Data Perspective in Digital Platforms: Three Tales of Genetic Platforms

Sirkka L. Jarvenpaa
McCombs School of Business,
University of Texas at Austin, USA
sirkka.jarvenpaa@mccombs.utexas.edu

M. Lynne Markus
Bentley University, USA
mlmarkus@bentley.edu

Abstract

Digital platforms play a critical facilitating role in the "changing models of biomedical research" and clinical care. Such platforms integrate disparate data sources and formats—including genetic, health, genealogical, and increasingly lifestyle data—into more accessible, searchable, and computationally efficient structures for basic scientific research, as well as for clinical care. Genetic platforms involve unprecedented data management challenges because of their scale and multidimensionality. Still, little research has been conducted on genetic platforms. Leveraging secondary data on three interlinked genetic platforms, we pursue a data perspective on platform evolution and entrepreneurial strategies. We contribute to the discussion on the design and evolution of digital platforms that considers responsible data use.

1. Introduction

"Genomics is at the crossroads where data and biology meet."
Li Ge, chairman of WuXiNextCODE, 2017.

During the twentieth century, digital technologies fundamentally disrupted businesses and social communities. During the twenty-first century, biology promises to alter life, death, and their pathways. Genetics opens doors for fundamental changes in the ways we understand diseases and their mechanisms. But these revolutions in biology at large, and in genetics in particular, are closely intertwined with disruptions in digital technologies.

Genetic platforms are enabled by digital technologies. In genetics, data represent the *blood line* that makes possible new research discoveries and that enable new treatment possibilities and care solutions in clinical practice. Genetic platforms facilitate the collection, storage, and analysis of large-scale genetic data with phenotypical and behavioral data. Since the mapping of the human genome project completed in 2003, data centric models have begun emerging that offer a compelling complement to basic scientific biomedical research and to the classic drug development models, as well as to care delivery [7]. Genome-wide association studies require access to large-scale genetic databases that allow the comparison of people who have a particular disease with those who do not have the disease. Personalized precision care requires the ability to integrate genetic data with detailed phenotypical data from medical records and family history. And consumer personal genetics companies like 23andMe are giving individuals direct access to their own genetic data, marketed as a form of entertainment about family history, which can be informative about an individual's health risks [27, 29]. Annas and Elias [2] predict that in a few years, "a majority of health plans will make it easy for their members to have their entire genomes sequenced and linked to their electronic health records and will provide software to help people interrogate their own genomes, with or without the help of their physicians or a genetic counselor supplied by the health plan."

Although genetic platforms populate daily news headlines, they have attracted little interest apart from the genetic data controversies and disputes related to medical, ethical, legal, security, and privacy concerns [3, 4, 30]. The concerns with data include informed consent, information privacy, the right to withdraw consent, the obligation to give feedback to study participants, benefit sharing arrangements, secondary uses of genetic data, and possible access to data by governments, public safety authorities, and insurance companies [5,10]. Although these issues are clearly important, the platforms themselves have drawn little interest. Even when the objective has been to commercialize genetic data, the focus has been on the loss of trust, restrictions on researcher and public access, and conflicts of public and private interests. The role of digital platforms has not been explored.

URI: http://hdl.handle.net/10125/50466
ISBN: 978-0-9981331-1-9

Meanwhile, the information systems and technology management literature has focused on digital platforms but not on genetic platforms and not on data more broadly speaking [13, 32]. Development toolkits that give access to and analyze data streams are discussed in terms of their applications and the size of the platform's potential market. But the toolkits and applications, rather than the data, are in the limelight. In health information technology literature, digital platforms and infrastructures have been examined in terms of coordinating access to care, facilitating knowledge transfer, and improving operational efficiencies [23]. What has been missing is the role of digital platforms in cultivating new scientific discoveries and treatment knowledge.

Admittedly, well-funded "triple-helix" partnerships involving government entities, corporations, and universities have tried to build genetic databanks for the scientific community, but some of these efforts have faltered in the face of restrictive national regulations and relentless controversies [30].

