratory Journal 373.
\end{flushleft}
clinical trial and a longitudinal epidemiological study. There is likely more heterogeneity between the data governance practices of different health research disciplines than there are between two separate sectors. The sector adopting a code also needs to internalize the costs of regulation by establishing an industry self-regulatory entity. The confidence of regulators and data subjects in the code will largely depend on the nature, mandate, and resources of this entity. Given the diverse range of commercial, academic, and non-profit entities involved in health research, negotiating cost sharing could be problematic.

A further weakness of the BBMRI approach is exclusion of international stakeholders. Finding rules for research within Europe doesn’t solve the biggest problem posed by the GDPR: the conditions of cross-border transfers outside of Europe, essential for international collaboration in health research. Unless most countries agree on a code of conduct, there is also a greater risk of policy conflict in the code of conduct context, where detailed technical rules are outlined. There is a possibility for organizations outside Europe to agree to comply with a code to facilitate cross-border transfers, but it remains problematic that such organizations would not have contributed to the drafting of the code.

Interoperability between data privacy regimes can be enhanced through international treaties, mutual recognition, and binding codes of conduct. Any success with these strategies is likely to provide greater clarity for international health research collaborations. Unfortunately, data privacy law does not always clearly apply to health research, and is not always fit-for-purpose where it does. As a result, even where harmonization of data privacy law is successful, rules for research may remain unclear and fragmented.

6 Tools to Enhance the Interoperability of Research

Researchers can adopt self-help strategies and tools to navigate data privacy regimes and enhance the interoperability of research data. Legal tools relating to data privacy compliance include broad consents, data sharing contracts, and privacy and security safeguards. Admittedly, researchers in some jurisdictions face limits on their freedom to adopt more permissive consents, contracts, and safeguards. Europe in particular takes a more collectivist rather than individualist approach to data privacy. Moreover, standardizing tools cannot remedy deficient data privacy
laws or resolve fundamental conflicts between jurisdictions. Nevertheless, standardizing these tools can provide greater certainty and interoperability internationally.

**Consent**

Traditional strategies to comply with data privacy law in health research are to anonymize data or to seek consent from participants. The former traditionally worked well because research is primarily concerned with producing generalizable knowledge, rather than making decisions that directly affect the rights and interests of participants. Privacy safeguards can therefore be applied at early stages of the research process to obscure individuals and focus on populations. The anonymization pathway, however, is being closed off by bioinformatics (which involves multi-variate data on the scale of billions of data points) and Big Data (where multiple data types are linked at the individual-level to allow mining for patterns).

A consent challenge for health research interoperability is the data privacy principle of purpose specification, reinforced by the health research ethics principle of informed consent (discussed in Chapter 6). The scope of a researcher’s freedom to seek broad consents to unspecified research uses or cross-border transfers (where applicable) appears to be limited by the European GDPR. An American view, by contrast, is that researchers and participants should be legally allowed to agree to a greater range of interoperability through the consent process, as long as the consent is voluntary and informed. This position is less paternalistic (or collectivist) than the European position. This shifts the debate from what is allowed or not allowed, to what is informed and not informed. Given the uncertainty over the potential uses, benefits, and risks presented by widely available research data, it can still be hard to determine how to satisfy the legal standard of informed consent. Factors considered when assessing the standard of voluntary and informed consent may include the critical nature of the good or services, and the sensitivity of the data. Individuals have little choice but to give up personal information to receive critical goods and services. Sensitivity speaks to the potential for harm from a data breach.

---

The tension is this. Data privacy laws may circumscribe the processing of data to a defined purpose, specified at the time of collection. Open science by contrast emphasizes that data be used for any research purpose, emphasizing “the benefits that are possible through unforeseeable future uses”. On its face, then, data privacy law appears to present a limit on the level of interoperability that can be achieved in science. There is some nuance in the sense that data privacy law is more concerned with processing that directly affects the rights and interests of individuals. Indeed, the GDPR states that uses for purely statistical and scientific purposes should be considered compatible with the original use. In this way, the GDPR strikes a balance between privacy and innovative use of data. Researchers may in some jurisdictions be able to adopt consents covering a broader range of future research purposes to promote interoperability. In jurisdictions that treat data privacy as a matter of inalienable personality rights, however, this may only take interoperability so far.

Data sharing agreements

Data sharing agreements aiming to ensure data protection compliance are problematic because the conditions and restrictions they contain tend to accumulate as datasets are aggregated and shared internationally. Researchers accessing these data may then derive a modified dataset (perhaps as the basis of a publication). Researchers are contractually bound, however, by the restrictive conditions of the data sharing agreement. These agreements tend to emphasize confidentiality, privacy, and security, without emphasizing equally important aspects of reproducibility and openness. They limit researchers’ ability to share derived datasets as broadly as the repository was able to share the original dataset, even though such datasets typically have a similar risk profile to the original dataset, and the expectations of participants are unlikely to differ from the sharing of one dataset to the sharing of another.

Phillips & Knoppers, supra note 106 at 109.

General Data Protection Regulation, supra note 39 Rec. 50.
Under the GDPR, there are model consent clauses that allow exchange of data between controllers outside the EU.\textsuperscript{386} While these standard contracts were originally limited to joint and several liability, a due diligence regime was later added, where “the data exporter and the data importer would be liable vis-à-vis the data subjects for their respective breach of their contractual obligations” and “the data exporter is also liable for not using reasonable efforts to determine that the data importer is able to satisfy its legal obligations.”\textsuperscript{387}

\textit{Privacy and Security Safeguards}

Data privacy law does not aim only to place barriers on data exchange. Rather, it aims to ensure that data exchange does not diminish protection. Naturally, if networks of researchers (and research institutions) can establish, adopt, and ensure adherence to privacy and security standards, then data should be able to flow more easily between them. Unfortunately, organizations may be unwilling or unable to meet standards if they involve an excessive compliance burden. Data access governance (see Chapter 7) can provide an interface between two organizations or contexts to ensure they have comparable privacy and security standards through audit and agreement.

\textit{Problems with Private-Ordering}

The problem of voluntariness is stark in health research contexts. Health researchers are often left to voluntarily adopt consent models and data sharing agreements that allow for desirable forms of data exchange and collaboration, but that comply with data privacy obligations. They are also often left to voluntarily balance privacy protections such as the stripping of identifiers, with scientific utility. First, there is an incentives problem. Individual researchers may be focused on completing their particular research study, and may not pay significant attention at the time of protocol and consent form design to the eventual interoperability of the data. They may also be reluctant to share their data because they are


\textsuperscript{387} Ibid at preamble para 5.
concerned about competition with other researchers who may scoop their findings, about misuse of the data to undermine their findings, about breaching legal or ethical obligations to protect participant privacy, or they may simply see no direct benefit accruing to them from sharing.

A second related problem is that researchers have imperfect authority to act as stewards over data. This authority is often shared, to an uncertain degree, with their institutions. Researchers must generally seek approval from a research ethics committee before starting their project. Research ethics committees often scrutinize data sharing plans as presenting potential privacy risks to participants. Researchers may also be required to work with their technology transfer office of institutional legal counsel, and are constrained by institutional policies, processes, and forms. The path of least resistance for researchers may simply be to limit sharing. Many public and philanthropic funders have data sharing policies or mandates, as do journals, but it is really the institutions that need to get behind researchers to ensure all strategies are adopted to share within the bounds of data privacy law.

Researchers generating data from banked samples may have little say in the design of participant consents. One challenge faced by the Human Cell Atlas is that gene expression research is heavily reliant on access to tissue and organ banks. Groups of cells must be destroyed to measure the (average) levels of RNA expression in the cell. Extracting such tissue directly from living participants would be highly invasive. Common sources of tissue for gene expression research are left-over tissue from surgery, unused transplant organs, and the recently deceased. These samples are often first collected by a biobank, which seeks an appropriate consent to sample sharing and related research, and imposes a material transfer agreement on accessing researchers limiting onward transfer of data derived from the samples. Tissue access agreements often restrict the extent to which researchers can share data generated from tissue. In turn, the Atlas cannot simply ask gene expression researchers to seek consent to sharing from participants, even prospectively. The researchers will not always have a direct relationship with the participants. They will be bound by the data sharing terms offered by the biobank. An additional challenge, especially for organ and pathology samples, is that re-consent cannot be obtained from deceased individuals. In seeking to ensure standards of confidentiality and security when samples are shared, biobanks can unintentionally restrict the openness and availability of derived data.
In Chapter 4, research license interoperability was defined in part as the ability to use data without case-by-case negotiation with the rights-holder. Restrictions and negotiation costs are potentially greater under data privacy law than copyright law. Instead of negotiating with a handful of data owners, researchers may be legally required to negotiate both specific research uses and specific cross-border transfers with potentially tens, hundreds, or thousands of individual research participants. Just as there are project-level datasets orphaned under copyright law – where the owner cannot be found or contacted – there are likely to be many orphaned individual-level datasets under data privacy law – where the individual cannot be found or fails to respond to re-consent attempts.
## Chapter 6
### Interoperability of Health Research Regulation

**Key Points:**

- Health research regulation gives researchers, research ethics committees (RECs), participants, and participant communities key roles in the governance of research. There is a conceptual muddle over if and how these roles translate into control over research data.
- Interoperability is undermined where researchers can only request specific permissions to share or re-use data from participants, RECs, or communities. This is illustrated by barriers encountered by the Canadian Open Neuroscience Platform in seeking approvals to make neuroimaging data more widely available.
- Canadian research ethics guidelines require informed consent to research as a general rule, but give RECs discretion to modify or waive consent requirements under certain conditions. US regulations explicitly enable broad consents. EU human rights and data privacy laws tend to limit the scope of broad consent. Cultural and structural differences between these norms may limit further international harmonization.
- The likely result of multi-party control over research data exchange and re-use is compounding restrictiveness and transaction costs.
- The ethical tension between openness and individual participant autonomy can be resolved to some degree by seeking consent to ongoing governance.
- Competing principles of openness and community control over data can be reconciled by engaging communities not in the governance of datasets, but in the governance of health research systems.

In this Chapter, I argue that health research regulation should allow and encourage researchers, individual participants, and communities to grant broader permissions for the exchange and re-use of research data. This shows appropriate deference to the freedom and creativity of scientists. Regulation should also encourage parties controlling research data to express these permissions in standard ways. Both strategies promise to improve researcher connectivity and facilitate collaboration. Health research regulation is not a formal term of art, but rather … the general ecosystem of activities, laws and
regulations that seek to shape the conduct of any and all types of health-related research involving human participants directly, or indirectly using data and biomaterials from them.388

It includes laws, regulations and guidelines – the latter often rendered binding by funding agencies or drug regulators – that promote ethical conduct and scientific integrity through principles of weighing benefits and risks, informed consent, and research ethics review. This last principle refers to the review and oversight of health research by research ethics committees (RECs - international), also referred to as research ethics boards (REBs - Canada), and as Institutional Review Boards (IRBs - United States).

Health research regulation should be tailored to research contexts and more supportive of the imperative of interoperability than external legal frameworks that only incidentally apply to research. This is not always the case. Instead, health research norms tend to restrict health research interoperability both directly and indirectly. They directly restrict the exchange and re-use of research data through narrow informed consent and research ethics review requirements. They indirectly restrict re-use and access to data by leaving important aspects of research governance, specifically data governance, to be determined not by policy-makers, but through multi-stakeholder negotiations between researchers, RECs, individual participants, and increasingly also participant communities.

Within the bounds of existing health research regulation, what steps can researchers take on their own to enhance the interoperability of research? There are two general strategies. First, they can clarify expectations about data exchange and related risks with participants. Second, they can adopt standard safeguards across research networks to manage risks and engender system-wide trust. Relying on the voluntariness of researchers to establish optimal interoperability, however, may be ineffective because of a collective action problem. The risks and compliance burden of distributing research data tend to fall to the researcher who generated the data, while the benefits are diffuse and tend to accrue to the research community and society in general. Voluntariness problems are compounded with labelling issues. If researchers or RECs

---

structure consent permissions narrowly or in non-standard ways, this can restrict downstream interoperability and scientific freedom. This labelling problem is compounded by policies promoting granular, dynamic participant choice over how their data are used and shared. The complexity of tracking, communicating, and enforcing such choices increases with their diversity. Given the complex and unpredictable scientific advancement, it is a misguided idea that a particular researcher, REC, or participant can clearly demarcate valuable from illegitimate data re-use, let alone do so in a few lines of text.

Building awareness about the importance of interoperability across stakeholders – including policy-makers, researchers, RECs, participants, and communities – is an important starting point to address these challenges. Education alone, however, is unlikely to resolve ubiquitous private-ordering problems of negotiation costs, unwillingness, and mis-labelling. National legal or policy approaches clearly limiting the number of parties who can claim control over research data, and how they exercise their control, may also be necessary.

This Chapter introduces health research regulation in Canada, the US, and Europe, and efforts to harmonize regulation internationally. It then provides a case study of how health research regulation complicates the release of research data on the Canadian Open Neuroscience Platform. Third, it discussed the ethical principles of informed consent and respect for communities, and how these tend to translate into unrealistic sharing of control over data between researchers, participants, and their communities. Interoperability must be an explicit priority for regulation seeking to strike a balance between openness and individual/community control.

