WORLD ECONOMIC FORUM

Sharing Sensitive Health Data in a Federated Data Consortium Model

An Eight-Step Guide

INSIGHT REPORT

JULY 2020

Contents

Foreword

Accessing sensitive health data at scale will advance research, innovation and patient outcomes

Genya Dana
Head of Healthcare Transformation, Shaping the Future of Health and Healthcare, World Economic Forum

Arnaud Bernaert
Head, Shaping the Future of Health and Healthcare, World Economic Forum

At the World Economic Forum, we think of data as the oxygen that fuels the fire of the Fourth Industrial Revolution. It is readily available and necessary but, if used improperly, it can generate dangerous and unwelcome results. Concerns over how to protect valuable data, especially sensitive, personal data, are at the core of many countries' and institutions' data policies. We see a complex and dynamic data policy landscape evolving around health data in particular; it is becoming more and more complicated to share data to the extent desired to advance research, innovation and patient outcomes. The need to rapidly provide access to health data, while protecting patient privacy and data security, has never been more urgent in the fight against the COVID-19 pandemic.

This paper is part of the Forum's work to create actionable resources for policy-makers, healthcare professionals and leaders of the Fourth Industrial Revolution to navigate complex and sensitive health data policies globally. The Forum is testing a federated approach – where data sets are accessed remotely without movement of data from its secure location of origin – as a practical way to access the disparate genomic and health data sets needed to accelerate the diagnosis of rare disease in patients in four countries. Federated data systems are not new per se, but they are starting to be used more frequently as a solution to accessing multiplying, disparate data repositories in a multinational and multi-jurisdictional world. Being able to quickly and securely access disparate data sets accelerates the ability to gather insights and inform care decisions for precision medicine approach, which uses data to drive more personalized and tailored diagnosis and treatment of disease in patients.

Offering practical advice on how to build a federated data consortium is only possible with the partners in the Breaking Barriers to Health Data project, key to the Forum's precision medicine portfolio of projects. Four genomics institutions in Canada, Australia, the United Kingdom and the United States have worked tirelessly to have the difficult conversations and build the governance model that inform this eight-step guide. We applaud their leadership. This guide also forms a critical input to the Forum's Data for Common Purpose Initiative focused on new models of data governance in the Fourth Industrial Revolution. A recently released Roadmap for CrossBorder Data Flows: Future-Proofing Readiness and Cooperation in the New Data Economy expressly recommends that governments should recognize Federated Data Learning as a valid means of cross-border data (insight) sharing and should not be blocked by legislation. Proactive efforts will be needed to motivate government officials, business leaders and civil society members to establish real-world pilots and to enable continuous and active experimentation with federated data systems, particularly in situations where they are most valuable.

Introduction

Accessing global health data through federated consortiums will reveal disease causes and cures

In the current era of the Fourth Industrial Revolution, data is our most valuable resource.[1] The five leading companies of our time – Alphabet, Amazon, Alibaba, Facebook and Microsoft – rely on data to fuel their successful enterprises. Data is also a resource in the healthcare ecosystem that can improve the standards, quality and outcomes of healthcare and healthcare delivery for patients worldwide.

But just how are health ecosystems using data? As volumes of healthcare data increase, genomic data and other types of sensitive health data provide a treasure trove of information on how to diagnose, treat and generally manage the most complex and destructive diseases – but only if we can look at data across the global population.

Genomic data is a particularly valuable type of health data because it represents the hereditary material in humans (and almost all organisms) called deoxyribonucleic acid (DNA), which stores the "master code" dictating how our bodies operate. More than 99% of genetic code is the same in all people, making it difficult to pick out "glitches" or specific small differences in the genetic code useful for research, diagnosis and treatment of disease without ways to comb through large amounts of data.

BOX 1

Why genomic data?

Genomic data represents our shared DNA and can be broken down into a machine-readable format in a process called genetic sequencing. During genetic sequencing, DNA is broken down into its four chemical bases (adenine, guanine, cytosine and thymine) for analysis. Each human DNA consists of about 3 billion bases.[2] Every human being has such DNA represented by billions of bases, but it is only possible to understand more about our shared DNA and, more importantly, how our DNA impacts or even predicts our health by mode of comparison using large volumes of DNA. This is because more than 99% of bases are the same in all people, making any differentiation more difficult to discern in smaller data sets. In contrast to a base, a gene is the unit by which an individual's one-of-a-kind combinations of DNA bases are inherited. Genes can vary in size from a few hundred DNA bases to more than 2 million bases per gene.[3]

Both in the sheer scale of genomic data and in the complex health data policy regulatory landscape, aggregating such data to improve patient outcomes is complicated. The human genome (your genome is the sum of the DNA in your body or the sum of your genetic data) represents roughly 100 gigabytes (GB) of data, which is equivalent to the size of about 100,000 digital photos. In 2011, our sequencing capacity hit 13 quadrillion bases, which was the equivalent of two miles of stacks of DVDs in data storage (which were used for storage in this era before data storage moved to the cloud). By 2018, however, the human genome (roughly 3 billion bases) fit on a single DVD disk – rather than on the hundreds of discs spanning two miles in 2011.[4] Storing the human genome is progressively getting easier, smaller in size and cheaper. Comparing genomic data to Silicon Valley's Moore's Law, which states that computers double in speed but half in size every 18 months, genomic data is outpacing Moore's Law by a factor of four in storage size.[5]

Aggregating large genomic data sets in ways that researchers and clinicians can use to improve patient outcomes is complicated, in part due to the flood of genomic data from national and institutional genetic sequencing efforts. The human genome (your genome is the sum of the DNA in your body or the sum of your genetic data) represents roughly 100,000 digital photos. It now takes approximately a day to sequence most of the genome of one person, and several hundred dollars, compared to 13 years and $1 billion in 2003. Countries and institutions are sequencing hundreds of thousands of people. In 2018, the UK announced the completion of 100,000 sequences from National Health Service patients. Accessing all of this data, however, remains a challenge due to a complex landscape of data protection laws and health data privacy regulations.

“ Federated data systems are a promising way to enable access to health data, including genomic data, that must remain inside a country or institution because of their sensitivity.