Framingham Genetic Medicine (FGM) provides one case in point [19, 24, 26, 31]. FGM was a for-profit venture with 20% ownership by Boston University. FGM aimed to digitize data from the Framingham heart study, conducted by Boston University under contract from the National Heart, Lung, and Blood Institute. Starting in 1948, the study gathered panel data on more than 5,000 residents of Framingham, MA. Boston University continued the study after the NIH contract expired; an additional 5,000 Framingham residents were recruited. Through the years, the study yielded valuable research results and a vast quantity of potentially valuable but largely manual data about the subjects' health and lifestyles. The FGM venture was formed in 2000 to digitize the data. FGM planned to conduct genetic "linkage studies" "similar to those being done by Gemini Genomics (Cambridge, UK) and deCODE Genetics (Reykjavik, Iceland)…" [16] and to fund the entire effort by providing data access to pharmaceutical firms for drug discovery research. FGM failed because NIH insisted that genetic data collected using public funds had to remain open for use by other researchers--a condition that would preclude a period of privileged access for fee-paying drug companies. Concerns about privacy were also raised, although the dispute about exclusive access to data is what ultimately resulted in the collapse of the effort.

Other failed efforts include the UmanGenomics in Sweden. The effort involved a grant of exclusive commercial rights (but not exclusive access) to an existing, publicly owned, research biobank; and consent for such use and access was secured from the participating individuals. New consent had to be secured for individual projects that exceeded the scope of the existing consent. Despite being heralded as a model of ethical conduct, fights over intellectual property rights brought the venture down [30].

Other more recent partnerships are following an open science and drug discovery approach (including genetic analysis), such as the effort at the Montreal Neurological Institute at McGill University. Here, researchers maintain the intellectual property rights to their research outputs [28].[1]

Publicly funded efforts increasingly are complemented by entrepreneurial firms with bold initiatives to build genetic platforms. Although left unexamined, these efforts are socially significant and important. The efforts can have ramifications beyond healthcare because they involve unprecedented big data management challenges, the application of artificial intelligence, and a whole host of social, legal, and ethical issues that together shape the platforms and their evolution.

2. Related Literature

Before examining three interlinked tales of genetic platforms being built by entrepreneurial firms, we briefly review selected literature on digital platforms and genetic data banks.

2.1. Digital Platforms

Digital platforms provide a shared set of services and architecture. The architecture includes technological modular systems and multiple actors in "multi-sided" roles [36, 38]. Digital platforms draw on various digital infrastructures, such as cloud computing and data analytics. Platforms take on many forms, but here we focus on platform ecosystems that are "more complex than either a product family or a multisided market" [35]. The study of platform ecosystems is important because these systems can lead to new markets, new industries, or in the case of science, new knowledge domains or even specialties. Recent research highlights how digital platforms can facilitate opportunity formation, creation, and scaling of entrepreneurial ventures [8, 18, 42].

Research on digital platforms has taken either a market or technological perspective [13]. A market-based perspective starts with a focus on *demand* and examines transactions, network effects, and competition; value is created from matching supply and demand and pricing. The focus is on competition between platforms and how economics of scope in

[1] (http://www.mcgill.ca/neuro/open-science-0/open-science-platform)

demand can create value. In the market perspective, the primary role of the platform is as a coordinating device (IOS and Android platforms are classic examples), and "the existence of the platform itself is also taken for granted, exogenous and unchanging" [Gawer, 2014, p. 1241]. Although the possible competition and collaboration between platforms is recognized, how that competition shapes the emergence and evolution of platforms is rarely examined.

A *technological* perspective takes a supply perspective and focuses on stable components, such as modules and functions in technological architectures. Value is created from reusing components for new combinations and other forms of co-creation that increase the growth of offerings. Variety in the innovation process expands the economies of scope and generates greater value through the platform. Just as in the market perspective, the platform is a coordinating device, but the platform adds value on the supply side rather than the demand side, promoting innovations among the technology development community through various toolkits and application programming interfaces (APIs). Here again, the IOS and Android platforms, as well as various maker platforms such as Shapeways, serve as examples.

Similarly to the market perspective, the technological perspective provides little insight into the emergence and evolution of platforms over time. When the focus is on evolution, it is limited to what happens with specific components or modules and does not consider the platform overall.

Neither the market nor the technological perspective focuses on data except as something enabled by the APIs or toolkits. In their review paper on digital platforms, Schreieck et al. [32] state that "no article explicitly analyzes the role of data as a boundary resource in platform ecosystems." The authors found this lacuna surprising because so many digital platforms are fueled by data sales.