1 Health Research Regulation: Introduction

Ethical reflection in the context of biomedical research can be traced back to at least the late 19th century. Moral concerns were already being raised over troubling research practices involving deliberate harm caused to vulnerable populations with no or questionable consent. The move towards formal ethical guidelines and regulations for biomedical research is rooted in

---

the Nuremburg Code, which emerged through the prosecution of German doctors who carried out unethical biomedical experiments on concentration camp prisoners during World War II. The Code highlighted the importance of the subject’s voluntary consent and risk-benefit considerations. This was followed in 1964 by the World Medical Association’s Declaration of Helsinki, which provided medical professionals guidelines for research ethics, in part to maintain professional self-regulation over research practice.\textsuperscript{390} The influential US Belmont Report, a response to other cases of ethically concerning research including the Tuskegee syphilis trials, articulated key ethical principles of autonomy, beneficence, non-maleficence, and justice.\textsuperscript{391} It also fostered the emergence of research ethics review systems. In short, modern health research regulation concerns the ethical conduct of research and its scientific validity, as invalid research on human subjects is itself unethical.

### 1.1 Canada

Canada has a dual track approach to health research regulation, with binding national guidelines for publicly funded research, as well as regulation of research conducted in the context of drug approvals. In the first track, the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS2) are research ethics guidelines promulgated by Canada’s three major public research funding agencies.\textsuperscript{392} They apply to all research conducted within federally-funded research institutions, imposed through a Memoranda of Understanding with the funding agency. Clinical trials supporting drug submissions filed for pre-market approval must respect Health Canada guidelines, which follow the International Conference of Harmonization’s Good Clinical Practices.\textsuperscript{393} Health Canada also refers to the TCPS2 as an


\textsuperscript{392} TCPS2, supra note 192.

\textsuperscript{393} Integrated Addendum to ICH E6(1): Guideline for Good Clinical Practice E6(R2), International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), 9 November 2016 [ICH Harmonized Guideline E6(R2)].
important ethics resource.\(^{394}\) I focus on the publicly-funded track, as data-intensive research is often upstream of drug trials. As clinical trials become increasingly data-intensive with the advent of precision medicine, however, interoperability of the second track, as well as between health research sectors, will grow in importance.

A variety of constitutional (e.g., *Charter of Rights and Freedoms*\(^ {395}\)), statutory, and common law rules may also apply to health research across Canada depending on the jurisdiction, institution, and type of research. Notably in the province of Quebec, which is a civil law jurisdiction, key legal provisions of the Civil Code\(^ {396}\) protect the physical and psychological inviolability of the person. Research that interferes with the integrity of the person is generally conditional on the “free and enlightened consent” of the individual.\(^ {397}\)

There are a number of design choices made in the TCPS2 guidelines that affect the interoperability of health research. The guidelines define research involving humans as including research involving identifiable samples and data. Information is identifiable where it “may reasonably be expected to identify an individual” (art 2.1). This means that principles of benefit-risk assessment and informed consent, overseen by mandatory research ethics review requirements, apply by default not only to experimental studies and clinical trials, but also to research solely involving human samples or identifiable data. Only research that relies exclusively on secondary use of anonymous information is exempt from research ethics review (s 2.4). Secondary use “refers to the use in research of information originally collected for a purpose other than the current research purpose” (s 5D). This exemption does not apply to research collecting identifiable information even if these data are subsequently “coded” (where direct identifiers are separated from research data and replaced with a unique code) or


\(^{396}\) *Code Civil du Québec*, 1 January 2018, RLRQ c C-1991 [CcQ].

\(^{397}\) *Ibid* art 10.
“anonymized” (where direct identifiers are irreversible stripped from research data). Given fundamental uncertainty over the identifiability of health research data, the TCPS2 appears to almost always require if not a full REB for re-use of research data, then at least an REB consultation to confirm the data are anonymous and the exemption applies.

By default, researchers using human data must seek consent from the individual (art. 3.1). An REB may waive or modify the consent requirement for secondary use of identifiable information if use of identifiable information is essential to the research, unlikely to adversely affect individual’s welfare, done with appropriate privacy safeguards, complies with any known preferences previously expressed by individuals about any use of their information, and where it is impossible or impracticable to obtain individual consent (art 5.5A). The general informed consent standard for research is “full and frank disclosure” of all relevant information, including detailed information about the management of identifiable information (arts. 2.7, 3.2), though these requirements can be modified by an REB. Broad consent – consent to future uses of data not fully specified at the time of consent often accompanied with ongoing governance by oversight bodies – is not explicitly recognized by the TCPS2, though the practice is widely practiced across Canada. In summary, Canadian researchers typically have to negotiate permission to re-use research data with an REB, and potentially also with the participants themselves. These issues have been addressed to some extent through the centralization of data in research community repositories overseen by data access committees (see Chapter 7).

1.2 United States

Health research regulations in the United States permit greater interoperability of health research than Canada. In the United States, the “Common Rule” governs federally funded research involving human subjects. Separate albeit similar regulations govern research investigating drugs, biologics, and medical devices regulated by the United States Food and


Drug Administration. My discussion here will focus on the Common Rule, as most data-intensive research is currently upstream of regulated products.

The Common Rule underwent important revisions in 2017. Research involving identifiable biospecimens is now included in the definition of human subjects research. Identifiable private information remains defined as “private information for which the identity of the subject is or may readily be ascertained by the investigator or associated with the information.” This definition is subject to regular review by federal agencies. Overall, it appears more permissive and clear as to the sharing of de-identified samples and data. The Common Rule requires a statement in the consent for any research involving identifiable samples or information as to whether or not there is a possibility of secondary use once identifiers are removed.

As for the sharing of identifiable samples and data, the revised Common Rule now recognizes broad consent. Broad consent “allows study subjects to sign off on future, secondary research that may be conducted with their identifiable information or identifiable biospecimens.” Broad consent is a new option for authorizing research, in addition to the former means of authorizing re-use: rendering the biospecimens or data non-identifiable, consent to a particular re-use, or an ethics waiver of the consent requirement under certain conditions. A complication arises where researchers ask a patient for broad consent and the patient refuses. The new Common Rule states that a refusal will preclude future ethics waivers, as it would be

---


403 Ibid 45 CFR §46.116(b)(9).

404 CenterWatch, supra note 401 at 3.
inappropriate to override the individual’s explicit refusal. Operationally, however, it is unclear if researchers or institutions have the means to track and enforce refusals over time. It is also unclear how long a refusal would remain valid.

How much information must a broad consent provide? According to the Common Rule, a broad consent form should include a general description of the information and biospecimens that may be used, the types of institutions and researchers that may access and use these resources, and “the types of research that may be conducted” according to a reasonable person standard. The Department of Health & Human Services Secretary’s Advisory Committee on Human Research Protections (SACHRP) has provided interpretive guidance on the broad consent rule, stating that the consent should include both a general description of the purposes as well as illustrative examples, especially if controversial types of research are foreseen (e.g., creation of cell lines, abortion/induced termination of pregnancy) and the areas for which specific mention is required by applicable laws (e.g., genetic testing, drug abuse/alcoholism, HIV/STDs, mental health).

The guidance also states that “[a]ttempts should not be made to present an exhaustive list of research areas or issues, as such a list cannot be comprehensive and could, in fact, mislead the individual considering the broad consent.” The approval of an Institutional Review Board (IRB) is still required for secondary use, to ensure that the proposed secondary research aligns with the language of the broad consent. The Common Rule may still allow for a blanket consent and public release of identifiable data as long as the participant is clearly informed that there will be no restrictions on use.

405 Ibid at 10.
408 Ibid.
409 CenterWatch, supra note 401 at 14.
In addition to establishing broad consent, there are also efforts to streamline IRB review under the revised Common Rule. A new rule was put in place that requires multi-centre collaborative research to be overseen by a single IRB at one of the institutions, though this provision has encountered push-back and in turn the application of the rule has been delayed.\textsuperscript{410} Only limited IRB review is required for certain activities, such as storing identifiable samples or data under broad consent, or for secondary research using those samples. This usually means one designated individual reviews the activity rather than the full board, with a focus on assessing the appropriateness of privacy and confidentiality safeguards.\textsuperscript{411}

Canadian health research regulation should be updated to more closely reflect the US Common Rule approach. The US Common Rule takes a clear policy stance with regards to enabling researchers to establish varying degrees of interoperability, to suit different research contexts. The TPCS2 does not prohibit broad consent or secondary use without consent, but it does require an ethics board to exceptionally approve both approaches on a case-by-case basis. This places excessive responsibility and administrative burden on Canadian REBs, who do not necessarily have the privacy and security expertise to make such determinations. This may result in cautious or inconsistent decisions by bodies mandated to represent the interests of participants. In terms of drafting, the Common Rule is more practical and detailed, whereas the TCPS is more principled-based, leaving too much to REB interpretation.

1.3 Europe

In the European Union, there are a number of regional human rights norms and conventions that apply to health research, in combination with national laws and research regulations. This legally binding human rights foundation distinguishes the European approach from the US and Canada. It begins with the European Union’s \textit{Charter of Fundamental Rights} and the European \textit{Convention on Human Rights} discussed above in Chapter 4, which guarantee

\textsuperscript{410} Ibid at 3.

\textsuperscript{411} Ibid at 17–18.
human rights to privacy and data protection. The Council of Europe’s *Oviedo Convention* is a legally binding treaty that outlines general principles for protecting the integrity and fundamental rights of individuals in biomedicine. For medical research, the Convention highlights the primacy of participants’ rights and interests, protection of individuals unable to consent on their own, and the importance of informed consent and privacy. It also prohibits genetic discrimination, prohibits the creation of human embryos for research purposes, and requires safeguards for the use of existing embryos in research. An *Additional Protocol on Biomedical Research*, though not widely ratified, outlines rules for biomedical research addressing ethics oversight; free, informed, express, and specific consent; and participant confidentiality.

By framing health research regulation in the language of human rights and international law, there is less flexibility for adapting rules at the level of guidelines or even national laws and regulations. There is also great deference to the rights of individuals in Europe, including rights to data protection. As a result, in the EU it is harder to make consequentialist arguments in favor of greater openness of health research data. The EU is also pressing for greater transparency and accessibility of research publications and data, albeit without compromising high privacy and security standards. Canada should resist the tendency to impose strict legal formalities of health research, as this tends to impose disproportionate compliance burden in low-risk research contexts.

---


414 Additional Protocol to the Convention on Human Rights and Biomedicine, concerning Biomedical Research, Council of Europe, CETS No 195 [Additional Protocol to the Convention on Human Rights and Biomedicine, concerning Biomedical Research].

1.4 International Harmonization

International harmonization of health research regulation, including research ethics review mechanisms, has followed the “trend towards international collaboration in both the conduct and the regulation of medical research and drug development.”\textsuperscript{416} The Declaration of Helsinki is influential soft law referred to in a number of national norms, though there is some controversy over the effectiveness and specificity of these references, in part considering the document’s numerous revisions.\textsuperscript{417} Drug regulators in the US, EU and Japan led the development of the International Conference of Harmonization, Guideline For Good Clinical Practice (ICH-GCP). The ICH-GCP guideline sets an international ethical and scientific quality standard for drug trials that involve human subjects.

The main objective of the ICH-GCP guideline is to facilitate the acceptance of clinical data collected in other countries by this harmonized standard. This should simplify drug approval procedures for research sponsors and could ultimately help avoid duplicating studies and decrease delays and costs.\textsuperscript{418} The World Health Organization and Council for International Organizations of Medical Sciences have developed the International Ethical Guidelines for Health-related Research Involving Humans (2016).\textsuperscript{419} The scope was intentionally drafted to capture a range of health research activities including “observational research, clinical trials, biobanking and epidemiological studies”, and research involving (only) health-related data.\textsuperscript{420} The aims of the Guidelines are not explicitly to promote harmonization but to uphold “universal” ethical


\textsuperscript{418} Hirtle, Lemmens & Sprumont, supra note 416 at 265.

\textsuperscript{419} International Ethical Guidelines for Health-related Research Involving Humans, Council for International Organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization (CIOMS/WHO), 2016 [CIOMS/WHO Guidelines].

\textsuperscript{420} Ibid at ix.
principles. Only in its section on online health research does it explicitly address the need for harmonization, stating that RECs “should emphasize the need to provide similar levels of protection to research activities that pose similar privacy risks.”

Divergence between health research regulation across jurisdictions is problematic because it may compel international collaborations to comply with the rules of the strictest jurisdiction, in order to allow researchers from all jurisdictions to join. There may also be policy conflicts between the regulations of different jurisdictions, especially as these regulations start to migrate away from establishing baseline privacy and autonomy protections, and towards striking a proportionate balance between privacy protections and openness (e.g., through transparency and data sharing requirements). Even in the context of a single jurisdiction and dataset, the conceptual muddle over who controls research data can lead to an accumulation of restrictions, conditions, or negotiation costs that in sum are more restrictive than any single party intended.

Health research regulation should be more amenable to international harmonization than intellectual property or data privacy. On the one hand, research communities and activities have a truly global character, and there are numerous prominent efforts to develop international soft-law, which promotes convergence. On the other hand, research ethics review mechanisms tend to be a national phenomenon. Even where regulations are harmonized on paper, they may still be subject to different interpretations by RECs. Moreover, with the rapid exchange of data between institutions and countries, there may increasingly be situations involving extra-territorial or extra-institutional application of ethical standards by guidelines, RECs, or data access committees. For example, Denmark’s national guidelines for whole-genome sequencing research, require Denmark’s rules to be imposed by contract on any international collaborators. For example, international collaborators must agree to respect the Danish policy on identification and return of incidental findings of clinical relevance to participants. Even where extra-territorial application is not imposed at the level of policy, it may be imposed by

---

421 Ibid at xii.
422 Ibid at 85.
423 Denmark, National Committee on Health Research Ethics, Guidelines on Genomics Research (2018).
RECs as a condition of releasing samples or data to foreign collaborators. At the very least, there is a need to promote awareness of interoperability concerns among researchers, RECs, participants, and their communities exercising sovereignty over research data.