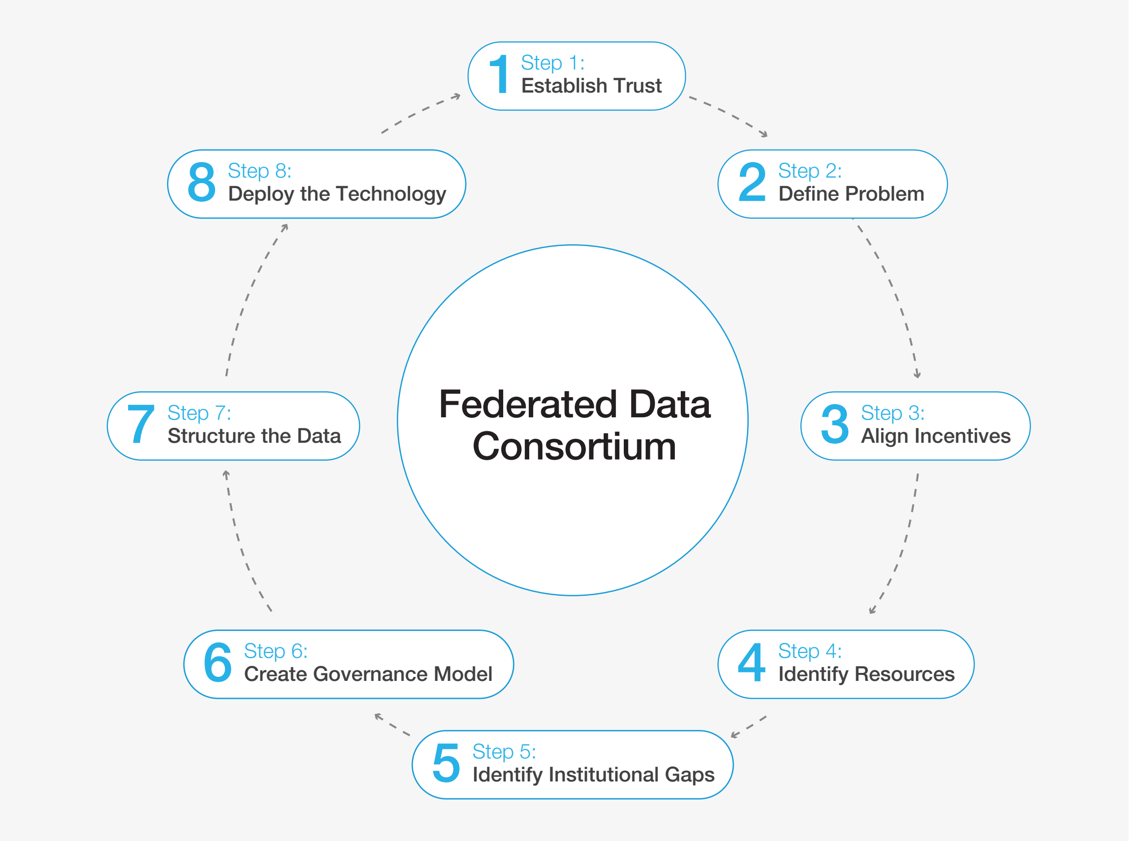
The World Economic Forum’s Global Precision Medicine Council, in its May 2020 Precision Medicine Vision Statement, cited the gap in data-sharing and interoperability as key to preventing the wider adoption of a more personalized approach to healthcare.[6] Precision medicine depends on the availability of health data in the aggregate. For genomic data in particular, the costs of storage and analysis are usually more expensive than the lab costs of sequencing. The cost to store, process and analyse the data can be justified in the global patient interest *if* the data can be used beyond its initial diagnostic capacity for a single patient.[7] Accessing and using sensitive health data and genomic information to its full potential requires care and creativity, with strong governance protocols to guide this process.

To tackle the challenge of governance of cross-border access to health data, the World Economic Forum led the Breaking Barriers to Health Data project, from July 2018 to July 2020. The project tested how a distributed federated data system could be set up and run sustainably across countries with clear governance optimizing for operational efficiency, patient privacy and data security. Federated data systems are a promising way to enable access to health data, including genomic data, that must remain inside a country or institution because of their sensitivity. Although examples of federating health and genomic data sets are growing, how to practically create the federated data system with a group of institutions was not as clear.[8]

Allowing access to data sets is not particularly difficult technically, but there are larger challenges in how to form the necessary relationships between institutions that enable trust and transparency, and sustained, predictable operations in a consortium model. In close partnership with Australia (the Australian Genomics Health Alliance), Canada (Genomics4RD), the United Kingdom (Genomics England) and the United States (Intermountain Healthcare), the Forum created and led a multistakeholder community that supported these institutions through the journey of determining how to maximize the benefits and minimize risks of federating genomic data to diagnose rare diseases.[9]

In order to federate data, a consortium of institutions must be formed. As outlined in Figure 1, this eight-step guide distils the learnings from the Breaking Barriers to Health Data project’s work to set up a federated data consortium for the purposes of diagnosing rare disease using genomic data from a global, distributed data set. Other institutions are also encouraged to adapt this federated data consortium model for additional use cases. Before creating such a data consortium leveraging sensitive health data, it is crucial to carefully plan for such a consortium and meticulously consider how to effectively craft – and implement – clear governance structures. Global federated data consortiums provide a tremendous opportunity to improve patient outcomes and healthcare delivery pathways but also require robust security, continually improving policy to provide safeguards against bad actors, data breaches or other types of preventable risk.

FIGURE 1 **Eight steps to follow to build a federated data consortium**


1 Step 1: Establish Trust
2 Step 2: Define Problem
3 Step 3: Align Incentives
4 Step 4: Identify Resources
5 Step 5: Identify Institutional Gaps
6 Step 6: Create Governance Model
7 Step 7: Structure the Data
8 Step 8: Deploy the Technology
Federated Data Consortium

Step 1

Establish and sustain trust

Generating trust is more important than ever and requires the right partners, thorough relationship building and support from leadership teams

The first step, and the singular component that appears to make or break a federated data consortium, is establishing trust with identified prospective partners entering a data consortium. Establishing trust between partners is also the most time-consuming component in establishing a successful data consortium.

The creators of a new data framework called Trust :: Data Consortium – which include the Massachusetts Institute of Technology, United Nations, White House Cybersecurity Initiative and the Forum – argue that today's social structures do not readily accommodate the new reality of integrated systems that can leverage autonomous, dynamic, digital feedback mechanisms. Our social structures struggle to adapt to digital methods, which can illuminate trust between data-sharing systems by transparently tracking when and how data is accessed or exchanged.[10] In other words, despite many technical solutions designed to encourage trustworthy behaviour between data-sharing partners once a consortium is up and running, establishing trust at the beginning of the relationship is nevertheless contingent on our everyday social structures and perceived social relationships.

1.1 Identify consortium partners

Before beginning to form social relationships with partners, however, it is important to select the correct partners for a data consortium. Identifying the best partners requires understanding of another institution's origin, strategic goals and its research objectives for prospective data consortium partners – and whether or not these align with similar metrics from your institution. A thorough vetting process at the beginning of the relationship cannot be facilitated with a quick website check or even a phone call but requires a series of in-person meetings. At the start of the Breaking Barriers to Health Data project, the Forum found that several iterations of discussion and reiteration of purpose were necessary with each prospective institution before it was possible to move on to discuss details of a partnership. Traveling in person to the location of a prospective partner institution eases the process of uncovering the day-to-day operations and team norms that will be contributed to the data consortium by a prospective partner.

At this recommended in-person meeting (or series of meetings), it is important to discuss: (1) what type of data each institution is currently collecting; (2) how each institution runs its day-to-day operations via a code of conduct or other guidance documents; and (3) how each institution either has control of or does not have control over its short-term and long-term funding. Nothing hurts a consortium foundation more than a group of potential partners saying "yes" to one another without understanding each other's motivations, institutional priorities and data assets. It is also important to ensure that promised actions or outcomes are achievable within institutional priorities or capabilities.[11]

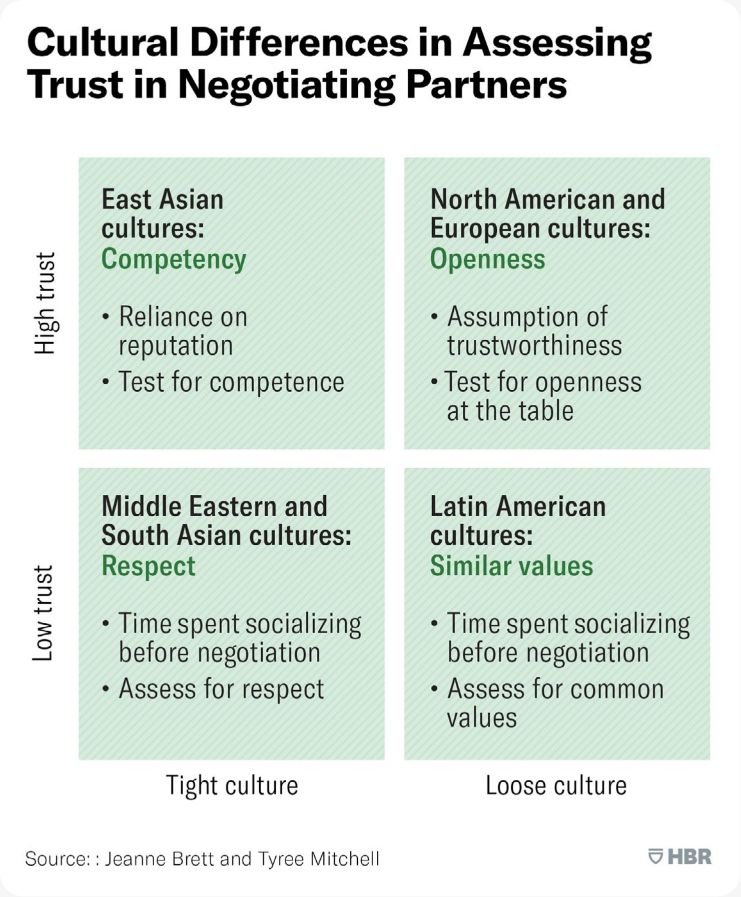
1.2 Encourage trust and prioritize relationship building

It is crucial to establish trust with prospective partners, but *how* to establish trust varies and differs based on region of the world. Depending on geographic location, openness, competency, respect and similar values offer different social cues of trustworthiness.