Admittedly, IS research at large is not devoid of a data perspective. Some research examines the creation of business value from large-scale and real-time digital data streams [25]. However, the focus is on specific applications and the effect of data streams on specific firms, rather than on platform ecosystems. Existing research also examines organizational data supply chains from the legal and societal perspectives, including privacy, ownership, and security [21]. Where data have been the focus, attention has been directed mainly to open government data or to data governance.

2.2. Genetic Databanks

Data governance was the focus of a study by Vassilakopoulou, Skorve, and Aanestad [39] on two different breast cancer genes. The authors chronicle the emergence of data repositories that involve varying governance based on public, private, and walled garden models. The oldest initiative was set up as a public commons to further the goal of open sharing of all existing datasets; in this initiative, "registration was open to all and access to registered users was unrestricted" [39, p. 7]. However, major labs stopped contributing, claiming that the quality of the data in the repository was poor. Meanwhile, the initiative involving private control has become the world's largest service—at least partially because of its use of multiple methods, advanced infrastructures, and rapid testing procedures.

The study is important because it can be used to begin to extrapolate a data perspective to complement the market and technological perspectives in the digital platform literature (see Table 1). The platforms are conceptualized by Vassilakopoulou et al. as *databanks, data commons,* and *data repositories*. Vassilakopoulou et al. emphasize *discovery* and advances in scientific knowledge (e.g., cancer biology) and clinical knowledge (e.g., better diagnoses of cancer susceptibility) as the key goals. The value is created through *large-scale* and *varied data* on inheritance and environmental influences from *diverse sources*.

TABLE 1. Perspectives on digital platforms

Literature[2]	**Economics**	**Engineering**	**Science (Genetics)**
Conceptuali-zation	Platforms as markets	Platforms as technological architectures	Platforms as databanks/ repositories
Perspective	Demand	Supply	Knowledge
Focus	Competition	Innovation	Discovery
Value created through	Economics of scope in demand	Economics of scope in supply and innovation	Large-scale data, varied data sets, diverse data sourcing
Role	Coordinating device among buyers	Coordinating device among innovators	Coordinating and quality control device among scientists/ clinicians
Empirical setting	ICT	Manufactur-ing and ICT	Personalized medicine

Although both market and technological perspectives view platforms as fixed and stable at the broader platform level, Vassilakopoulou et al. shed light on the evolution of databanks, including how one databank stimulates the growth of another and how databanks compete. Evolution is influenced not just by the arrival of new actors and their datasets but also by

[2] The columns of economics and engineering are adapted from Gawer (2014).

sociotechnical design decisions. Decisions that affect data quality influence the evolution of platforms.

Vassilakopoulou et al.'s study also raises many other issues, such as intellectual property protection and legal and ethical concerns that shape the databanks. Technology is largely in the background, although the importance of computational techniques in improving data quality, access, and analysis and in increasing the benefits to the researchers is acknowledged. However, fragmentation of the data persists because of the use of home-grown protocols, processes, and tools for gathering, storing, and interpreting genetic data across different research groups. And data sharing in the research communities remains selective.

Other research on genetic databanks echoes concerns over quality of genetic data. Lee [20] points out that the likelihood of errors in genetic data is relatively high, "accentuating the need for manual oversight and verification…. Correcting, updating, and adding value to existing data records remain critical challenges." Data quality issues become even more complex when health and medical records need to be merged with genetic records to support clinical research and, ultimately, clinical practice. For instance, Gainer and Cagan [12] report that the codes used by clinicians to designate patient diseases in electronic medical records often describe possible rather than definitive diagnoses: They primarily serve administrative and billing purposes and might not be accurate enough for research purposes. For example, the Partners Personalized Medicine initiative required a sizable data science effort to develop algorithms for proper disease classifications of medical records, according to reports.

Below, we explore further the data perspective by examining three interlinked entrepreneurial initiatives to advance genetic platforms.