2 Case Study: Health Research Regulation and the Canadian Open Neuroscience Platform

The Canadian Open Neuroscience Platform (CONP) demonstrates the challenges of enhancing interoperability of research data under current research ethics guidelines and regulations. The CONP is a collaborative network across 31 national and international partner institutions of basic and clinical neuroscientists and neuroinformatics researchers across Canada. The CONP aims “to remove the technical barriers to practicing open science and improve the accessibility and reusability of neuroscience research to accelerate the pace of discovery.” The CONP has a number of goals including the promotion of innovation publishing and knowledge dissemination practices; providing training in open science; establishing repositories for sharing the data and software code supporting publications; and developing standards to facilitate the discovery and combination of these resources. Its flagship product is the CONP Portal, an open science repository that will host 1) a searchable catalogue of neuroscience datasets, 2) software tools, and 3) datasets related to animal, human, and phantom (i.e., computer simulated) brains.

The CONP Portal proposes two tiers of access. The first is a public tier that anyone can access provided they create an account and agree to an online terms of use (click-wrap agreement). The second is a “registered access” tier, which is only accessible to researchers (faculty or students) affiliated with a recognized research or educational institution, and who also individually agrees to an online terms of use. The purpose of this registered access tier is to enable rapid sharing of potentially identifiable neuroscience data with trusted researchers in cases where there are privacy concerns about releasing the data publicly. The Prevent Alzheimer’s Disease (Prevent-AD) datasets, for example, are drawn from a cohort of cognitively


425 “About | Compute Canada”, online: <https://www.computecanada.ca/about/>. 
normal, at-risk adults followed over time to identify early biomarkers that predict the onset of Alzheimer’s disease. The project plans to release clinical, neuroimaging, and genetic data annually on the CONP Portal.\(^{426}\)

The CONP’s robust conception of open science is that neuroscience data should be freely available to anyone, anywhere for any research purpose. Health research regulation can pose barriers to extending the interoperability of research data. First, it is more difficult to determine if certain types of neuroscience data are identifiable when they are publicly released, and therefore if such regulation applies. The TCPS2 guidelines, for example, direct Canadian REBs to consider identifiability in the context of a specific research project.\(^{427}\) It is therefore uncertain how identifiability should be assessed when data are released publicly or in research community repositories.\(^{428}\) The data will be made available to anyone over a long period of time. It is hard to predict what technologies, or auxiliary personal information, may become available in the future to facilitate re-identification. Also unclear is whether or not researchers using data from public or community repositories are required to seek an additional REB approval of their project. The resulting uncertainty leads to the conservative assumption that research projects sharing or re-using these data are subject to ethics review, and must obtain consent or meet the criteria for an ethics waiver.

For example, it is unclear if the structural Magnetic Resonance Imaging (MRI) scans the Prevent-AD study plans to share through the CONP Portal are identifiable, either alone or in combination with other demographic or clinical data. De-identification techniques such as “de-facing” reduce identifiability but may not render the scans non-identifiable.\(^{429}\) The Prevent-AD study did receive an ethics approval to share de-faced MRI data publicly, but only for


\(^{427}\) TCPS2, supra note 192 ch 5.

\(^{428}\) The difficulty of objective determinations of identifiability where data are broadly available over long periods of time are discussed in Chapter 5.

participants who provided explicit re-consent. In the words of the REB, either consent or “complete anonymization” was required for public release. In not prohibiting public release, the REB acknowledged the privacy risks were low. But in requiring re-consent, the committee appeared to express doubts structural MRIs could ever be anonymized.

A further concern for ongoing studies releasing data at regular intervals is that the data remain coded rather than anonymized. This is necessary in order to follow-up with participants to collect additional information over time. The original research team retains a deliberate mechanism to re-identify participants, in order to collect additional information in the future. In Chapter 5, I discussed the European data protection law issue of whether or not pseudonymized data can even be considered anonymous in the hands of third party researchers. Similarly, under the TCPS2, if data are considered “coded” in the hands of one researcher, can they be considered “anonymized” when they are shared with a third party researcher? This may determine in part if external researchers require a research ethics approval at their respective institution. The Prevent-AD study, for example, is an ongoing, longitudinal study, so the data remain coded with the original researchers. This may have contributed to the REB’s concerns that the data were not completely anonymized. If data released publicly through CONP are considered coded and not anonymized vis-à-vis users, research ethics review of proposed re-uses would be required. This would defeat the aims of the CONP to reduce duplicative oversight and project-by-project negotiation of data use permissions.

Other conceptual challenges are raised when researchers seek permission not for a specific research project, but rather for the public or otherwise broad release of research data. It is unclear what standard of informed consent applies where researchers seek consent not for a specific research project but for public or broad data release. Moreover, is consent sufficient to justify public sharing? Rudolf Amman et al. argue that genomic data should be made publicly available for unrestricted use, and is straightforward to achieve with informed consent. A letter in response pointed out that “some research participants have concerns that uses of their

---

430 Letter to PreventAD from McGill Research Ethics Board (20 August 2018).
431 Amann et al, supra note 383.
data might not fit with their norms or values or might disadvantage certain populations” especially in light of current political economies of data, structural inequalities, and unaccountable data use by commercial companies.432

This exchange reflects a long-standing debate over the ethical validity of both broad and blanket consents. Broad consent is consent to future research not fully specified at the time of consent tied with ongoing governance (e.g., the future research project will have to seek approval from a REC and/or data access committee).433 Blanket consent is consent to future research without restriction. Arguably, public release implies a blanket consent, as there is no means of ensuring oversight of research using the data. Neither form of consent is explicitly recognized under the TCPS2, even though broader permissions are essential for interoperability. Ethics norms should provide flexibility to enable different approaches to be adopted in different circumstances. Blanket consents should be permitted, but only where there is both a high standard of informed consent, to ensure understanding of the consequences of public release of data, and low risks of re-identification and potential harm. The CONP Ethics and Data Governance Framework encourages neuroscientists participating in open science to release data as openly as possible, while respecting data subject rights and interests. This means clear consent information about the aims, benefits and risks of open science, appropriate privacy safeguards that do not undermine the scientific value and accessibility of data, and above all a commitment to the quality and interoperability of data to ensure its maximum societal impact.434

3 Who, Exactly, is in Charge of Research Data?

The categories of stakeholders making sovereignty claims over research data are expanding. The question on everyone’s mind, according to Charlotte Haug is “[w]ho should


control how data are distributed and used by others? The patients themselves? Doctors and researchers? Research institutions or governments? One can also add RECs, families, and communities in this list. The result is a conceptual muddle over what party or parties control the exchange and re-use of health research data. Interoperability is particularly challenging where multiple parties claim to control the same data. This means that researchers who seek to access and use a single dataset, may need to negotiate with multiple parties. As we saw in the context of licensing research data (Chapter 4), even negotiating re-use with a single data “owner” can be extremely restrictive.

The more parties claiming control over the same data, the greater the transaction costs of negotiating new uses. In some cases it is difficult to identify or re-contact one of the relevant parties, precluding exchange and re-use altogether. For example, if researchers are required to ask an individual participant to consent for a specific re-use of that individual’s data, the research will not be possible where the participant cannot be identified or re-contacted.

Moreover, sharing control between multiple parties tends to result in restrictive decisions. This usually means each party holds a veto over re-use. As a result, the preferences of the party with the most restrictive views will hold out. For example, if an individual participant is happy to share his or her data widely, but the individual’s community is not, the community may veto the individual’s permissive stance. A different risk described by Eric Juengst and Eric Meslin is that a conceptual muddle over who controls data can create policy stalemates that allow other non-moral considerations such as source of funding, relevant expertise, and material resources dominate by default, usually to the advantage of the stakeholders who control other forms of power.

This section discusses the growing emphasis in research ethics as well as in health research regulation on individual and community control over data. While these trends have primarily evolved in the context of experimental research, they are increasingly being analogized to governance of research data in ways that threaten interoperability. Nations are also making


important sovereignty claims over research data, particularly genetic data. This is an additional layer of control confronting international researchers and collaborations, but one beyond the scope of this Chapter.\(^{437}\)

### 3.1 Individual Participant Control

Individual autonomy is a core legal and ethical principle in the governance of experimental health research. It is less clear how far the principle should extend to health research data. An initial challenge is the standard of informed consent for the sharing and re-use of data. The traditional standard of informed consent for clinical research is “full and frank” disclosure of the research purposes and any material risks.\(^{438}\) For health research data, however, the potential research purposes, collaborators, risks, and governance safeguards cannot be fully specified at the time of recruitment. In Canada, the standard of informed consent for data sharing is largely left to the determination of REBs on a case by case basis. Interoperability is limited where there is an insistence on detailed and specific consent. The principle of individual autonomy is also increasingly being interpreted to mean individuals should be offered granular choices that can be dynamically modified over time. The revision of the US Common Rule seriously considered the addition of such provisions, thought it ultimately rejected this approach because of concerns about effects on the research enterprise.\(^{439}\)

There are a number of calls for giving individuals ownership-like rights in their health and genetic information. These calls have various motivations,

---

\(^{437}\) National sovereignty and benefit-sharing requirements could become an increasingly important barrier depending on the outcomes of negotiations under the Nagoya Protocol to the Convention on Biological Diversity, a treaty that aims to stop biopiracy and provides countries with control over the export of their biological resources. See United Nations, Secretariat of the Convention on Biological Diversity, *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from Their Utilization* (2010). See also “Science with borders: A debate over national rights could inhibit research”, (14 January 2019), online: STAT <https://www.statnews.com/2019/01/14/science-with-borders-a-debate-over-genetic-sequences-and-national-rights-threatens-to-inhibit-research/> (Foreign researchers who want access to protected samples must first seek government permission. A number of countries are insisting the protocol applies to genetic sequence data of viruses and pathogens, though this is not in the explicit wording of the treaty).

\(^{438}\) *Halushka v University of Saskatchewan*, 1965 53 DLR (2D) 436.

ranging from concerns over individual autonomy, privacy, and dignity, to offering a more palatable alternative than corporate ownership, to creating a basis for data-based market transactions, and to dissatisfaction with existing regulatory and administrative data protection frameworks.440

Contreras argues that such rights present a new threat of a tragedy of the anti-commons to biomedical research and public health monitoring. The tragedy of the anti-commons was originally articulated by Michael Heller and Rebecca Eisenberg, who cautioned that patents on DNA sequences threatened to fragment ownership, increase transaction costs of research, and hamper innovation. This is because “each upstream patent allows its owner to set up another tollbooth on the road to product development, adding to the cost and slowing the pace of downstream biomedical innovation.”441

Similarly, expansion of the informed consent doctrine could have a propertizing effect on any data generated from a clinical or research intervention, if data are considered under the exclusive control of the individual.442 Granting individual participants stronger property-like interests in their data would also lead to greater fragmentation and transaction costs, but on an order of magnitude greater than the patent anti-commons. Indeed, modern research cohorts include millions of participants. Property-like rights in data would have important consequences on health research and public health monitoring. They would lead to greater upfront costs in seeking consent, and downstream risks to the value of data resources and analyses should individuals seek to exercise the right to exclude and request withdrawal of their data.443 Not all forms of individual control will diminish interoperability. Individuals may have a higher tolerance for broad permissions than regulators. To the extent individuals can directly access a

440 Contreras, supra note 439.


442 Contreras, supra note 439.

copy of their data, they may be able to directly or indirectly “port” their data to other researchers, users, or platforms.\footnote{Thorogood et al, supra note 85.}

There has been a movement towards offering individuals more granular control over their health and research data. As a starting point, this may mean separating the decision to seek healthcare or to participate in a research study from the decision to allow sharing or re-use of the resulting data. In data privacy law, this is referred to as a restriction on “tying”.\footnote{Schwartz & Peifer, supra note 335 at 120.} A criticism of research consents is that participants are presented with a non-negotiable consent of adhesion: they must accept or reject the proposal, but they are unable to modify the proposal in ways that would suit their own preferences. Thus the consent document is not always an accurate proxy of the participant’s expectations or preferences. A step further would be to give participants granular choices over who may access their data or what their data may be used for. An analogy is the granular privacy settings found on social media sites like Facebook. Participants could be given options not just between a list of types of researchers, organizations, or uses, but also a choice between giving a broad consent or being asked to re-consent (through a consent portal or app) for each future research use. It is also technically possible with sufficient technical infrastructure to extend individual control over data through time, as an “adaptive” or “dynamic” consent. Jane Kaye et al.’s dynamic consent model is conceived of as a two-way communication tool, allowing for both ongoing participant engagement, as well as adaptive participant control over access and use of their samples or data.\footnote{Jane Kaye et al, “Dynamic Consent: a Patient Interface for Twenty-first Century Research Networks” (2015) 23:2 European Journal of Human Genetics 141.} Perhaps the pinnacle of participant-centric data sharing is the platform developed for rare disease communities by the Genetic Alliance called the PEER Platform (Platform for Engaging Everyone Responsibly).\footnote{Genetic Alliance, “Platform for Engaging Everyone Responsibly”, online <http://www.geneticalliance.org/programs/biotrust/peer>.

\footnote{444} Thorogood et al, \textit{supra} note 85.  
\footnote{445} Schwartz & Peifer, \textit{supra} note 335 at 120.  
\footnote{447} Genetic Alliance, “Platform for Engaging Everyone Responsibly”, online <http://www.geneticalliance.org/programs/biotrust/peer>.}
a blanket or broad consent, to a tiered consent, to a dynamic consent where they must give a specific authorization for each proposed research use.