The *Harvard Business Review* explains that across North America and Europe, openness with information, allowing for the ability to "trust but verify" is the most common strategy. Thus, to evaluate a potential partner for a health data consortium or other data consortium in either of those regions it is crucial to allow partners to verify that each one is acting in a transparent and honest manner. As Figure 2 illuminates, asking open questions that some members of your respective institution already know the answer to (and thus informally double-checking sources of information) is the surest way to establish a base level of trust in North American or European cultures. In East Asian countries, however, reputation often is required to establish competency by demonstrating a successful track record. In Middle Eastern and South Asian countries, the mantra may shift from "trust then verify" to "verify then trust" to confirm such respect. Lastly, the *Harvard Business Review* posits that in Latin America, determining shared values is crucial, resting on social interaction and business interaction.

It is crucial to be aware of cultural differences in trust to ensure that new data consortiums are providing equal opportunity to institutions in other regions of the world, particularly regions not often included in genomic data or health data initiatives. The Forum found that it was easier and faster for institutions from the same region of the world to establish trust between one another due to similarities in how they assessed trustworthiness. Yet this ease of trust will inevitably lead to geographically niche consortiums that do not as drastically expand the variety of data or discoveries sought when entering a consortium in the first place. While it may take longer to establish trust with institutions that are geographically and culturally different, effort must be made for genomics and personalized medicine to realize its long-term value.

FIGURE 2 **Harvard Business Review diagrams show different frameworks for establishing trust in global cultures**



Cultural Differences in Assessing Trust in Negotiating Partners

	Tight culture	Loose culture
High trust	**East Asian cultures: Competency** • Reliance on reputation • Test for competence	**North American and European cultures: Openness** • Assumption of trustworthiness • Test for openness at the table
Low trust	**Middle Eastern and South Asian cultures: Respect** • Time spent socializing before negotiation • Assess for respect	**Latin American cultures: Similar values** • Time spent socializing before negotiation • Assess for common values

Source: : Jeanne Brett and Tyree Mitchell

1.3 Secure leadership support

After establishing trust, understanding the level of support from the leadership at each potential partner organization is paramount. If there is not a clear green light or direct support from the chief executive officer, or equivalent, for the proposed consortium partnership, it will be difficult, if not impossible, for an individual institution to facilitate the level of concerted coordination needed in its staff to share data initially and for the long term.

Entering a consortium to share data is not a side project or a regular, everyday partnership; rather, it requires continued reinvestment in staff and funding (as mentioned in more detail in Step 4) and continued learning at all levels of an institution. Unless the leadership of an institution with health data, including genomic data, is committed to using a federated approach to derive research and clinical utility from its data, the internal will to establish a federation will not materialize in a successful data consortium.

What are we missing? (in the framework)
WORLD ECONOMIC FORUM
COMMITTED TO IMPROVING THE STATE OF THE WORLD

Step 2

Jointly determine the problem for a federated approach

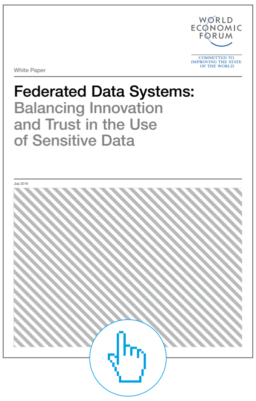
A federated approach is helpful for solving specific problems when the solution requires leveraging distributed data sets

Federating data is an appropriate solution if there is a clear problem that distributed data access can solve. Federating data via the creation of a consortium is an opportunity when solving a problem relies on access to high volumes of data, but it also requires navigating different data policy laws, security and privacy protocols, and data interoperability challenges.

2.1 Identify the problem a federated data consortium could solve

The Breaking Barriers to Health Data project, for example, aims to solve the shortage of genomic data available to clinicians and researchers to diagnose rare disease. Many clinicians globally continue to have insufficient data to provide a diagnosis for paediatric rare disease patients, which not only severely limits the accuracy of other diagnoses but also hinders the development of treatment plans for the nearly 475 million people living with a rare disease across the globe.[12] The average time to diagnosis is seven years and only 5% of people with a rare disease have a US Food and Drug Administration-approved treatment.[13] Increased access to genomic data can improve these statistics because 80% of people with a rare disease have one caused by a genetic or genomic variant.[14] Due to the complicated data policy landscape globally, genomic data cannot readily be transferred across borders or institutions to a centralized data lake or other data pools for easy access. The Forum has found that a federated data system's approach of remote data access without movement of data from its secure location of origin could help clinicians and researchers get enough access to disparate data sets to ask for the variants needed to assist with, or confirm, a diagnosis.

2.2 Determine if a federated approach is the appropriate data system solution


WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

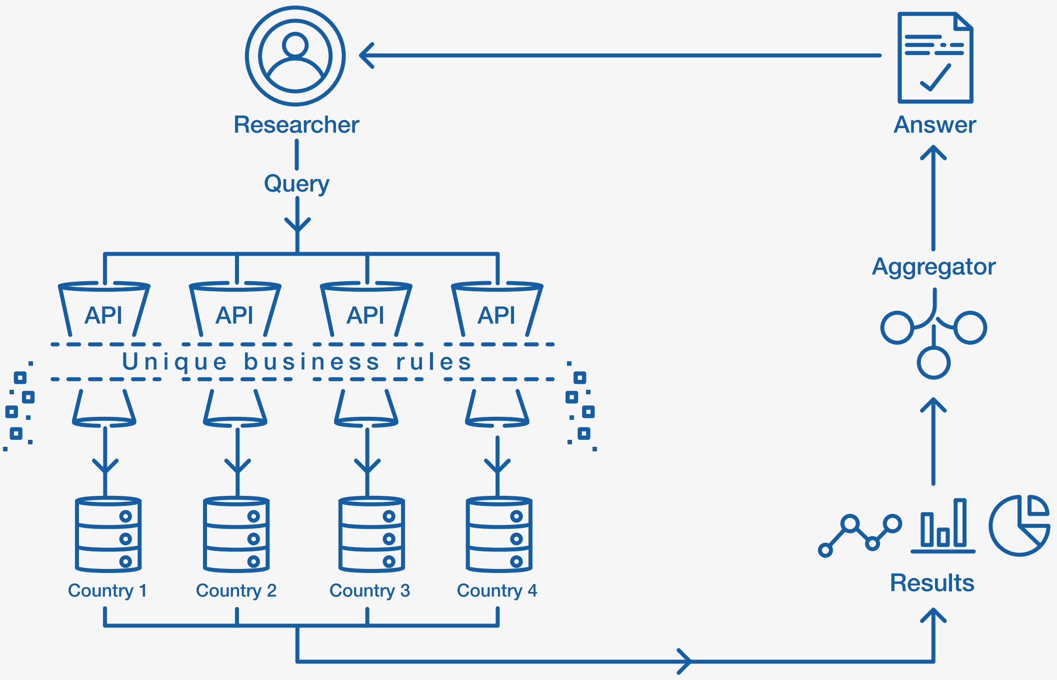
White Paper

Federated Data Systems:
Balancing Innovation
and Trust in the Use
of Sensitive Data

July 2019

What is the power behind a federated data system approach? What are its core capabilities to solve tough problems via expanded data set availability? In a Forum white paper, Federated Data Systems: Balancing Innovation and Trust in the Use of Sensitive Data, the technology for federating is described as a database architecture that solves a core problem in data: protecting sensitive data while still allowing such data to be used in innovative and trustworthy ways.[15] A federated data system uses multiple interconnected nodes, enabled by application programming interfaces (APIs), to provide secure yet open access to geographically disparate data systems and data formats (see Figure 3). A federated data system leverages and relies on a common architecture on nodes allowing for a common set of privacy, security, authentication and auditing features to enable all data sites to adhere to the same "rules" and core principles. A federated data system provides an opportunity to allow remote access to geographically distributed data sets while still guaranteeing that shared principles on security, interoperability and performance are consistently applied.