Any data-intensive research consent model that increases individual choice of participants, potentially decreases the representativeness of the data. Certain types of people may be more or less likely to contribute or agree to data sharing. This may impact the validity and generalizability of research. Some criticize giving more choice and control to participants on psychological grounds. The perception (illusion?) of control may encourage people to share more, perhaps in ways they did not intend.\textsuperscript{448} Others criticize dynamic consent from the perspective of psychological overload. Participants do not necessarily have the time, knowledge, or interest to constantly manage and negotiate the exchange and re-use of their research data.\textsuperscript{449}

Implementing systems to ensure respect for individual choices, especially across distributed research ecosystems, is non-trivial. Every option provided to participants increases the operational burden for researchers and research data repositories for recording, tracking, communicating, and enforcing those choices. Given the resources and infrastructure required to enable dynamic consent, the pressing policy question is not whether or not to respect individual preferences, but how much to invest in consent infrastructure. The research community aims to build data infrastructure to support the aggregation, sharing, and analysis of research data. The community would understandably be reluctant to invest in building consent infrastructure that undermines these goals. In short, granting individual participants granular and dynamic control over their data threatens the interoperability of research data. To support interoperability, health research regulation should explicitly permit broad consent models and work on strengthening associated governance, rather than fetishizing granular, individual choice or building health research infrastructure that risks being empty or unusable.

\textsuperscript{448} Hartzog, \textit{supra} note 274 at 95–96.

The conceptual muddle over control, however, is not limited to whether policy-makers, researchers or participants should determine the scope of data sharing and re-use. I now turn to the growing move to recognize community control over research data.

3.2 Participant Community Control

3.2.1 Respect for Communities

A second emerging principle in research ethics is respect for communities. Charles Weijer et al. formulate this principle as conferring on researchers “an obligation to respect the values and interests of the community in research and, wherever possible, to protect the community from harm.”450 This principle is prominently featured in the 2016 CIOMS/WHO Guidelines:

[researchers, sponsors, health authorities and relevant institutions should engage potential participants and communities in a meaningful participatory process that involves them in an early and sustained manner in the design, development, implementation, design of the informed consent process and monitoring of research, and in the dissemination of its results.451

International and national health research guidelines include robust protections primarily for indigenous communities. These protections are motivated by their distinct cultures, the importance of their self-determination, and concerns that research may adversely affect their members and values.452 Protections include consultation in protocol development, research document, development of consents and governance documents, and involvement in manuscript preparation. The Canadian TCPS2 guidelines include a chapter on research involving the First Nations, Inuit, or Métis Peoples. This chapter requires community engagement where research is likely to affect the welfare of the community. Review by a community REB or the leadership of an Indigenous community are considered appropriate forms of engagement.453

451 CIOMS/WHO Guidelines, supra note 419 Guideline 7.
452 Weijer, Goldsand & Emanuel, supra note 450 at 277.
453 TCPS2, supra note 192 arts 9.1,9.2.
More recently, the Canadian Institutes of Health Research (CIHR), a federal funding agency, has published a draft of general guidelines for developing partnerships with patients across the research life cycle.\footnote{Canadian Institutes of Health Research Government of Canada, “Draft Ethics Guidance for Developing Research Partnerships with Patients - For public consultation - CIHR”, (7 November 2018), online: <http://www.cihr-irsc.gc.ca/e/51226.html>.

\footnote{Jantina de Vries et al, “Regulation of Genomic and Biobanking Research in Africa: a Content Analysis of Ethics Guidelines, Policies and Procedures from 22 African Countries” (2017) 18:1 BMC Medical Ethics 8.}

Mechanisms have also been developed in lower income countries to seek community consent before starting research.\footnote{Government of Canada, supra note 454.}

In addition to protection of community members and values, engagement can also improve both the quality and acceptability of research. It demonstrates respect for communities and their traditions and norms; ensures the relevance, social value, and acceptability of research; encourages the recruitment and retention of participants; allows for mutual education about values between researchers and communities; and helps to ensure an equitable distribution of benefits and burdens.\footnote{Patricia A Deverka et al, “Hopeful and Concerned: Public Input on Building a Trustworthy Medical Information Commons” (2019) 47:1 J Law Med Ethics 70 at Abstract.}

Engagement may be especially important for minorities or marginalized groups, where researchers have limited understanding of the stigma and discrimination they face. Forms of engagement can be characterized across a spectrum (\textbf{Figure 6.1}).

\textbf{Figure 6.1. Spectrum of community engagement}\footnote{Patricia A Deverka et al, “Hopeful and Concerned: Public Input on Building a Trustworthy Medical Information Commons” (2019) 47:1 J Law Med Ethics 70 at Abstract.}
3.2.2 Challenges of Community Engagement

Community engagement presents a number of conceptual challenges, the primary one being: how is a community defined? The Canadian TCPS2 chapter on involvement of aboriginal peoples in research defines community generically as “a collectivity with shared identity or interests, that has the capacity to act or express itself as a collective.”\textsuperscript{458} Communities may be centered around territories (e.g., an aboriginal community governing reserve lands), or organizations with formal mandates and leadership.\textsuperscript{459} Communities may also be centered around more informal identities and interests such as tribes, neighbourhoods, patients suffering from a similar disease, as well as religious, ethnic, or racial communities.\textsuperscript{460}

Another challenge is selecting appropriate community representatives for engagement.\textsuperscript{461} Representatives may have conflicts of interest (financial or non-financial). There are few processes in place for identifying and handling such conflicts. Power inequities or peer pressure within communities may be reinforced through engagement. On the one hand, it is desirable to include individuals with previous experience with comparable studies. On the other hand, there is a risk that patient representatives become professionalized, which may distance them from the currents of average community members. Agreement on how exactly decision-making is shared is probably the most controversial issue, with concern that the role of patient representatives is often tokenistic.\textsuperscript{462} How are disagreements between researchers and communities negotiated and resolved? Who has the final say?

Practically, engagement involves costs and delays for both researchers and representatives. Engagement usually involves upfront investment by researchers, relationship-building, and delays. If it is done honestly, there is no guarantee that at the end of the day the

\textsuperscript{458} TCPS2, supra note 192 s 9A.
\textsuperscript{459} Ibid s 9A.
\textsuperscript{460} Juengst & Meslin, supra note 436 at 76–77.
\textsuperscript{462} Ibid at 55.
research will even take place. The work of community representatives is usually uncompensated, and they face significant time and cost barriers to participating fully. Data collection in the name of community engagement itself may be “research” requiring respect for scientific and ethical norms and oversight. In order to have assurances that community engagement is sufficient, there may be a need for oversight, such as REC review of a detailed community engagement plan. A number of design questions can arise for such plans: Will engagement be open, collaborative, or multi-stakeholder? How early should engagement begin? What engagement modality should be adopted? What issues should it address?

3.2.3 Community Data Governance

The principle is commonly extended to include community control or participation in governance of research biobanks and databases. The norm of community engagement has also found application in the context of research biobank and data governance. The approach of the Canadian Institutes of Health Research (CIHR) for genomic research involving aboriginal communities is a DNA on Loan model where the community explicitly retains ownership of the samples and data, though they may still be managed by a research institution. Historical exploitation of aboriginal communities has engendered suspicion of all forms of resource expropriation. Some aboriginal communities also harbor beliefs that prohibit some forms of sampling and research uses. A growing chorus of authors and health research guidelines call for

---

463 TCPS2, supra note 192 art 9.2 (recommends engagement should begin before finalization of the research protocol and governance documents, so that these can be influenced by the community).

464 Ibid art 9.2 (issues covered can include research priorities, preferred trial designs, willingness to be involved in the preparation and conduct of the study; benefit-sharing of successful products or financial gains; design and understandability of consent and governance documents; defining potential/acceptable individual benefits and risks, which requires a good understanding of the research context and community values/preferences; and transparent descriptions of engagement processes as part of governance inter alia).


466 Weijer, Goldsand & Emanuel, supra note 450 at 278.
extending community protections including control over data to less-cohesive communities, though this is not without substantial problems.\textsuperscript{467}

Jane Kaye \textit{et al.} argue for expansive community engagement, “to bring the voices of data contributors into decision-making around data sharing”.\textsuperscript{468} They call for including more research constituencies and all territories (not just developed countries) in global data sharing governance, arguing certain voices tend to be excluded as data move through global networks through a series of contractual transactions. Overall, however, community engagement has primarily been developed for governing clinical research, and has not been as robustly conceptualized for data governance.

There are a number of prominent examples of type 2 community engagement relating to data governance. These include surveys and workshops to better understand how publics feel about the governance of bioinformatics, research, clinical, and health data. YourDNAYourSay is an international survey of over 10,000 individuals – translated in 15 languages and delivered with engaging videos – which seeks to explore perspectives about the sharing of genomic and health-related data.\textsuperscript{469} One preliminary finding is that (even) people from Western, English-speaking countries know very little about genomics in the first place, unless they have had professional or personal experience or exposure through popular culture.\textsuperscript{470} Less than half were happy to share data broadly, in part due to suspicion of commercial use.

The US Medical Information Commons project held community focus groups, with strong representation from minority groups, to identify ethical and social concerns over the

\textsuperscript{467} Ibid at 277.
\textsuperscript{468} Jane Kaye \textit{et al}, “Including All Voices in International Data-sharing Governance” (2018) 12:1 Human Genomics 13 Intro.
\textsuperscript{470} Anna Middleton, “Interview at the EAPM Conference on YourDNAYourSay (28 November 2018)”, online: <https://www.youtube.com/watch?v=HZVMQNavTE8>.
governance of networks of biomedical databases available for research and clinical purposes.\textsuperscript{471} The focus groups revealed concerns over commercial access to data, and fears this would exacerbate the already unequal access to healthcare services in the US. Respondents, especially minority respondents, also expressed distrust in research institutions and frustration in having limited control over their own data. In light of these concerns, the focus group recommendations encompassed “opt-in consent, transparent data policies, public representation on … governing boards, and strict data security and privacy protection.”\textsuperscript{472} Experiments are underway with more radical forms of community engagement. Community representatives can for example be directly incorporated into governance and oversight systems. In fact, patient groups and participant cooperatives now manage research databases and engage researchers in governance rather than the other way around.\textsuperscript{473}

Implementing community engagement in data-intensive, networked research environments presents a number of conceptual and practical challenges. The conceptual challenges are analogous to those facing informed consent to broad sharing and re-use of research data, discussed above. Consent cannot be fully informed when future users and uses of research data cannot be fully identified in advance. Similarly, community permissions cannot be fully informed in open science contexts. Broad consent is one common solution in the individual context, where a broad permission is paired with ongoing governance. Likewise, communities can be asked about the acceptability of a broad scope of users and uses, as well as the acceptability of the accompanying governance infrastructure in place to ensure data are used responsibly.

In terms of interoperability, community engagement can cut both ways. On the one hand, community engagement presents an opportunity for greater connectivity between researchers, as communities may be supportive of open science norms of data accessibility and collaboration. In

\textsuperscript{471} Deverka et al, \textit{supra} note 457.

\textsuperscript{472} \textit{Ibid} at Abstract.

turn, their support can help to convince regulators and oversight bodies (e.g., RECs) that such practices are desirable. On the other hand, community engagement may represent another source of policy divergence, if communities insist on distinct access and use policies for different datasets. The latter is more likely. Eric Juengst and Eric Meslin explain that distrust in national governments and “philosophical skepticism of the cogency of individual autonomy” encourages the vesting control in communities between these poles. Health research governance is affected by broader political movements questioning economic and technological progress. There is growing skepticism that scientific progress will continue to mean progress for all. Radical visions of community engagement call for shared governance of health research between researchers and communities, potentially combined with mechanisms to offset background power disparities in negotiation between stakeholders.474 Juengst and Meslin point out, however, that community control tends towards the restrictive:

> appeals to community interests out of a commitment to group solidarity rarely help support policies in favor of wider genomic sample and data-sharing. By definition, communities have to define themselves against the rest of the human population, and structure their allegiances accordingly.475

### 3.2.4 Reconciling Respect for Communities and Interoperability

Risks to interoperability are lower when communities are presented with standard governance models as a take it or leave it proposition. Risks are higher where community engagement involves tailored access and use restrictions, oversight requirements, and benefit-sharing arrangements. Seeking sustained community input over time can ensure governance continues to respond to changing risks and benefits, and changing community values and preferences, but adaptive governance can also disrupt interoperability. Maintaining some diversity among data governance models is desirable. Governance diversity may reflect important community sensitivities essential to their participation, and may allow for valuable experimentation. At the very least, community consultations over data governance should highlight interoperability concerns, so they can consider trade-offs between community values and down-stream interoperability. Recall that interoperability does not mean harmonization.

474 Juengst & Meslin, supra note 436 at 85–86.

475 Ibid at 76–77.
Another way to reconcile the tension between community engagement and interoperability is to engage communities in establishing international standards, rather than project- or dataset-specific governance. Jane Kaye et al. argue that “[g]ood practices for data sharing must evolve towards an interoperable set of standards, permitting sharing across borders. Good practices must also ensure that all views are heard and taken into account when defining these standards.”

Multi-stakeholder global governance is a growing field holding promise for resolving complex issues, while also presenting conceptual and operational challenges. It is commonly associated with Internet governance, but can equally be applied to health research. Multi-stakeholderism is characterized by two or more classes of actors (e.g., states, intergovernmental organizations, firms, and civil society organizations) engaged in a “governance enterprise” of a common issue of a public nature, where authority is distributed among the actors. It often involves inchoate and fluctuating rules and procedures. Likewise for health research data infrastructure, governance “will have to become more transparent, representative, and responsive to the voices of many constituencies, with all the prospects for discord, compromise, and delay that this implies.”

Another approach at reconciling community engagement and interoperability is to treat the governance of research data and data infrastructures as but a small part of a greater system of research governance. As the CIHR guidelines illustrate, there are numerous opportunities for inclusive governance across the research life cycle: priority setting and planning, research proposal development, funding and ethics review, etc. Communities could be engaged about the overall priorities of research, rather than about each specific data governance issue.