FIGURE 3 | Federating data using APIs



Elements of a federated data system

- A federated data system allows authorized users to perform queries on the data within a federated network of organizations. The results retrieved from each organization in the federation are then aggregated and returned to the individual who submitted the query. The data never leaves the organization that holds it. Instead, the data is "visited" and only the computed answers to the query are brought back to the federation system.
- Federated data systems use foundational, shared technology architectures, including operational components of security, auditing, authentication and access rights, among others. Agreement on which functions of this architecture are shared and which are left to local control is a critical component in setting up the federation that will allow access to the data.
- A central component of federated data systems is the use of APIs, which are managed using this shared technology architecture. The use of APIs and the foundational architecture enables a scalable, secure and reliable means of accessing the local data stores of the federated organizations, even though they likely use a variety of technology systems and data formats.
- Most importantly, the use of APIs allows the definition and enforcement of specific governance policies (including honouring local laws) by each organization within the federation. The use of APIs within a federated data system allows for crucial governance control to reside with each local entity in the federation, based on the overall agreement of the federation.

The benefits and constraints of federated data systems

Benefits

- Enable local control with global scale and efficiencies
- Addresses both privacy and security concerns
- Facilitate the ability to discover new data
- Enable the ability to analyse larger datasets to gain richer insights
- Reduce financial and operational costs
- Facilitate cross-border data sharing by respecting local governance and legal regulations

Constraints

- Add extra complexity to decision-making processes
- Create new types of "infomediaries" where the risks and liabilities are unknown
- Redress and remediation measures for accidents or acts of negligence with federated data are not yet developed

BOX 2 How other consortiums leverage federated data systems effectively

Other institutions like CanDIG, a data consortium in Canada, leverage a federated data system model to conduct health research using genomic data.[16] CanDIG allows researchers from across Canada to access and study national genomic data sets without violating provincial health data and privacy regulations. Each data access site handles its own data sets and users and controls who can access the federated data system and how often. A federated analysis is built and facilitated using APIs ensuring that the insights derived from localized data sets are shared and accessible across the consortium.[17] The common problem of a lack of access by external researchers or clinicians to genomic data from patients is solved by allowing researchers from across Canada to access and analyse genomic data without the data ever being copied or privacy being compromised.

The Global Alliance for Genomics and Health (GA4GH) is a non-profit alliance with the mission to accelerate progress in genomic research and human health by cultivating common frameworks and approaches for effective and responsible genomic and health-related data sharing.[18] The GA4GH provides many helpful frameworks and technical guidance for genomic-data-specific consortiums at www.ga4gh.org.

Step 3

Align on incentives and organizational capacity

For each institution participating in a data consortium, uncover why they are participating and their capacity to contribute

It is crucial that each data-owning institution in a proposed data consortium share with the group its incentives for joining and respective organizational capacities to contribute to the overall success of the consortium.[19]

3.1 Uncover different motivations or goals for participation

Having different incentives for wanting to participate in a data consortium is acceptable and should be expected. Often organizations will start with the goal of improving patient outcomes using genomic and health data. When pressed to further elaborate on what this means in practice, the Forum found that the actual institutional incentive and motivation for joining such a consortium can go well beyond the perceived benefit to patients alone.

As an example, three leading motivations to join the Breaking Barriers to Health Data global data consortium focused on more accurately diagnosing a rare disease are:

- **Discovery**: We need to increase the volume of data sets to solve rare-disease cases and diagnose diseases for those that otherwise would remain undiagnosed. Global coordination is essential to diagnose as many rare diseases as possible. Each country in the world simply cannot hold the volume of data needed to help every patient with a rare disease.

- **Improve and expand applications of genomics**: The success of using genomics to drive patient care is contingent on collecting large volumes of genomic data. Federating data sets in institutions or countries enables researchers to share the workload with access to the diversified data sets on which precision medicine relies.

- **Encourage international collaboration**: International collaboration is easy to talk about but much harder to execute with tangible outcomes. Institutions want to practise how to facilitate global genomic data access, starting small and building models that encourage replication, modification and sharing.

It is incredibly helpful to understand at the beginning of a prospective data consortium – before investments are made in building the data system – why each partner institution is motivated to participate. A consortium must be able to satisfy one or more goals of each partner, otherwise certain partners will not reap enough benefit to justify continued engagement.

Specific to the Breaking Barriers to Health Data project, each prospective institution initially thought its incentive to participate centred on patient best interest, but the underlying, long-term motivation extended beyond the “moral good” of such patient benefit. Joining a data consortium not only benefits the patient but also can achieve additional goals such as expanded international prestige or increased discovery in the research setting.

Understanding each institution’s capacity to realistically achieve its goal is essential to a data consortium’s long-term success. For instance, if an organization has limited size and volume of data, it may not yet be able to contribute sufficient resources to foster a successful consortium. Having the capacity to participate in a consortium requires that data collection and storage processes be optimized and finely honed, which is not a one-time, easy task but needs continued commitment and work.

3.2 Transparently share each institution’s ability to contribute to success

It is also imperative to uncover each individual institution’s capacity to uphold the goals of the prospective data consortium. For instance, it may be necessary to collect and store new types of data to achieve the goals of the other partners in the data consortium. Transparency, both in current data set characteristics and in future data set collection plans, will help the consortium understand if and when it can achieve its goals. Sharing future plans can also safeguard against one institution being the indefinite majority data owner. A data imbalance between data consortium partners creates a structural inequity and can lead to potential power conflicts at the decision-making level.

Collecting information from each prospective partner on future data-collection schemas and growth trajectories for data sets is an effective way to gauge current and future institutional capacity. If a prospective partner is only collecting a specific, niche type of data without plans to expand in volume or scope, it may not be able to help make a consortium successful. While a business plan, strategic plan or growth plan can suffice for securing accurate information on a given institution’s future data capacity, a data audit is likely necessary as well to understand the specific volumes of data that could be accessed in the short term in a data consortium.

Transparency, both in current data set characteristics and in future data set collection plans, will help the consortium understand if and when it can achieve its goals.

3.3 Perform a data audit with all prospective partners

It is crucial to understand the types of data each institution has capacity to share in the short term versus the long term, via a data audit. Whether via a survey or surveillance mechanism, each institution should share detailed information on both the type and volume of data already in its localized database.

As a general guide, data should be separated into categories showing at a minimum clinical data, genomic or omic data, and data from unaffected family members in their respective database. Within these categories, listed below are additional subcategories for types of necessary data.