In conclusion, health research regulation continues to wrestle with a conceptual muddle over who should control research data. Researchers are increasingly expected to share control with individual participants, and the communities from which those participants are sampled.

476 Kaye et al, supra note 468 Conclusion.
478 Kaye et al, supra note 468 Conclusion.
Interoperability will be restricted if a researcher has to negotiate with multiple parties in order to share, access, or re-use a dataset. Where individuals and communities are given control over data, the interoperability of data will be determined by their voluntary choices. Where individuals and communities decide to apply parochial restrictions on data, this can have unintended consequences on ethical compatibility across datasets. For this reason, participants and communities need to be engaged about the value of interoperability. They may still opt for restrictions and conditions, but at least such decisions will take into consideration their effects on downstream research.
Chapter 7
Access Governance

There are two general lines of solutions pursued to address legal barriers to the interoperability of health research data: access governance and access technology (see Chapter 8). Structurally, access governance typically consists of policies, processes, and safeguards. Access policies articulate who can access research data and for what purposes. Access processes involve due diligence review of access requests by a data access committee (DAC) to ensure the trustworthiness of requestors and/or the scientific quality of their proposed project. Researcher trustworthiness is often assessed with reference to institutional affiliation and project-specific Research Ethics Committee (REC) approvals. A data transfer agreement is the primary safeguard, and can include provisions that cover intellectual property rights and interests, academic credit, privacy and security obligations, and a description of authorized uses (e.g., the approved project only). Access governance systems act as interfaces that ensure consistent protection when researchers exchange data between different legal and organizational contexts. Access governance is a generic solution to legal interoperability, as it can deal with a range of considerations including intellectual property, academic credit, data privacy, research ethics, and respect for participant community values.  

Governance is often contrasted with formal law and regulation. General characteristics of governance include

…being initiated by actors other than the state … non-reliance on law for authority; principles-based modes of conduct to guide action; absence of formal sanction… as well as diverse expression across various contexts.  

The primary goal of access governance is to strike a balance between openness and control, between maximizing the organizational and societal value derived from data while minimizing proprietary, privacy, and ethical risks. Data access governance is a key component of

\[\text{footnotes}\]

479 Canadian Council of the Academies & The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation, Accessing Health and Health-Related Data in Canada (2015) at xxiii.

480 Laurie, supra note 388 at 1.
data stewardship, the view that organizations responsible for data should not only aim to secure and protect data, they should also aim maximize data’s societal value.\textsuperscript{481}

While researchers have long provided colleagues access to data in \textit{ad hoc} fashion, many research funders, journals, and regulators now systematically promote or require the deposit of research data in community repositories, which then provide open and timely data access to the research community. Entities promoting data sharing include research funding agencies and organizations (e.g., US National Institutes of Health, the Wellcome Trust, the Bill and Melinda Gates Foundation, and the European Research Council); biomedical journals\textsuperscript{482}; and even professional societies and associations.\textsuperscript{483} Data sharing aims to enable creative re-uses of data and meta-analyses, reduce duplicative effort in data generation, and improve reproducibility through validation studies, so as to accelerate research and improve human health. In many countries, routinely collected healthcare data are also increasingly being made available to researchers.\textsuperscript{484}

The broad sharing of health research data promises many benefits, but it can also involve risks. Health research data can reveal sensitive information about data subjects and their relatives, presenting risks of discrimination and stigmatization. Broad sharing of health research data can also raise concerns for the researchers or organizations who produce data. Academic researchers may want to ensure they have an opportunity to publish first on their data, or that they receive appropriate credit for their efforts generating, curating, and sharing data. Commercial research companies may be concerned their data will be misused by competitors. Access governance helps to promote timely access to data by researchers, while offering protection for these various interests.

\begin{footnotesize}
\textsuperscript{481} Canadian Council of the Academies & The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation, supra note 479 at xii.


\textsuperscript{483} ACMG Board of Directors, “Laboratory and Clinical Genomic Data Sharing is Crucial to Improving Genetic Health Care: a Position Statement of the American College of Medical Genetics and Genomics” (2017) 19:7 Genetics in Medicine 721.

\textsuperscript{484} OECD, supra note 50.
\end{footnotesize}
As access governance systems multiply, however, it becomes increasingly important to maintain the interoperability of access policies, processes, and safeguards. Otherwise, researchers interested in accessing multiple resources may face exponentially increasing transaction costs and compounding or conflicting restrictions. Fortunately, the interoperability of access governance systems can be enhanced through various strategies, including standardization, centralization, and delegation.

1 Elements of Data Access Governance: Policies, Processes, Agreements and Oversight

This section gives an overview of the general structural and normative components of two common access models: open and controlled access.

1.1 Controlled vs Open-access Data Access Models

The nature of data – and the associated ethical, policy, and legal issues – largely determines the access model, which can range from open to controlled to closed. Open-access models generally make data available to any user, anywhere, over the Internet without financial or technical constraints. Where there are no practical, legal, or ethical restrictions/conditions, data can be made openly accessible and in the public domain. These two terms have slightly different meanings. Open access typically means data are available to anyone, for free, over the Internet. Public domain typically means there are no legal restrictions or conditions placed on use of the data. Open data are both open access and public domain. At the opposite end of the spectrum is data held securely by a single organization or researcher for proprietary, data privacy, or research ethics reasons. The organization may negotiate access by external researchers on a one-to-one basis, but the associated transaction costs drastically limit interoperability.

The Human Genome Project, which sequenced the whole human genome for the first time in the course of thirteen years, shared the sequence data openly. Subsequent publicly-funded projects sequenced more individuals, and combined these data with richer demographic and clinical data, prompting concerns about the privacy of data subjects. Controlled-access models emerged to ensure data could still be shared broadly with qualified researchers while also ensuring confidentiality and security safeguards were in place. Controlled-access models are
common in biomedical and genomic research aiming to strike a proportionate balance between openness and control. This model is used where there is a desire to share data broadly to maximize their value, but where there are proprietary, legal, or policy reasons why the data steward cannot make the data publicly available.

Controlled-access models typically have the following characteristics. Any qualified and trustworthy researcher can request access to data on a non-discriminatory basis. Access is managed by a research ethics committee or increasingly by a specialized DAC. The DAC reviews requests for data access. DACs may be composed of individuals with a range of relevant expertise including familiarity with the scientific area, privacy and security, and research ethics. As William Lowrance notes,

some of these groups are formally constituted, have terms of reference and hold regular meetings. Others, are casual, rarely meeting but existing to be consulted from time to time by the custodian and in a position to address serious problems should any arise.\textsuperscript{485}

These individuals may be part of the scientific team that generated the data, though the independence of members is often included to avoid conflicts of interest. On one hand, real or perceived conflicts of interest may arise where the researcher who collected the data controls who is allowed to access and use data. On the other hand, the team generating the data has important expertise allowing it to assess project feasibility. As a middle ground, a member of the study team familiar with the data and the field may be involved in an advisory role. Given the importance of respect for communities under health research regulation (see Chapter 6), DACs can also include representatives or explicitly consider community values during review.

DACs may review the credentials and affiliations of the applicant, the scientific and ethical aspects of the applicant’s research proposal, and the applicant’s organizational accountabilities (is the proposed research project approved by and overseen by a research ethics committee? Does the researcher’s institution follow appropriate privacy and security processes?) Typically controlled access models also include a data transfer agreement (see below).

1.2 Data Transfer Agreements

Another component of both controlled and some open access models is a data transfer agreement (i.e., data access agreement, data use agreement). This agreement establishes conditions governing the accessing researcher’s use of data. The terms of data transfer agreements may address data subject privacy and ethical protections, including obligations to keep data confidential, to respect consent-based use limitations, to ensure appropriate security safeguards are in place, to refrain from unauthorized linkage of individual-level data, or to refrain from attempting to re-identify participants. The terms may also include terms to protect the rights and interests of the researchers or organizations producing data, such as publication embargoes to allow the producing researchers the first attempt at publication, or intellectual property clauses that give data producers rights in downstream commercialization. Still other clauses may address the interests of science and society, such as requirements to publish findings open-access, or to redeposit derived datasets for further sharing.

While data transfer agreements are technically legally binding if they meet the basic requirements of contract formation, their enforceability in the case of violation has rarely been tested.486 It may also be practically or legally impossible to enforce a breach of the data access agreement if identified.487 Practically, it is unlikely that many data stewards would have the resources to enforce these agreements as contracts, especially where data users are in different jurisdictions. Some “self-help” solutions may be possible. In the case of a serious breach that cannot be remedied, access could be formally suspended or revoked. Reporting the incident to other interested parties may trigger other accountability mechanisms (e.g., to an employer/institution, a supervising research ethics committee, a funding agency, or another DAC).488 Some form of public or community naming-and-shaming may be adopted, though this


could create liability risks under defamation laws. In some cases, there may be recourse against data users under statute law. For example, data users may independently breach data privacy law obligations to keep personal data confidential or to employ adequate security measure to protect the data.

1.3 Monitoring of Data Use

DACs may additionally develop tools and mechanisms to maintain oversight of downstream data uses. For instance, the data users may be required to provide periodic reports about how data are being used, or publications resulting from the data use. Implementation of ongoing oversight, however, requires infrastructure and human resources that may be burdensome for some data stewards to provide. There may also be important burdens (e.g., reporting or transparency obligations) placed on data users that discourage access in the first place. Monitoring may be more feasible where access governance is combined with technological protections, such as access provided within a secure cloud computing platform, where it is easier to monitor user behavior.

2 Examples of Data Governance Models

The International Cancer Genome Consortium/25K Initiative (ICGC) was a large-scale genomics research initiative aiming to generate and share 25,000 whole genome sequences from 15 jurisdictions to better understand the genetic changes occurring in different forms of cancer. The ICGC adopted a tiered access approach, with open access for data unlikely to be linked to other data that could re-identify individual participants, and controlled access for more sensitive data like raw sequence and genotype files (though the exact data types in these two categories

---


490 EAGDA, supra note 486.

491 “About Us”, online: <https://icgc.org/about-us>.
More sensitive data can only be accessed through the Data Access Committee Office (DACO) to protect the privacy and expectations of study participants, and scientific community norms of attribution and publication priority. The DACO reviews the purpose and relevance of research proposals, and the trustworthiness of applicants to protect participant privacy and data security. The ICGC adopted a plain language access agreement restricting users from establishing intellectual property on primary data or attempting to re-identify individual participants, with signatures from the principal investigator and institutional signing official. Recognizing that requirements for ethics review vary from country to country, the DACO asks applicants to indicate if their study of ICGC data requires local ethics approval. The structure of the DACO has been influential and has informed the gatekeeper models of numerous other projects.

ClinicalStudyDataRequest is a portal “facilitating access to patient-level data from clinical studies” carried out by pharmaceutical companies and academic researchers. The portal involves independent review of proposals as well as protections for participant privacy and confidentiality. A major differentiator of this access model from the publicly-funded genomics research context is protection of commercial interests. For pharma-sponsored trials, the data sharing agreement requires users to keep all information provided confidential, in part to protect commercially sensitive information. The user must also agree to give the sponsor an exclusive license to any new intellectual property generated from the study. The agreement also requires users to publish or otherwise publicly disclose their results, which helps to ensure research is pursued for verification rather than commercial purposes.


Ibid.

“ClinicalStudyDataRequest.com”, online: <https://clinicalstudydatarequest.com/>.

“ClinicalStudyDataRequest.com”, online: <https://clinicalstudydatarequest.com/Help/Help-Data-Sharing-Agreement.aspx>.
The MSSNG/Google Autism database contains raw genomic sequence and health-related data from children with autism and their families. The goal of MSSNG is provide as open access to the data as possible to accelerate science, but also to protect the privacy of participants. One of the interesting aspects of MSSNG is that in principle the database is accessible to citizen scientists, which may include experts from fields like computer science or amateurs. The MSSNG DAC does review proposals on the basis of scientific merit and security safeguards, complicating applications by citizen scientists. One interesting angle of the project is how it handles variation in regulatory requirements internationally. Ethics review is only required “if applicable”, noting that some countries (namely the US), and private sector companies generally do not always require ethics review for secondary use of genomic data.

3 Is Access Governance Effective?

A common criticism of data sharing access models is that even well-resourced DACs lack the ability to hold users to account. In the traditional access model where research data are distributed and downloaded, there is usually little means for DACs to monitor compliance with data transfer agreements, or to sanction a breach should one be discovered. A different criticism of access governance systems is that they are too expensive and slow. Not all research teams or repositories have the guidance, resources, or expertise to provide both responsible and timely data access. Adequate support from funding agencies, institutions, or community repositories is key in this regard.

There are a range of access models that modulate the balance between efficiency and protections. Project Data Sphere, for example, is a platform containing HIPAA de-identified cancer imaging datasets. Users have to identify themselves and their purpose for joining on an access request form, agree to the terms and conditions of the platform, and then they have access to all the datasets on the portal. Access is granted largely based on the characteristics of the applicant, rather than assessment of a proposed research use on the basis of scientific and ethical considerations.496

The Global Alliance for Genomics and Health has proposed a streamlined “registered access” model, similar to Project Data Sphere, but limited to “bona fide researchers”, determined by their affiliation with a recognized academic institution. Registered access focuses more on the trustworthiness of researchers and their institutions than on case-by-case assessment of the proposed project.\textsuperscript{497} This has been likened to the security concept of role-based access. Qualified researchers could then be granted access to a dataset, to a platform containing multiple datasets, or to a network of datasets where the custodians all agree to recognize the same credentials. This model is being piloted by the Canadian Open Neuroscience Platform (CONP) (see Chapter 6). Such innovative approaches to data access oversight could be adopted to improve the efficiency of the oversight process by establishing safeguards proportionate to the lower risks posed by some health research datasets.

Transparency of access governance can help to hold data stewards and DACs accountable for efficiently and appropriately managing access. It is preferable if access policies, processes, and terms/conditions are publicly available. Moreover, the principle of transparency could also be extended to the individual access decisions themselves: who is granted access to data, for what purposes, or the provision of reasons when access requests are denied. There may be limits on how much detail are provided to protect commercially or academically sensitive information. Third-party external audits are a potential alternative where the interests in transparency cannot be adequately balanced with data user concerns.\textsuperscript{498} Indeed, there are also compelling meta-arguments that the governance policies and processes of open science should also “walk the walk” of openness, in order to be accountable and open to evaluation.

4 Promoting Access Governance Interoperability

Access policies and processes that determine who can access data and under what conditions are often heterogeneous. This is somewhat inevitable, given the background legal and policy variation surveyed in previous Chapters. Different data privacy law regimes and health


\textsuperscript{498} Floridi et al, supra note 489, s 4.2.
research norms may impose different default restrictions and conditions on data. Researchers may make different commitments to participants during the consent process about how their data will be shared and used (or not). Under a dynamic consent model, the participants themselves may set the access policy for their data, much like one can establish privacy settings on social media. Background ownership rights in research resources may differ, which can in turn be modified by funding agency, institutional, or researcher policy. Even where the rules are the same, the processes may differ. Procedurally, DACs may have different compositions, or may use different forms, authentication mechanisms, and deliberative processes.

Fragmented access governance has the potential to undermine the interoperability of health research. Non-interoperability between access governance is most evident when a researcher seeks to access and combine multiple datasets governed by different DACs for a single analysis. The researcher may have to contend with confusing differences between processes, resulting in compounding transaction costs and delays. Recall from Chapter 4 that having to negotiate permissions with individual owners of research outputs on a case-by-case basis is a hallmark of legal non-interoperability. A similar argument can be made where researchers must negotiate access on a case-by-case basis with multiple DACs. The researcher may have to deal with accumulating restrictions or incompatible conditions across multiple data transfer agreements.

Fortunately, there are strategies that can enhance the interoperability of access governance systems. These include standardization, centralization and coordination.

**Standardization**

Standardizing access policies, processes, and agreements can improve the quality, predictability, and interoperability of governance. Admittedly, this is hard to do without concerted upstream harmonization of legal tools (e.g., consents) and perhaps also legal systems. Often, harmonization involves re-construing consents and data sharing plans in a standard language of access and use. Indeed, aiming for standard access governance can guide upstream harmonization. A number of international health research collaborations have established standard access policies. They then ask data contributors to prospectively or retrospectively harmonize their governance by seeking updated institutional approvals and/or participant
consents. This allows multiple datasets to be made available through a central access process. In short, standardizing access governance paves the way for centralizing access governance.

A key challenge for standardizing access policy is to generally define and agree upon the kinds of researchers who are allowed to access data, and what kinds of research may be conducted with data. These two questions are sometimes distinguished as “access” restrictions and “use” restrictions, though they can also be intermeshed. For example, it is often hard to distinguish a commercial user from a commercial use. Standardizing access policy can proceed incrementally by defining modular attributes that define a qualified researcher. These attributes may include institutional affiliation, position (e.g., faculty member), area of expertise, ethics training, etc. Data stewards are then free to select which combination of these standard attributes is required to access their data. Similarly, using a standard language to describe use permissions and restrictions (e.g., no commercial use, disease-specific research only) can facilitate interoperability of use policies.

Standardizing data use agreements across health research resources is also desirable. Standard agreements are more predictable, easier to understand, and more likely to be compatible when researchers aim to access and combine multiple datasets for a single analysis. The license interoperability discussion in Chapter 4 covers much of this ground, save to say that interoperability is important in both open access and controlled access contexts.

Centralization

Centralization of access governance has historically proceeded by establishing a central repository where researchers deposit their data, turning over responsibility to the repository to manage access. Central repositories aggregate datasets from multiple researchers and benefit from efficiencies of scale on matters such as technical infrastructure design, data curation, and security. Access decisions may also be more independent when made by a specialized data steward, rather than the researcher who created the dataset. An alternative is to only centralize

---

499 Woolley et al, supra note 349.
500 Knoppers et al, supra note 233.
access review, but to allow the data to be stored and controlled at numerous distributed sites. One of the key advantages of centralized access control is it allows researchers to request multiple datasets in a single access request, reducing transaction costs when pooling multiple data sets.

Centralization usually depends on a certain level of standardization, but this is not essential. A central access committee, for example, can manage access to various resources with heterogeneous access policies, though this can become a burdensome interpretive process. This problem has sparked interest in the automation of at least some parts of data access review. In a recent initiative developed a standard vocabulary for data uses, and a tool to allow the creation of machine-readable consent forms.\textsuperscript{501} The proposal is that DACs could use an algorithm to see if data use requests are compatible with the dataset’s use profile.

Centralization raises accountability issues in the case of a data breach. If data are provided to a party that should not have received authorization, who is responsible? The organization managing access, or the organization who originally collected the data? In terms of data privacy law, the answer will differ depending on jurisdiction.\textsuperscript{502}

\textit{Delegation}

Access processes can be broken down into modules, and responsibility for some of these modules can be delegated to one or more third parties. Delegation is akin to third party certification, which is the confirmation of certain characteristics of a person or organisation through review or assessment.\textsuperscript{503} Delegation is an effective tool where the same access process has to be repeated multiple times, by multiple repositories. It reduces duplication of effort because the process only has to be carried out once. Repositories are interested in delegating two aspects of access processes: authentication and authorization. Authentication is a security process that confirms the identity of a researcher (or in digital contexts the identity of a user of a browser

\textsuperscript{501} Woolley et al, \textit{supra} note 349.

\textsuperscript{502} See discussion in Chapter 5 of Canadian personal information protection law accountability principle, and the EU GDPR concept of joint controllership.

or application). An example of delegated authentication in a consumer setting is a Facebook or Google single sign-on, where one online identity allows you to rapidly access many other websites. The CONP Portal delegates authentication of users to ORCID, which is a unique and persistent digital identifier for researchers.⁵⁰⁴

Authorization involves confirming that a user has appropriate credentials to access an otherwise restricted server or resource. A basic credential in research contexts is institutional affiliation. The local institution is best positioned to confirm a researcher’s affiliation, and what position or other credentials he or she holds. DACs traditionally had to rely on institutional web pages and email addresses to confirm affiliations and other credentials. The CONP Portal delegates authorization to eduGAIN, a network that allow universities to authorize and authenticate their own staff and students to access educational and research resources.⁵⁰⁵

Access governance is a due diligence and contractual interface that enables data exchange between distinct regulatory contexts. A variety of governance models have been deployed in health research contexts to enable timely access to research data without widely reported misuse. There is some concern about the strength of access governance in the absence of monitoring mechanisms and sanctions. Beyond everyday frustrations with the limited resources for DACs and resulting delays and inconsistencies in access decisions, interoperability can be threatened between access systems where policies, processes, and data agreement terms are incompatible. Standardization, centralization, and delegation can streamline access across research ecosystems without diminishing protection of data. Access governance is now being complemented by access technologies, which fundamentally change the very nature of researcher-data interactions and associated risk profiles. These technologies are discussed in the next Chapter.

⁵⁰⁴ ORCID, Home Page https://orcid.org/
⁵⁰⁵ eduGAIN, Home Page https://edugain.org/.
Chapter 8
Access Technologies: The Uncertain Promise of Algorithmic Access

Researchers are now turning to technological solutions that promise to circumvent technical and legal barriers in order to enhance health research interoperability. The basic idea is to design application programming interfaces (APIs) that allow a researcher’s algorithm to travel to and run on the data without the necessity of copying, download, or direct access. Algorithms could even be run across secure datasets held at numerous different institutions. But do these technologies raise more practical and legal issues than they solve?

1 Algorithmic Access

1.1 Technical and Legal Barriers to Data Sharing

Data sharing traditionally refers to the distribution of data by researchers, healthcare institutions, and commercial companies to (other) researchers. This involves three parties. Data producers are entities that collect and generate data, such as researchers or hospitals. Data stewards are entities that protect and responsibly distribute data resources. Data stewards may be the producers themselves, or trusted third parties managing data on behalf of producers. Data users are entities that use data for research, clinical, commercial, or personal purposes. The primary means of data sharing is for users to directly download the data from producers or stewards.

There are technical and legal barriers to data sharing. Technically, health research datasets are increasingly too large to download.\textsuperscript{506} Legal barriers to distributing rich individual-level research data outside institutions and jurisdictions are surveyed in previous chapters. Data producers may risk intellectual property (IP) rights and interests when distributing data. Commercial genetics testing laboratories, for example, often rely on trade secret protections for

their databases. This IP protection is lost if the data are disclosed to a third party without adequate safeguards. Users of research databases may risk infringing copyright or database rights, even where those databases are publicly available (see Chapter 4). The data privacy laws of some jurisdictions impose restrictions or conditions on cross-border or extra-institutional transfers of personal data (see Chapter 5). The transfer and re-use of identifiable human research data is generally subject to informed consent and ethics oversight requirements under health research regulation (see Chapter 6). The potential academic and commercial value of data can raise additional practical concerns for data producers. Academically, they may worry other researchers will scoop their ideas, or will not provide them appropriate recognition for generating valuable data. Datasets may inform downstream commercial applications or may be commercially valuable in themselves. Access governance – which basically consists of due diligence review of data requestors and data transfer agreements (see Chapter 7) – plays a central role in managing these issues, but its effectiveness is limited as data sharing scales.

1.2 Algorithmic Access

Faced with such technical, legal, and practical concerns over health research data exchange, interoperability strategy has started to shift from the data layer to the technology layer. One component of this strategy is a paradigm shift to an algorithmic access approach. In this paradigm, a data steward keeps the dataset in a secure computing environment. Data users, such as bio-informatician researchers, submit search queries or analytic code to the data steward, who then runs the code on behalf of the user. Data users never directly access the individual-level data. The steward only returns algorithm outputs to the user, such as aggregate results or even just performance metrics (e.g., “your algorithm correctly identified cancer in a tumor slide 97% of the time”). The feasibility of algorithmic access has been successfully demonstrated in crowdsourced competitions involving confidential medical data. The Digital Mammography DREAM Challenge, for example, enabled the analysis of more than 640,000 de-identified digital

---


mammography images from over 86,000 individuals, without transferring or giving direct access to the images.Only the Challenge organizers had direct access to the image datasets.

A second paradigm shift is database federation. This involves the networking together of numerous databases held securely behind firewalls across numerous research or healthcare organizations. This approach contrasts with the traditional centralization approach, where data producers deposit data in a central database. Federated networks allow each organization to retain control over their data, while also enabling access to large, virtual cohorts of individuals. Cloud computing can greatly facilitate federated networks. Multiple databases, each legally controlled by a different party, could be stored in the same cloud. This allows control over databases to remain distributed across multiple organizations, while the analysis of these databases would all occur in the same computing environment. Another advantage is that data in the cloud can be collocated with software tools, computing resources, and related services.

The analysis of data can be federated where numerous secure databases in a network enable algorithmic access. In other words, a federated analysis is where a data user distributes a research algorithm across a network of secure databases. The outputs of the analysis at each node of the network are then combined and returned to the data user, without any of the databases having to release individual-level data. Federated search functions in a similar manner. In a federated search model, data users can submit a search query to a network of databases, and receive confirmations from databases containing the information of interest. Just as Google Search returns hits from a network of websites, federated search returns hits from a network of health research databases.

Consider the following use cases for federated search in the context of genomics. First, a researcher wants to determine if a particular genetic variant is linked to higher rates of cancer.

---


Second, a researcher wants to recruit patients with a rare genetic variant from a network of hospitals into a precision medicine clinical trial. In both cases, the researcher can submit a search query to a network of genetic databases for a particular genetic variant. The search query might look something like this: “have you observed a patient or patients with nucleotide C at position 3,000,001 on Chromosome 13”?512 This search query is circulated to the network of databases, and each returns a response. The response may simply be a Yes/No, or it may contain greater detail, such as other health or demographic characteristics of the individual carrying the genetic variant. Once a researcher has found the datasets or individuals of interest, he can follow-up with the relevant entities to request more detailed data access, to submit an algorithm to be run on the data, or to re-contact the individual for recruitment into a clinical trial. Federated analysis and search can function even if the underlying representation of the sequence data or structure of the underlying database are different at each site, as long as every organization in the network adopts an API with a standard technical structure.513

2 The Advantages of Algorithmic Access

Federated analysis and search promise to reconcile the long-standing tension between openness and privacy in health research. A broad community of problem-solvers can exploit the knowledge and patterns contained within data, without actually having to copy or access the data. These approaches are most appropriate where it is technically or legally infeasible to distribute the data. Confidential medical data that was previously inaccessible may be rendered analyze-able through algorithmic access, and analyze-able at scale through federated analysis. These approaches may also lower the risk of security and privacy breaches. Where there is only one copy of a dataset in a single, secure database, data leaks are less likely. The traditional approach of creating and distributing multiple copies of data over the Internet increases vulnerability of data to malicious actors, sloppy data handling, and hackers. A federated network

513 Ibid at “Beacon Protocol”. 
approach can provide for high-levels of interoperability specifically for research purposes, without high-levels of general interoperability presenting inherent privacy and security risks.