FIGURE 4 **Potential tiers of data to include in a federated genomic data consortium**

Type of data	Sub-categories of data	Detailed examples to categorize sub-tiers of data
Clinical data	Clinician/researcher captured data, data contributed by patients, facial imaging data, linked family-member information	Family history, medical history, past medical interventions, genotype information, diagnosis, self-reported family history, phenotype data, socio-economic data, photographs
Genomic/omic data	Clinome, exome, genome, variants identified in genomes and exomes, transcriptome, metabolome, lipidome	*Requires standardization of raw read data to achieve interoperability but may be stored in an unstructured format*; labelling for genomic position, gene name and function, predicted pathogenicity, frequency of in population and control data sets
Data from unaffected family members	Clinical data, facial imaging data from family members, omic data	Same as clinical data and genomic/omic data but from family members who are unaffected or relevant for pedigree findings

This sample showing tiers of data was created by the Genomics4RD team in Canada under their pre-existing governance structure

Addressing the types of data *already* collected and stored by a given institution as well as plans to expand to *new* types of data is crucial to align expectations amongst potential partners. Differing types of data may not be a roadblock to a successful data consortium but it is important to recognize that the process of collecting, structuring and storing new data sets can be time-consuming. In the absence of a data audit, difference in types of data, if persistent in the long-term operations of a consortium, ultimately limits the potential for deriving new insights via federated querying and decreases the capability of a data consortium to deliver on its initial goals.

Step 4

Identify resourcing – team leadership and funding

After deciding to partake in a consortium, running a consortium requires a dynamic team and steady funding

Finding the right person (or team of people) within an institution to lead day-to-day consortium oversight and securing a steady funding source will contribute to the long-term viability of a data consortium.

4.1 Find internal champions within each partner institution

While support from the chief executive officer or leadership equivalent, as referenced in Step 1, is needed to create a successful data consortium, identifying and selecting the team of internal champions who will set up and run the consortium at each partner institution is also a necessity.

The process of setting up and running a data consortium requires acting amid ambiguity and inevitable internal roadblocks. Each internal champion (or a team of champions) per partner institution must know how to navigate the multiple branches, wings or teams within a given institution.

Participating in a data consortium requires work across multiple teams: policy and legal teams, technical teams, and research or clinical teams. This internal champion(s) also ideally carries enough institutional clout to drive forward decisions that would otherwise stall or completely hinder the creation of a data consortium.

There are not always clear “right” or “wrong” ways to proceed when participating in the creation of a new data consortium; the internal champion will need to make difficult decisions without always having all of the ideal information.

4.2 Secure a funding source to ensure continued participation

A clear funding source at each institution is also a necessity that can otherwise cause a data consortium to prematurely collapse before achieving its goals. Three streams of funding are needed: (1) to ensure data sets are structured and interoperable; (2) to build and implement an application programming interface, or API (further discussed in Step 8); and (3) to manage any data system upgrades or improvements in technical components. Funding does not necessarily need to be secured from an outside source but could be in the form of an internal investment or pledge to sustain the consortium operations at each institutional site.

If each prospective partner institution relies on different modes of funding (e.g. one private institution and one publicly funded institution), it may be useful to divide up funding responsibilities if funding will only be available to certain partners later. For instance, in the Breaking Barriers to Health Data project, one partner with a robust funding stream at the inception of the consortium built the initial federated data system API and distributed the API to the other partner institutions. In this approach, the other partners agreed to fund additional technical components and upgrades as soon as their additional funding sources were delivered.

4.3 Develop an economic framework to justify investment

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

White Paper

Global Data Access for Solving Rare Disease

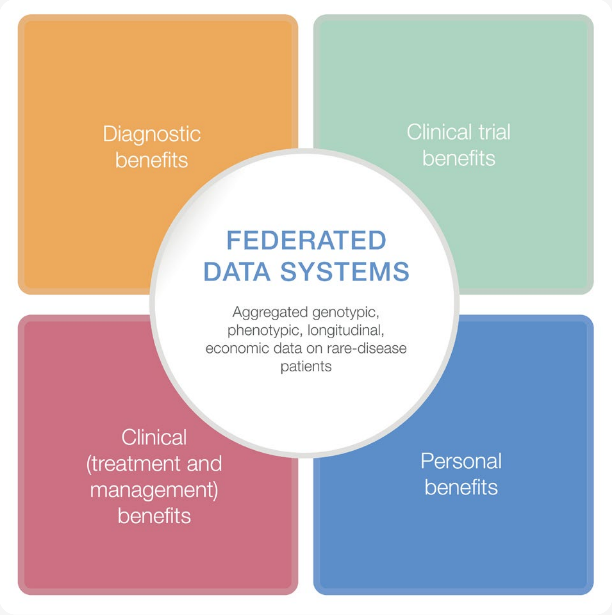
A Health Economics Value Framework

February 2020

If there are concerns regarding the potential for a long-term return on investment in a data consortium, economic models can be developed to estimate both the quantitative and qualitative returns of participating in a consortium. For the Breaking Barriers to Health Data project, the Forum created an economics framework for rare disease diagnosis and published Global Data Access for Solving Rare Disease: A Health Economics Value Framework in February 2020. The Forum's economic framework showcases the economic rationale for a data consortium and exemplifies downstream opportunities for financial returns when participating in a global data consortium.[20]

Successfully allocating internal funding and creating economic frameworks, however, does not guarantee long-term funding for the data consortium if a change in leadership occurs. Data consortiums often end due to previously allotted funding streams drying up when the leader or leadership team leaves. Thus, seeking outside funding or even setting up a separate entity to secure consortium finances can be the safest route. Nonetheless, creating, managing and running a separate legal entity specific to a data consortium adds additional work to the team selected to oversee the consortium's success.

FIGURE 5 **Potential areas for benefit and return-on-investment via participation in a cross-border federated data system for rare disease**


Diagnostic benefits
Clinical trial benefits
FEDERATED DATA SYSTEMS
Aggregated genotypic, phenotypic, longitudinal, economic data on rare-disease patients
Clinical (treatment and management) benefits
Personal benefits

Definitions of types of benefit

1. **Diagnostic benefit:** The identification of pathogenic or likely pathogenic variants in known disease genes.
2. **Clinical benefit:** Changes in the medical or surgical management of patients as a result of the diagnosis being made. These changes relate to improvements in health outcomes via assignment of therapies (therapeutic benefit) or improvements in the management of patients in the absence of therapy assignment (management benefit).
3. **Clinical trial benefit:** Changes related to the improvement of clinical trial operations.
4. **Personal benefit:** The presence of nonclinical outcomes that are important from a personal point of view to a person with a rare disease or who is affected by a rare disease. These outcomes may relate to the intrinsic value of information, the knowledge about the condition and the opportunity to make plans for the family or the future.

Step **5**

Identify institutional differences or gaps in policy

Differences in institutional processes are to be expected in leadership, legal, and technical teams

Steps 5 and 6 provide detail on how to develop a clear governance policy that will help the data consortium achieve its goals while also mitigating the risks that may emerge in the day-to-day operations. While it is expected that each institution entering the data consortium will operate differently internally – using different consent models or different data structuring norms – such differences must be uncovered early in the partnership via interviews conducted across multiple levels of the institution (leadership, technical, legal and beyond).

5.1 Partner with an outside, neutral organization to guide transparent discussions

For the Breaking Barriers to Health Data project, the Forum served as the impartial body to conduct interviews at each institution in person and, subsequently, share the answers in a standardized format with all prospective partner institutions. It is recommended to choose an outside body that is impartial, without ties to any prospective partner, to maintain neutrality in what can otherwise be difficult discussions about an organization's internal operations, capabilities and policies.

In the recommended in-person interviews at each prospective partner institution, an impartial organization needs to uncover three key areas of information, at a minimum:

1. Data collection and consent norms: How is the data collected? Do patients know why the data is being collected?

2. Operational norms and standards: How would each institution operationalize the consortium internally?

3. Technical standards: How are data sets managed to guarantee data security, data integrity and patient privacy?

The Forum developed a set of sub questions within these three categories to guide discussions and prospective partner interviews (Figure 5). It is important to ask *all* of these questions to the leadership team, technical team, legal team and beyond. For instance, it is expected that the legal team may not know the answer to technical questions but the answers they provide can illuminate any potential institutional knowledge gaps. Such knowledge gaps should prompt further discussions internally so that each institution is aligned on their technical and regulatory policies before entering similar discussions with prospective partner institutions.