Because a single data steward retains control over the data, there is also an opportunity for greater trust and accountability. Under traditional data distribution models, user accountability for data privacy law compliance is a major concern for repositories and oversight committees who have to make decisions about whether or not to release data. 514 In a federated analysis model, data stewards would no longer need to worry about restrictions on extra-institutional or cross-border transfer, where applicable. They would no longer have to assess the trustworthiness of researchers requesting access to data, or the adequacy of the privacy laws applicable in the accessing researcher’s jurisdiction. They would no longer have to negotiate detailed confidentiality and security contracts. By extension, patients only have to place their trust in a single institution to protect their privacy, rather than a wide and anonymous network of institutions. Where data stewards retain control over data, patients may in turn exercise more meaningful control over their data. For example, it is technically more feasible for an individual to request their data be erased when a single institution keeps one copy under its control and responsibility, than when copies are distributed to multiple organizations and researchers globally.

Taking the philosophy of decentralization to its logical conclusion, control over data could be held by individuals rather than organizations. In the consumer space, this is the idea behind the SOLID project, led by World Wide Web creator Tim Berners-Lee. 515 SOLID is short for Social LInked Data. Individuals hold their data in a secure Pod, which can be stored on their personal computer, mobile phone, or hosted by a trusted provider. The individual’s POD is hyperlinked to the PODs of their friends and family, and they can choose to make their data accessible to apps via a personal API. This is an alternative to consumer data being stored in centralized databases of app developers, which risks creating proprietary data silos with limited


portability for reasons both technical (data are stored differently) and competitive (consumer lock-in, reduced innovation among app developers). 516

This approach is favored by data privacy law in two ways. First, the individual’s data never leaves the POD, and therefore does not need to be distributed to companies or other countries. Second, individuals can use their individual right to access their personal information in order to populate their POD with content. PODs beg the following question: when designing a federated network, why stop at the institutional level (research institution or hospital), when you can go to the individual level? These two approaches are not necessarily incompatible. Institutions like hospitals are naturally positioned to collect rich-individual level data for legitimate purposes. These institutions will remain as natural data aggregators, even if individuals also want to pursue complementary data management approaches. Many individuals will also lack the interest or sophistication in the dynamic management of their PODs. Indeed, individuals may prefer to delegate management of their data to a trustworthy organization that provides them appropriate rights of control.

The Global Alliance for Genomics and Health (GA4GH) advocates for a federated data ecosystem in genomics where “data may be distributed across many databases and computers around the world, they must be virtually connected through software interfaces that allow seamless, authorized access.” 517 The GA4GH is a public-private consortium of over 500 healthcare centers, research institutions, patient groups as well as technology and pharmaceutical companies that develops technical standards and policy frameworks for the global exchange of genomic and health-related data. 518 Genomics has long had an international, collaborative dimension. The sheer scale and complexity of the human genome demanded a global effort to sequence it through the Human Genome Project. Identifying disease-causing mutations in rare

517 GA4GH, supra note 3.
518 “GA4GH”, online: <https://www.ga4gh.org/>.
disease often requires the sequencing of patients and their family members from around the

globe.

The opportunity perceived by the GA4GH is that genomics is moving from the cloisters
of academia into healthcare, pharmaceutical research, and consumer sectors. As genomic
sequencing is adopted as a standard of care for certain conditions (such as rare disease and
cancer), the number of individuals sequenced will begin to scale into the 10s, 100s, or
millions.\textsuperscript{519} This data will be linked with rich data on healthcare outcomes, and best of all will be
paid for by healthcare systems. The technical and legal barriers to making these databases
available to researchers necessitate a federated approach. The International Cancer Genome
Consortium, which coordinates large scale cancer studies, has adopted a cloud-based, federated
approach where tens of thousands of patient and tumor sequence pairs are analysed together, but
each generating site retains autonomy over its data.\textsuperscript{520} Similarly, the World Economic Forum’s
Breaking Barriers to Health Data initiative aims to create a scalable federated model to share
data across sectors (academic, public and private) and countries, in support of precision
medicine.\textsuperscript{521}

Google also believes that the success of artificial intelligence (AI) in healthcare will
hinge on federated approaches given concerns over privacy and data control.\textsuperscript{522} Google piloted
this technology by training AI on text-messages held on individual phones without the texts
having to leave the phone.\textsuperscript{523} They are now exploring this approach in healthcare. The central
benefit is that data can be utilized by a community of researchers without proprietary or privacy
concerns. This is all the more important for AI, which relies on vast quantities of data that cannot

\textsuperscript{522} MIT Technology Review, supra note 2.
\textsuperscript{523} Ibid.
be collected by any single healthcare institution. There is also concern about the aggregation of healthcare data by private tech companies, given their questionable data governance practices. A federated approach would allow algorithms to be trained on numerous local hospital datasets, and then amalgamated later into a master algorithm.\textsuperscript{524} The hype around algorithmic access and federated networks, however, downplays a number of practical and legal issues. Indeed, recall that interoperability cannot be achieved at the technology layer alone. Data, humans, and legal systems still matter.

3 \hspace{1em} \textbf{Practical Barriers to Algorithmic Access}

Algorithmic access and federated networks promise institutions holding data enhanced security and control. Presumably, this should make it easier for institutions to make sensitive data available to researchers. Building computing and networking systems to enable algorithmic access, however, is likely to require even greater levels of collaboration than traditional data sharing.

The problem of incentives that plagues traditional data sharing is likely to persist or worsen with a shift to algorithmic access. Certainly, greater interoperability of health research promises to offer great scientific, medical, and societal benefits. But data sharing also raises questions about how those benefits and the associated costs are distributed. Data sharing policy in health research has long had to handle tensions between the interests of data producers and data users (at least their academic-competitive interests). The Toronto Statement by major funders and research centers pushed for timely pre-publication release of data produced by community resource projects. The Statement emphasized the value of data release to allow the broader research community to exploit the data, so as to accelerate research.\textsuperscript{525} It permitted time-limited publication embargos to protect the producing-scientists’ right to publish first, and encouraged explicit collaboration between funding agencies, data producers, and data users.

\textsuperscript{524} \textit{Ibid.}

Aligning incentives with optimal societal outcomes is difficult enough in the context of publicly funded, large-scale, community resource projects. Where data is produced in diverse funding and institutional contexts, it is even harder. For example, healthcare-generated genomes present an incredible opportunity for large-scale research, but healthcare institutions and professionals may lack sufficient incentives, technical infrastructure, personnel, and expertise to collect standardized and high-quality data that can be combined with data from other healthcare institutions and used for research. Collecting and curating data for research will always be at best a second priority for healthcare systems.

What policy-makers enthusiastic about algorithmic access seem to miss is that it requires data stewards to become not only data distributors, but also data curators as well as computing and analysis providers. This is almost certain to involve increased costs. Hospitals have to be willing to ensure data are structured in standard ways. Because the analysis has to be run on locally controlled computing infrastructure, the hospital needs to adopt interoperable computing infrastructure and to hire personnel expert in networks and data science (e.g., bioinformatics or AI). Interoperable file formats, computing environments, and network interfaces are required to enable the automated exchange of search or analysis queries and data outputs across a federated network. In short, data producers must incur non-trivial costs in terms of data curation, computing infrastructure, and personnel. Absent commercial benefit-sharing provisions, they are likely to receive little to no direct benefit for doing so. Self-interested entities would be unlikely to accept this arrangement in the absence of significant external funding or support. One way to address cost and expertise issues is for a trusted third-party service provider to act as a data steward on behalf of multiple data producers, to achieve economies of scale. The most likely candidates would be commercial cloud service-providers, though transfer of technical (if not legal) control to the cloud involves its own host of data governance and trust issues.

The credit and commercial-benefit sharing concerns of data producers are also likely to be exacerbated by federated networks. Data producers interested in academic recognition may

---

526 MIT Technology Review, supra note 2 (In the context of AI for example, “federated learning requires every hospital to have the infrastructure and personnel capabilities for training machine-learning models.”).

527 Thorogood et al, supra note 282.
impose conditions on algorithmic access including co-authorship, citation, or acknowledgement. Such requirements may become more difficult for users to meet as the number of data stewards in the network grows. Perhaps the trickiest policy issue with federated analysis is commercial benefit-sharing. On one hand, absent commercial benefit-sharing arrangements, data users are most likely to benefit directly from data sharing, and not data producers. Healthcare systems and future patients may benefit in the long run, but this may not be sufficient motivation for healthcare providers or researchers to make data available. On the other hand, strict benefit-sharing conditions imposed on data access are likely to discourage researchers and companies from seeking access. Researchers may prefer to seek out other data sources that offer favorable (i.e., no) benefit-sharing conditions. Given the uncertainty around biomedical discoveries leading to commercially-valuable findings, it is difficult for both data users and data stewards to evaluate the likelihood of downstream commercialization and to negotiate terms.

The problem grows even more complicated as researchers seek to analyze multiple data resources across a network. Any commercialization conditions imposed on algorithmic access by data producers could quickly accumulate as the number of nodes in the network increases. This can lead to a problem similar to the phenomenon of royalty stacking in technology contexts, where royalties on licenses for multiple, related technologies exceed a fair price for a single license on the entire stack of technologies. In federated analysis, even modest commercialization clauses imposed by each node could rapidly accumulate to exceed a fair price for access to the equivalent overall dataset. Even if royalty-stacking does not occur, the transaction costs of negotiating data access with multiple nodes may discourage scientifically and socially beneficial uses, leading to a tragedy of the anti-commons.

There may also be problems where no benefit-sharing conditions are placed on users of federated networks. As the size of a network increases, algorithmic access could potentially provide a concentrated and unfair benefit to a single data user. An analogy could be drawn with Facebook. It is unclear if any particular individual’s personal information would have a recognized market value. Yet, there is something unconscionable about the result that such power and wealth are concentrated in the hands of a single company who can generate results from analyzing the personal data of billions of individuals. With algorithmic access, the difference is that the researcher doesn’t control all the individual-level data, but the researcher may still control the results, where the real value lies.
In the case of AI, the impact of commercialization may also have a structural effect that displaces the societal and market value of the same institutions or parties who provided the data to train the algorithms. For example, an algorithm trained on tissue slides curated by pathologists at multiple healthcare institutions aims to displace or disruptively modify the market niche of those same professionals. This may have contributed to the furor at Memorial Sloan Kettering, where an AI company was granted exclusive license to train its algorithms on 25 million tissue slides curated by the hospital’s world renowned pathologists.\textsuperscript{528} The professionals doing the essential data curation needed to train the AI may wonder why they have no ownership claim in the resulting software. Their frustration may turn to indignation if the AI makes them redundant. Even if you disagree that the holistic role of professionals can be reduced to an algorithm, there are technologists and venture capitalists for which this is a matter of faith.\textsuperscript{529}

These data producer–data user disputes also permeate into broader societal issues about unequal access to healthcare and healthcare technologies. Where significant public funding or data from many patients are needed to develop new medical products or technologies, there is a good argument that those products or technologies should be made available back to society on accessible and affordable terms. Without controls over benefit-sharing at the point of algorithmic access, or at higher levels of policy-making, federated analysis could exacerbate inequality as profits flow to rich technology and pharmaceutical companies and their shareholders, and the technologies (at least initially) are only made available to the wealthiest patients.

Organizations in a federated ecosystem retain control over their data resources perpetually. This means that concerns over costs, incentives, credit, and commercial benefit-sharing remain perpetually open to re-negotiation. As a result, algorithmic access is likely to require even higher levels of cooperation than traditional data sharing. Federated networks were proposed in part as a technological solution to address data producer reluctance to relinquish control over data. The human layer of interoperability, however, remains central. To put it


succinctly: if organizations are unwilling to relinquish control over their data, would they really be willing to relinquish control over their computing systems?

4 Legal and Regulatory Barriers to Algorithmic Access

Federated analysis also fails to be a panacea for legal and regulatory barriers. While technological advances resolve and minimize some such barriers, they also generate new ones.

4.1 Copyright and Database Rights

Federated analysis may avoid restrictions emanating by default from copyright and database rights, because there is no need for copying or distribution of data. Likewise, algorithmic access obviates the burden data producers face selecting an appropriate license when releasing data (see Chapter 4). One of the justifications for IP regimes is to address a market failure resulting from the fact that information is generally non-excludable. Because of the ease of copying, especially in the digital age, it is difficult and costly to distribute data to authorized (i.e., paying) parties while excluding unauthorized free-riders.

Technological protection can also address this market failure. Algorithmic access may also enable sharing of research resources by a repository where there is uncertainty over third-party rights or permissions to copy or redistribute. The technical protection of algorithmic access models may actually outperform the legal protections of copyright, as least from the perspective of data producers. Technical protections offer practical and potentially indefinite protection. One need not be concerned about gaps in foreign legal protection because data are not exchanged across borders. Technical protection can even extend to research resources without underlying IP rights. Another benefit of algorithmic access is that the data resource may receive copyright law protection against circumvention of technological protections.


The problem with algorithmic access, however, is that it risks disrupting the balance established under copyright law between the rights of owners and users of data. Copyright law includes many exceptions and limitations on owner rights. Algorithmic access displaces copyright regimes with technical systems and contracts that favor control by the data producer. In short, algorithmic access may resolve some technical and practical challenges with copyright, but not fundamental issues over the balance between data producer rights, and the interests of both users and society in data flows. In short, IP problems are not solved by algorithmic access, instead, they are displaced from the legal to the human layer, as data owners can maintain a more absolute control over data through technical means.