All prospective partners of a data consortium should clearly share information on these three categories, including suggested detailed questions in the diagram below, as the answers guide the creation of the consortium's overarching governance framework (Step 6).

FIGURE 6 | **Suggested types of data required to establish a federated genomic data consortium**

A federation enabling access to sensitive health data should be guided by three goals: patient trust, ethical use of sensitive data, and trustworthiness between members. The operational components needed to achieve these three goals, however, are distinct. Each box below investigate central operational components that each partner needs to clearly answer in order for a federation to operate.

Component 1
How is data collected?

- Do patients know why the data is collected?
- Do patients agree to share the data?
- Do patients know why the data is being shared?
- Do patients understand how it is being shared?
- Do institutions communicate results as appropriate back to patients?
- Does data collection follow pre-existing regulations and rules?

Component 2
How are we going to operate as a federation?

- From whom will you accept a query?
- For what purpose will you use the data?
- What results will you allow to be returned?
- Who adjudicates disputes or ambiguous situations?
- What is the threshold to entry for a new member of the federation?

Component 3
What are the technical standards?

- How do you ensure security of queries within the transfer process?
- How is patient privacy protected?
- Do you make your build standards transparent?
- What is the change management process?
- What level of interoperability needs to be reached (API vs data sets)?
- What are your mechanisms for ensuring data integrity?
- How is data harmonization currently achieved?

These components recognize pre-existing norms in genomic data sharing, including:

The ability of all to 'share in scientific advancement and its benefits.' (Univeral Declaration of Human Rights, Article 27, 1948)

The priority of minimzing harm by respecting all persons and their human dignity. (Framework for responsible Sharing of Genomic and Health-Related Data, The Global Alliance for Genomics and Health)

5.2 Identify unforeseen differences in institutional practices

It is important to identify any crucial differences in data collection, operational norms, or technical standards that need to be discussed further or even modified before a group of partners can agree to work together via a consortium model. It will only be possible to build a governance model if each institution reaches a consensus internally on their answers and policy norms, and transparently shares this information with the group of prospective partners for the data consortium.

Many institutions will have noticeable gaps in policies where other institutions by contrast have a strict policy. For instance, in the Breaking Barriers to Health Data project, the Forum identified surprising differences in patient consent policies. While one institution had a sophisticated dynamic consent model fully deployed, allowing patients to select exactly how and when their data is used, another institution only had a broad consent policy that did not allow patients to voice preference on data use. This difference in policy prompted conversations internally on whether the institution with the broad consent policy would consider revamping and changing its policy later or preferred to keep its current policy. Differences in cultural norms can dictate differences in consent policies and consent policy preference. Yet facilitating these conversations can also help an institution improve its own policies by providing increased awareness of innovative new approaches implemented by global leaders in the same discipline.

The answers gathered to these key questions will ground the governance model, which is discussed in the next step.

Step 6

Create a consortium governance model

A custom governance model is achievable with strong foundational principles paired with detailed operating standards

A governance model dictates how the consortium will operate. A good consortium governance model provides (1) foundational principles to guide future decision-making or any instances of ambiguity and (2) clear standards to act as the "guard rails" to oversee the day-to-day operations and to make sure the consortium is effectively activated and used to achieve the group's shared goals (as established in Step 3). Overall, a governance model also helps maintain and reinforce trust amongst partners as the consortium grows and changes.

6.1 Leverage pre-existing policies

There are many policies that already exist for data sharing and even a few for sensitive health data sharing that institutions can draw from to create a data consortium. For instance, the GO FAIR Data Principles are commonly used as the industry standard for health data consortiums in Europe. GO FAIR is a bottom-up, stakeholder-driven and self-governed initiative that aims to implement the FAIR data principles, making data Findable, Accessible, Interoperable and Reusable (FAIR). Additionally, the GA4GH's Framework for Responsible Sharing of Genomic and Health-Related Data provides clear principles to consider when sharing sensitive health data.[21] Creating a governance model can take a hybrid approach by sourcing foundational principles from pre-existing frameworks and also by creating additional standards from scratch to be specific to a group of institutional partners.

6.2 Part 1. Decide on foundational principles to ground your governance model

The questions answered in Step 5 can guide discussions that will identify areas of agreement that are so unanimous they can be articulated in foundational principles to ground the governance model of the consortium.

For the Breaking Barriers to Health Data project, the Forum completed a simple grouping process, identifying where pre-existing policies at each partner institution were similar to one another and where pre-existing policies at each partner institution were different. The consortium adopted two specific sets of pre-existing foundational principles that aligned with their areas of agreement and can guide future consortium decision-making: the FAIR Guiding Principles for scientific data management and stewardship and Canada's Roadmap for Open Science.

Areas of disagreement required further discussion to decide where new common ground could be reached to include in the governance model. For instance, although one area of disagreement centred on how to handle intellectual property generated from discoveries via the consortium, partners on the Breaking Barriers to Health Data project agreed to adopt the "open science" principles articulated in Canada's Roadmap for Open Science with a slight modification. All members of the consortium agreed to share intellectual property and give credit to the consortium as a whole. Also, no patents will be sought on findings or derivatives of findings in order to adhere to the open science guiding principles. If an institution had a pre-existing intellectual property policy internally, that policy continued to stand for any discoveries made using data *outside* the consortium's data set.

6.3 Part 2. Create detailed standards that are specific to the consortium

> “As genomics and the wider healthcare ecosystem continue to change, it should be expected that respective institutional policies as well as the ‘industry standard’ policy will change.”

The goal when creating custom standards is to provide enough detail for all consortium partners to maintain consistency in how they each facilitate data collection and consent norms (Step 5 key area 1), manage operational norms and standards (Step 5 key area 2), and guarantee technical standards (Step 5 key area 3).

While developing the consortium’s standards, it is especially necessary to discuss if any pre-existing institutional policies are likely to change. As genomics and the wider healthcare ecosystem continue to change, it should be expected that respective institutional policies as well as the “industry standard” policy will change.

For instance, the return of results to patients is an area where consortium standards are expected to change for genomic consortiums. Each person’s genome has about 3 million-4 million genomic variants representing specific changes in their DNA sequence.[22] Yet deciding which genomic variants have what effect on an individual’s biology is difficult. There are five categories of variants: pathogenic (disease-causing), likely pathogenic, unknown significance, benign (not disease-causing), and likely benign. The American College of Medical Genetics and Genomics (ACMG) lists 59 variants we know are pathogenic with defined phenotypes and clinically actionable pathways that can improve patient well-being.[23] There is division in the global genomics community on whether or not “likely pathogenic” results or other findings that lack clinical actionability should be reported.

On the Breaking Barriers to Health Data project, consortium partners separated and defined (1) primary findings (relevant to main diagnosis using genetic testing; variant reported back is known to be pathogenic), (2) secondary findings (additional, health-related finding; genomic variant must be linked to serious conditions for which there is good evidence that knowing about such a condition could influence the delivery of healthcare such as BRCA1 or BCRA2), (3) carrier status (finding of an autosomal recessive genetic characteristic such as sickle cell anaemia or cystic fibrosis), and (4) incidental findings (secondary health-related finding not included in variants in three prior categories). Yet each institution currently has a different policy on whether or not each of the four categories is holistically reported back to the patient, which is also a common incongruence in countries.[24] As we increase education on the value of genetic testing, it is expected that patients will begin to expect to receive information beyond a primary finding and rather expect to also receive secondary findings or even carrier status findings. Such a development will require the consortium to develop new standards that apply to the consortium as a whole rather than each institution following a different policy developed internally.

A clear-change management process, with triggers for when standards may need to be revisited or what constitutes a significant enough change that the consortium needs to convene to make decisions, is also essential to include with clear standards in a governance model. A clear decision-making structure within the consortium should be set in place for this purpose.