4.2 Data Privacy Law

Algorithmic access circumvents many data privacy law restrictions and conditions because personal data no longer need to be disclosed to external institutions or countries. In traditional data sharing contexts, data privacy law presents a number of challenges, outlined in Chapter 5. For one, the data steward must determine whether or not data are identifiable before release. This assessment can be difficult in the context of bioinformatics, Big Data, and open data. Where the data are identifiable, they are generally subject to obligations of confidentiality by default. Disclosure of personal data to outside institutions or across borders may variously require the data subject’s consent (potentially including specific information about recipient jurisdictions and related risks); a regulatory decision declaring a foreign jurisdiction’s laws adequate; or detailed contracts outlining purpose limitations, confidentiality requirements, and security safeguards. Data stewards may be required to carry out a privacy risk assessment before transferring personal data to researchers, potentially involving difficult review of the adequacy of the laws of the recipient jurisdiction and the security policies of the recipient institution. Data producers providing algorithmic access do not disclose individual-level data, so they avoid these transfer-related concerns. It would also be easier for data stewards to demonstrate that research is subject to appropriate safeguards, required by Article 89 of the GDPR, where data are not distributed. Data users submitting algorithms to federated networks will also have reduced compliance obligations, as they will no longer need to establish local privacy and security safeguards traditionally imposed by data producers in data access agreements.
Federated networks could also allow individual patients and participants to make more informed decisions. These individuals only have to assess the trustworthiness of a single research or healthcare institution, rather than the trustworthiness of a wider data sharing network. Breaches would be less likely, and where they did occur at the local institution, they would be more likely to be identified and reported to individuals, who can take steps to mitigate harms. It is also easier to erase an individual’s data on request where copies are not widely distributed.

It is less clear, however, if algorithmic access would resolve data privacy consent and purpose specification challenges. The EU GDPR, for example, generally requires that the purposes of processing be specified when seeking consent, though it does state that individuals should be able to consent to broad areas of research, because uses cannot always be fully specified in advance. This breadth has been questioned by subsequent interpretive guidance.

Uncertainty over broad consent would remain a problem even if data remained sequestered, because the personal data would still be processed. The data steward executes an algorithm on the data on behalf of the data user. In Canada, personal information protection laws may be technically less of a legal barrier to a broad consent, as they tend to delegate this determination to research ethics boards (REBs). For example, under Ontario’s Personal Health Information Protect Act, personal health information may be used for research without consent on the condition there is a research plan approved by a research ethics board, and the researcher has entered into an agreement (to protect confidentiality and security, to limit use to ethics-approved purposes, and to not attempt participant re-identification or re-contact). All of these issues apply equally to internal and external researchers, except the need for an agreement on security safeguards (presumably already covered by that researcher’s employment contract). In practice, however, Canadian REBs may still interpret the scope of consented purposes strictly as they seek to comply with both privacy law and health research ethics guidelines.

Middleton et al, supra note 469.

European Commission - Article 29 Working Party, supra note 325.

Some residual data privacy risks of algorithmic access must also be acknowledged. First, there is a risk that algorithmic outputs themselves contain identifiable information. This relates to a standard privacy risk for researchers when they publish the results of any analysis. Aggregate statistics or even trained machine learning algorithms may allow sophisticated users to infer individual-level information. The difference here, however, is that the responsibility for safeguarding the privacy of research outputs falls to the data producer, rather than the data user. Moreover, output datasets may be as numerous and diverse as the questions and algorithms users dream up. Data users or third parties may maliciously or accidentally reverse engineer individual-level data from results. Unintended memorization of data has been shown to occur during the training of algorithms. These risks can be reduced by adding safeguards such as auditing of outputs and limiting feedback to researchers. But this returns us to the practical issue raised above: who pays for all of this?

4.3 Health Research Regulation

A federated analysis paradigm does not appear to ease barriers to interoperability stemming from health research regulation. Where research involves identifiable information, health research regulation generally requires informed consent and oversight by a research ethics committee (REC). Where data does not need to be transferred to researchers, there is a reduced need to remove identifiers from data. Ironically, this may mean that health research regulation is more likely to apply in an algorithmic access context than in a data distribution context, where a rigorous attempt would normally be made to strip the data of identifiers.

Algorithmic access tends to increase institutional control over data, and by extension it can also increase patients’ control over their data. As I discussed in Chapter 6, implementing a granular or dynamic consent is operationally challenging where multiple copies of data are widely distributed because consent preferences would need to be captured, tracked, monitored, updated, and enforced not only by one institution but also across a diffuse research network. If there is only one copy of the data, however, the participant could trust one institution to ensure

respect for their preferences. It would also be more feasible for participants to update their preferences with that institution dynamically over time, or to request that their data be erased. Increased individual control may make research more ethically acceptable from an autonomy perspective, but it may also limit data availability or introduce bias into research. Algorithmic access may also narrow the ability of researchers to seek a waiver of the consent requirement from a REC for secondary use. Ethics waivers are usually only available if it would be impracticable to re-consent participants. Where data are kept by one institution that maintains a stronger relationship with participants, they will be easier to re-contact, so it will be harder to satisfy this impracticability criteria.

The general principle that RECs oversee research involving identifiable human data presents a puzzle for federated analysis. Traditionally, the location of the institution where research takes place determines the applicable health research norms. In turn, the applicable regulations determine whether or not research ethics review is required. In the context of federated analysis, however, it is not clear “where” the research is taking place. The data user submitting code to a network may be subject to oversight by his or her local REC. Each data steward in the network executing the research analysis on behalf of the data user may technically be subject to oversight by its local REB. The data user poses a question or provides an analysis algorithm to one or more data stewards, who carry out the research on the requestor’s behalf. If the network is multi-jurisdictional, different health research norms may apply at different sites. If multiple institutional RECs assert jurisdiction, this could quickly lead to potentially duplicative oversight that hinders research without necessarily increasing protections for individual participants.536

Coordinated oversight is an issue for all multi-centre data-intensive research, especially internationally, not just for federated analyses. The general solution is greater coordination between ethics oversight bodies within and between jurisdictions. There is already a move in single jurisdictions towards single REC oversight of multi-centre trials.537 There is even greater

537 US Common Rule, supra note 402.
justification for such a move in the case of data-intensive research. To the extent the procedural and substantive aspects of research ethics review can be standardized across institutions and jurisdictions, there is opportunity for trust and coordination. Three models are proposed to enable such coordination of ethics oversight. Mutual recognition is where two institutions reciprocally recognize each other’s ethics approvals. Delegation is where jurisdiction over multiple sites is given to a central REC by agreement or legislative fiat. Federation is where a central REC is created with representatives from multiple sites. It is challenging to scale any of these strategies internationally because of the lack of an international legislator, discrepancies between health research regulations, and (perhaps most importantly) a lack of trust that ethics oversight rules and processes in other jurisdictions are interoperable.

Ideally, a data user seeking to run a federated analysis would obtain a single ethics approval, recognized by all of the nodes in a federated network. Presumably, federated analysis would need to be subject to some sort of centralized access oversight body. The institutions in the network could all delegate access oversight by agreement to the centralized body. In turn, the centralized body could request evidence of a single ethics approval from a data requestor. A twist, however, is that the centralized body would not oversee data access; it would oversee algorithmic access. While the security and privacy risks of algorithmic access are lower than researcher access, there may still need to be policies and processes in place to assess the scientific credentials of applicants, and the scientific quality of their proposed use. There should also be some level of transparency for participants about who is being given algorithmic access to the network, for what purposes, and under what conditions.

### 4.4 Other Legal Concerns

Federated analysis reduces privacy and security risks for individual participants, but it may create new security risks for researchers’ ideas and algorithms. Data users must submit their scientific queries or algorithms to a network of institutions. They must therefore trust that each of

---


539 Dove et al, *supra* note 536.
the recipient institutions will respect their academic or commercial interests, which may be tied to or revealed by their query/algorithm. These interests could be safeguarded through confidentiality contracts and software licensing agreements. The data users may insist that each node agrees to only use the query or algorithm for the purposes of running the search or analysis, to keep queries and algorithms confidential, and to employ adequate security to ensure there are no breaches exposing the data user’s query or algorithm to unauthorized third parties. Data stewards could also use these agreements to ensure they have proper permissions to run IP-protected software, and to protect the security of their computing environments from malware intentionally or accidentally submitted by users.

There are also scientific challenges with not providing users direct access to the data. The users cannot confirm the quality of the data. This means that data stewards have the responsibility to curate, organize, and document data on behalf of data users. If done poorly, this raises concerns about scientific validity. If done well, this raises fairness and sustainability concerns for the data steward when all of the direct benefits flow to the data user. A compromise is to provide data users with access to a subset of the data as a sandbox so they can familiarize themselves with the data.

Novel security issues may also emerge relating to data integrity across federated networks. There is growing recognition of adversarial attacks in the context of AI, where interested parties may manipulate the integrity of data to exert subtle influence over the outcomes of model training.540 In a federated analysis, there is a risk that data stewards or unauthorized hackers could manipulate underlying data or the results transmitted back to the data user. The financial incentive to do so may be higher where federated analyses or AI training outcomes relate to industry-sponsored trials, or to billing and insurance.541

541 Ibid.
Federated analysis requires the computing environments of different nodes to be interoperable. As I discuss in Chapter 2, a long-standing interoperability challenge for health research networks is at the data layer. As researchers continuously dream up new studies, and are primarily focused on getting publishable results, how can they be expected to generate data that is scientifically and technically interoperable? As healthcare institutions focus on treating patients, how much of their focus and resources can they put into generating interoperable data for the research community?

For federated analysis, data interoperability is still pressing, especially as data users are unable to harmonize data sequestered at multiple nodes on their own. What makes federated approaches more complex is that they also require additional focus on interoperability at the technology layer. There is a greater need for the computing environments at each node to be interoperable, which in turn demands greater willingness for institutions and professionals to cooperate at the human layer. Organizations may already be unwilling to release data, let alone to ensure the data are interoperable and of high quality. While federated analysis reduces concerns about loss of control or about breach of confidentiality and security, it appears to involve an even greater organizational commitment to openness than data release. Organizations need to invest not only in data collection and curation, but also in computing infrastructure, resources, and expert personnel who can vet and deploy the external software. Recognizing the collective action problem of traditional data sharing, the interoperability of federated networks will depend on the policy-making efforts and funding of research funding agencies and healthcare systems. The experience of the US approach in moving towards interoperable electronic health records, discussed in Chapter 2, illustrates the mix of definitions, incentives, and legal mandates that may be needed to establish interoperability not only of data and technology, but also of human systems.

There are some technical tools and approaches that can ease the complexity of establishing interoperability at the technological layer. Container technologies such as Docker bundle a software model and its dependencies into a light-weight package that can be transferred
and executed reliably across different computing platforms.\textsuperscript{542} Cloud computing could allow for data to be collocated in the same computing environment, but to still be legally controlled by different institutions. Relationships between research/healthcare institutions and commercial (mostly US-based) cloud providers, however, raise their own set of concerns over control, privacy, and cross-border transfers.\textsuperscript{543}

A shift to a federated approach will also depend on technical standard-setting for computing and interfaces. Before even a well-resourced and expert-staffed node can receive and execute algorithms, data and technical standards need to be defined and agreed upon by the community. In technical realms, this process is usually carried out by international consortiums of stakeholders called standard-developing organizations (SDO). An example of an SDO focused on enabling health research is the Global Alliance for Genomics and Health (GA4GH), which aims to develop technical standards and policy frameworks to enable the exchange of genomic and health-related data across countries and sectors (see above). To provide some examples of technical standards, the GA4GH promotes standard file formats for genomic data. Sequence Alignment Map (SAM) is a file format enabling storage of genomic sequences aligned to the reference genome.\textsuperscript{544} If every node in a network stores their data in SAM format, then it is easier to design algorithms to analyze the data or interfaces allowing researchers to query the data. Standard approaches are also being developed for compression of genomic data (CRAM) and encryption (CRYPT4GH), to facilitate storage and security respectively, without undermining interoperability.

Standardization is also very important for the interfaces between systems. Everyone needs to agree on a common interface to allow communication between systems. Interfaces are needed to enable federated analysis and federated search of genomic data. The GA4GH Workflow Execution Service API standard is designed to enable a software package (i.e.,

\textsuperscript{542} Docker, “What is a Container?”, online: <https://www.docker.com/resources/what-container>.

\textsuperscript{543} Thorogood et al, supra note 282.

workflow) to run reliably across a variety of computing environments. The API can be used to “request that a workflow be run, pass parameters to that workflow, get information about running workflows, and cancel a running workflow.” For search, the GA4GH has developed the Beacon API standard, an interface that allows researchers to submit search queries about whether or not certain genetic variants are found in a given genomic databases.

Standardization raises a constellation of meta-governance issues in balancing interoperability, accessibility, and ownership. Who should be involved in establishing a standard? Who should be able to use a standard and under what conditions? How can communities determine who holds IP essential to the implementation of a standard? Should conditions be placed on royalties from such standard-essential patents? Standards-development is often led by a consortium of private sector institutions looking to collaborate on standards necessary to establish competitive markets in software and technology. Open standards are those that allow any organization or individual to contribute to the standards-development process, and typically also make sure the standards are then made freely available to anyone to use. This means no fees or royalties are charged to use the standard. To ensure a balance between incentivizing the development of valuable technologies, and ensuring standards are available to the world on fair terms, SDOs must develop detailed IP disclosure and licensing policies.

Techno-solutionism aims to avoid barriers at the messy human and legal layers. Certainly, the interoperability of health research can be improved at the technology layer, by enabling algorithmic access and federated analysis. Paradoxically, however, establishing and maintaining such technological interoperability will require even greater coordination at the human layer. Different organizations will need to be willing to give up a certain level of control over their computing environment, and to expend resources on establishing infrastructure, implementing standards, and retaining expert personnel. Moreover, the process of establishing


546 Ibid.

technical standards demands human and organizational collaboration to agree on standards-
development governance processes and to contribute on a pre-competitive basis to
standardization.