FIGURE 7 **Sample federated data governance model**

Sample Federated Data Consortium Governance Model

The Breaking Barriers to Health Data consortium aims to be community-led, self-governed, open and inclusive across countries of operation. In particular, this consortium adheres to the GO FAIR Implementation Network, which encourages the creation of consortiums that are committed to defining and continuing to build specific tools to encourage the Internet of FAIR Data and Services (IFDS). This consortium also adheres to Canada's Roadmap for Open Science, which encourages a shared commitment to all stakeholders in science, transparency in data, inclusiveness, collaboration and sustainability.

The governance model details the following operating standards:

a. Data use and data access minimum ground rules
b. Data formatting standards and typologies
c. Data security
d. Patient consent
e. Benefit sharing
f. Intellectual property guidelines
g. Consortium membership responsibilities

We provide details below of one standard in this federated data consortium standards, as an example, on data use and access. To view the entire governance model from the Breaking Barriers to Health Data with detail on each standard, please see the Appendix.

Standard a: Data use and data access minimum ground rules

Having clear ground rules on how data is collected and what can be done with it is important to maintain patient trust, patient privacy and the security of the consortium. Consortium members agree that:

- Data is collected for research within the disease area of focus (e.g. rare disease); it is not used for querying additional ailments outside the agreed disease scope
- Data will be queried to achieve greater volumes of diagnoses for people with a rare disease
- Data will not to be exported. It will be visible to members of the consortium granted access to the federated data system but will not be exported, downloaded or otherwise duplicated without explicit patient consent
- Data relevant to the federated system will not be accessed by third parties without explicit, unanimous approval by all members of the consortium and without explicit consent from patients contributing data to the consortium
- Data will not be used for the purposes of identifying patients or attempting to identify patients in any circumstances. Failure to follow this principle will result in immediate and permanent removal from access to the consortium*

*Many people living with a rare disease can be easier to re-identify due to the low number of other individuals in the world with the same variants linked to a specific rare disease.

Step 7

Structure the data

Structured data ensures the data can be used as effectively and efficiently possible

In order to enable the consortium to achieve its aims and solve the problem defined in Step 2, institutions in a data consortium must make sure their respective data is structured in a way that can be queried by a federated data system.

On the Breaking Barriers to Health Data project, the Forum found that even after each institution completed and shared results from a data audit, there was a lack of clarity on how each institution's data was organized and annotated. For instance, while the technical team could explain why the data was structured and stored in a specific way, it remained unclear if that would continue to be the case or if additional data collected would be stored in a new way. Additionally, data sets collected in the past often weren't structured the same way as data sets collected more recently. Such structuring appeared to be largely dependent on who was in charge of a given institution's technical team and when, but also varied based on leadership changes and a lack of change management. All institutions were eager to strive for ways to improve their data structuring so the data could be used as effectively and efficiently as possible in additional analyses.

Structuring data can be time-consuming if data is not initially collected, structured or stored with an eye to remote access or sharing capabilities. Nonetheless, structuring data in a way that can be federated is a requirement to participate in a data consortium.

The Global Alliance for Genomics and Health (GA4GH) is the international technical standard-setting organization improving cohesion in sensitive health data structure. A common GA4GH data standard is included below as a recommendation for adoption in a federated data consortium using genomic data. Adopting data standards like the ones developed by the GA4GH or other leading data standard-setting bodies provides a basis for mutual understanding among people and organizations and increases the ability of disparate organizations to seamlessly connect and share multiplying genomic data sets.

FIGURE 8

Example Genomic Data Structure Standards

The GA4GH recommends that all genomic and health-related data be coded so that it can be: (1) anonymized by each genomics institution in a consortium; and (2) re-identified with the case of clinically relevant findings needing to be reported back. In 2019, the GA4GH developed **Data Use Ontology (DUO) codes**[25] that allow users to semantically tag genomic data sets with usage restrictions, so they can become automatically discoverable based on a health, clinical, or biomedical researcher's authorization level or intended use.

DUO has three main features:

1. DUO provides a shared understanding of the meaning of data-use categories. Each DUO term was developed with community consensus and includes a human-readable definition, which can be expanded by adding optional comments or example uses. This allows data stewards in different resources to consistently tag their data sets with common restrictions on how this data can be used.

2. DUO is distributed as a machine-readable file that encodes both how the data can be used (data-use categories) and how a researcher intends to use the data (additional terms that define intended research usage). This file is publicly available, versioned and written using the World Wide Web Consortium (W3C)-standard OWL Web Ontology Language and following Open Biological and Biomedical Ontologies (OBO) development principles. DUO-enabled data sets are automatically discoverable for secondary research within databases such as the European Genome-Phenome Archive (EGA) at the European Molecular Biology Laboratory (EMBL)'s European Bioinformatics Institute and the Centre for Genomic Regulation. A researcher can query EGA, or any database that has implemented DUO, and receive only data that matches his/her intended use and/or authorization level.

3. DUO can be implemented alongside an advanced search algorithm, such as the Broad Institute's Data Use Oversight System (DUOS), which allows authenticated users to query and gain access to data sets pertaining to their research. For example, an industry researcher working on cancer would be matched with any data set that is allowed *for commercial use* and *for cancer research* and offered the opportunity to fetch them automatically.

: DUO leverages and extends previous work, such as Consent Codes[26] and the Automatable Discovery and Access,[27] as well as all existing terms in dbGaP, the US National Institutes of Health (NIH) database of Genotypes and Phenotypes.

Step 8

Deploy the API technology

With the right partners, a clear governance model and the technical proof of concept, queries are ready to be sent

A federated data system uses multiple interconnected nodes, enabled by application programming interfaces (APIs), to provide secure yet open access to geographically disparate data systems and data formats. An API simplifies the ability to retrieve data from many types of databases and applications, including those at remote locations. Access to data via APIs enables permission-based access with different layers of granularity.

8.1 Implement the API to activate queries

Once the governance model is finalized and data is structured based on agreed standards, it is up to each institution to jointly adopt an API to begin sending queries via a federated data system. The technical team of each institution ensures that the APIs for federating data are programmed such that they operationalize the principles and standards agreed on in the consortium's governance model (Step 6). The technical teams of each partner institution in the consortium must strive to improve the data consortium with agreed, consistent system upgrades. As the goals and objectives of the consortium grow or additional partners join the consortium, changes will need to be made and a clear change-management process needs to be in place.

8.2 Track success with KPIs

Lastly, it is crucial that the data insights accessed via the federated data system and subsequent clinical or research findings be effectively tracked in line with agreed on key performance indicators (KPIs). Which KPIs can be traced or tracked are dependent on the build of the API, which makes deciding on such indicators difficult at a previous step.

Depending on the varying goals for joining a data consortium, specific success measures could be established and clearly tracked before the technology build, but only if the technical team has the capacity to be deeply involved in the governance model conversations, which is not always the case.

For example, in the Breaking Barriers to Health Data project, the consortium will track how many new rare disease diagnoses are made as a result of consortium membership as its first KPI. From there, additional metrics for success may include the economic impact of clinical discoveries made via the consortium, such as the financial savings of shortening a rare disease patient's "diagnostic odyssey". If federating data allows for faster and more precise identification of patients to enrol in clinical trials, clear returns on investment for pharmaceutical development could be demonstrated.

Many consortiums may not be able to decide on KPIs until after the consortium is up and running given the time allocation required to create a governance model and deploy the API technology, yet a consortium will only be able to track how effectively it is achieving its goals if such considerations to track performance, however possible, are made on the technical side of the data system.

Conclusion

Participating in a sensitive health data consortium is the only way to maximize volumes of data already collected, sitting in silos around the globe. However, creating a consortium in practice requires an ongoing process with months of informal negotiations ultimately resulting in the creation of a clear governance model and a well-functioning consortium.

From finding trustworthy partners (Step 1) to determining a common problem where federating data is beneficial (Step 2), to aligning on incentives and capacities (Step 3), to identifying resourcing (Step 4), to designing and deploying a governance model (Steps 5 and 6), to structuring data (Step 7), to deploying the API technology (Step 8), creating a new health data consortium requires a custom process to ensure success and long-term viability.

Wanting to join a data consortium given the benefits of gaining access to an increased volume of data may be an enticing offer, but to actually set up and run a data consortium, especially a distributed one across country borders, requires time, grit and difficult conversations to arrive at a specific but adaptive governance model.

When the Forum set out to set up and test a federated data system model specific to genomic data for the Breaking Barriers to Health Data project, not many guidance documents existed. This document represents the learnings and takeaways from the project. It is our hope that this guide will spur international collaboration in the global patient interest and encourage additional health data consortiums to consider and adopt governance mechanisms.

As health data continues to unlock new innovations and also produce new risks to patient privacy or data security, it is impossible to implement a singular policy safeguarding against all the potential hazards of participating in a health data consortium. Yet it is possible to take the consortium governance model development process seriously to create and encourage a cohesive, symbiotic relationship between institutions with otherwise differing models of consent, operations, security and technology; it is possible to optimize for the best outcomes possible to optimize for specific outcomes by being intentional in policy.

As data continues to play a central role in science discoveries, medical discoveries and beyond, it will be the institutions participating in collaborations and consortium initiatives that are able to lead the way in innovation to improve outcomes for patients globally.

This paper is part of a series by the World Economic Forum's Centre for the Fourth Industrial Revolution focusing on data policy in a post COVID-19 world.

Appendix

These documents are available on the World Economic Forum's Breaking Barriers to Health Data project page.

Governance model for Breaking Barriers to Health Data Consortium

Federated Data Systems: Balancing Innovation and Trust in the Use of Sensitive Data

Global Data Access for Solving Rare Disease: A Health Economics Value Framework

Breaking Barriers to Health Data one-pager

Acknowledgements

The World Economic Forum's Platform for the Future of Health and Healthcare and Precision Medicine team would like to extend their appreciation to the Australian Genomics Health Alliance, Genomics4RD, Intermountain Healthcare and Genomics England for undertaking the challenge of establishing a cross-border federated data system for the benefit of people living with rare disease globally. We also thank our entire project community, which includes over 80 subject-matter experts from industry, academia, government, civil society and beyond.

We would like to thank our reviewers:

Tiffany Boughtwood
Australian Genomics Health Alliance

Jamie Ferguson
Kaiser Permanente

Brad Power
Reengineering Cancer Treatment

Christina Waters
RARE Science

Durhane Wong-Rieger
Rare Diseases International

Author

Lynsey Chediak
Lead, Precision Medicine, Shaping the Future of Health and Healthcare, World Economic Forum

Endnotes

1. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data
2. The US National Library of Medicine provides an excellent overview of DNA and the role of "bases": "DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 % of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences." https://ghr.nlm.nih.gov/primer/basics/dna
3. The Human Genome Project estimated that humans have between 20,000 and 25,000 genes. Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes (less than 1% of the total) are slightly different between people. https://ghr.nlm.nih.gov/primer/basics/gene
4. https://www.sciencedirect.com/science/article/pii/S1532046413001007; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5958914/;
5. https://www.sciencedirect.com/science/article/pii/S1532046413001007#b0060
6. The full list of governance gaps preventing wider adoption of precision medicine are: (1) data-sharing and interoperability, (2) ethical use of technology, (3) patient and public engagement and trust, (4) access, delivery, value, pricing and reimbursement, and (5) responsive regulatory systems. Read more in the full report here: http://www3.weforum.org/docs/WEF_Global_Precision_Medicine_Council_Vision_Statement_2020.pdf
7. To read more on the return-on-investment of aggregating genomic data with a focus on a rare disease use case, read the World Economic Forum's "Global Data Access for Solving Rare Disease: A Health Economics Value Framework": http://www3.weforum.org/docs/WEF_Global_Data_Access_for_Solving_Rare_Disease_Report_2020.pdf
8. The Global Alliance for Genomics and Health (GA4GH) is internationally recognized for spurring the creation of several genomic data consortiums. The Beacon Project (https://www.nature.com/articles/s41587-019-0046-x), CanDIG (https://www.distributedgenomics.ca) and Matchmaker Exchange (http://www.matchmakerexchange.org/) are excellent examples of federated data systems. To view a comprehensive list of international genomic data initiatives, many of which are also federated data systems, visit: https://www.ga4gh.org/how-we-work/driver-projects/
9. Rare disease was chosen as the case study for the project due to the "diagnostic odyssey" by which people with a rare disease routinely wait 5-7 years on average for a correct diagnosis. Genomics as a diagnostics tools provides an opportunity to decrease this number
10. "Trust: Data: A New Framework for Identity and Data Sharing"; https://trust.mit.edu
11. "Research: How to Build Trust with Business Partners from Other Cultures", Harvard Business Review, https://hbr.org/2020/01/research-how-to-build-trust-with-business-partners-from-other-cultures
12. https://www.weforum.org/projects/breaking-barriers-to-health-data-project
13. Menzel, O., 2019. Value Creation from Diagnosis to Custom Drug: The Rare Disease Example. Black Swan Foundation, 4–6 December; Facts and Stats about Rare Disease, World Rare Disease Day: http://globalgenes.org/wp-content/uploads/2015/12/2016-WRDD-Fact-Sheet.pdf (link as of 2/2/20)
14. Facts and Stats about Rare Disease, World Rare Disease Day: http://globalgenes.org/wp-content/uploads/2015/12/2016-WRDD-Fact-Sheet.pdf (link as of 2/2/20). While this figure is widely referenced by global patient advocacy organizations, the publicly available epidemiological data in the Orphanet database contains information on 6,172 unique rare diseases, with 71.9% genetic in origin and 69.9% exclusively onsetting in paediatric patients
15. http://www3.weforum.org/docs/WEF_Federated_Data_Systems_2019.pdf
16. https://www.distributedgenomics.ca
17. Building a federated analysis is specific to each data consortium. APIs allow for the movement of insights derived from data sets without the localized data leaving its location of origin, but the process of analysis (often called federated learning) itself is reliant on the specific API build and the type of data offered in a data system.
18. https://www.ga4gh.org/aboutus/
19. In some instances, the "data owner" may not be a research or clinical institution but the person who shared her or his data in the first place. This guide is intended to aid institutions that have clear consent processes and retain ownership of their data sets. For one example of a new data ownership model, see LunaDNA: https://www.lunadna.com
20. http://www3.weforum.org/docs/WEF_Global_Data_Access_for_Solving_Rare_Disease_Report_2020.pdf
21. GO FAIR Principles: https://www.go-fair.org/fair-principles/ ; GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data: https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/framework-for-responsible-sharing-of-genomic-and-health-related-data/
22. https://www.genome.gov/news/news-release/Genomics-daunting-challenge-Identifying-variants-that-matter
23. https://www.coriell.org/1/NIGMS/Collections/ACMG-59-Genes
24. https://www.nature.com/articles/gim2017157.pdf?origin=ppub
25. DUO, Ebispot: https://github.com/EBISPOT/DUO (link as of 30/03/20)
26. Dyke et al., "Consent Codes: Upholding Standard Data Use Conditions", PLOS Genetics, 21 January 2016: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005772 (link as of 01/04/20).
27. Woolley, Brookes et al., "Responsible Sharing of Biomedical Data and Biospecimens via the "Automatable Discovery and Access Matrix" (ADA-M)", *npj Genomic Medicine*, 3 (17), 23 July 2018: https://www.nature.com/articles/s41525-018-0057-4 (link as of 01/04/20).

WORLD ECONOMIC FORUM

COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744
contact@weforum.org
www.weforum.org