3. Three Tales of Genetic Platforms

The initial focus in our study was on controversies surrounding genetic databanks and their commercialization. We followed newspaper articles in regional and national newspapers. The articles caught our attention because we were already studying issues such as data protection and data responsibility. DeCODE Genetics was acknowledged as a bold scientific venture that had gone farther than any other in the commercialization of genetic data, and it had become a reference point for most discussions of genetic data commercialization [41]. Following up on the deCODE story led us to NextCode and WuxiNextCode. We then searched for the customers of NextCode and found our third platform example. Hence, the platforms were not chosen randomly but instead resulted from an inquiry that followed the principles of the snowball method. To understand the entrepreneurial genetic platforms, we relied primarily on secondary data. The data perspective emerged in our study when we triangulated our analysis with the existing literature on digital platforms.

3.1 deCODE: Genetic population database for Icelanders

This entrepreneurial venture had its start in 1996 in Iceland when deCODE's iconic founder, Kari Stefansson (KS), formerly a neurologist at Harvard Medical School, received $12 million from seven U.S. venture capital (VC) firms to build a trio of linkable databanks that leveraged Iceland's wealth of medical, genetic, and genealogical information. The building of three linked databases was an ambitious and high-risk vision to generate new scientific discoveries, drugs, and treatments. Hence, the company's market entry point was data.

Initially, deCODE Genetics sought to control access to the databases it built about Icelandic citizens using government financial support. deCODE's founder justified the exclusivity arrangement by referencing the expense and commercial risk of genetic and drug discovery research. deCODE's arrangement with the Icelandic government gave the company an interest in any commercial product resulting from research using the data [22]. The plans were to sell data access and research to pharmaceutical companies.

The first task involved turning Iceland's genealogical records into searchable form. These records stretched back 1,000 years and were in the public domain. In addition to automating genealogical records, deCODE planned to build the first population-wide genome database in the world by collecting samples from the entire population. In addition, the new venture proposed to automate the country's medical records.

Hoffman-La Roche Pharmaceuticals provided initial venture funding and bought the rights to manufacture any drugs developed by deCODE. deCODE filed an IPO in 2000 but had mixed results because of the international controversy that then surrounded its data plans. KS explained: "Because we were a commercial entity from the start, we had both the regulatory entities and the scientific community in Iceland and abroad concerned" [1].

To build the proposed (but never completed) health database, KS solicited the help of Iceland's legislative entity, the Alþingi. He convinced the entity to pass a statute authorizing the transfer of personal medical records (dating back to 1911) from doctors and health

centers to deCODE. In the process, deCODE would computerize Iceland's system of medical records [14].

In 1998, deCODE received from the Icelandic government a 12-year exclusive license for construction and commercialization of the database. The licensing terms included deCODE's exclusive access to the database, exclusive rights to generate findings from it, and the right to sell the findings to parties chosen by deCODE. According to the statute, only assumed, rather than informed, consent was required from citizens. Individuals could opt out by filling out official forms, but any data already in the database would not have to be removed [9, 11, 14].

Concerns over ethics, privacy, and security generated more than 700 news articles on deCODE in Iceland alone, and many more overseas. The exclusive privatization agreement authorized by Iceland's government authorities led to outcries from researchers concerned about their future access to the data. A statement from the chairman of the Icelandic Medical Association's ethics council read [33]: "When you put genealogical information into the data bank and also genetic data, then the data bank knows more about you than you know about yourself." Many members of the medical community refused to turn over their patients' health records. The Data Protection Commission was not convinced of the adequacy of its security. Citizens as well as concerned parties outside of Iceland were angered by the lack of informed consent protocols. The storm culminated in 2000 when a 15-year-old girl filed a lawsuit because her dead father's medical records were to be entered into the database, which she considered a violation of her privacy. The lawsuit triggered the Icelandic Supreme Court to rule the Health Sector Database Act unconstitutional. The court ruling halted the further construction of a centralized health database.

When deCODE's centralized database plan failed, the firm switched to a distributed approach that leveraged individual research projects and their data requirements to collect data samples. deCODE enlisted the cooperation of the informal owners of the relevant health data by inviting local physicians to participate as researchers in its projects. These physicians then brought their patients to the studies. The company ran tens of research studies in parallel "under the strictest standards of informed consent" [41, pp. 94-95]. The firm reported 95% participation rates in the studies, and 90% of participants signed the broader consent form. Such high rates of participation were unheard of around the world, including in the United States. By 2002, statements were made suggesting that, to some extent, "a [health] database now exists inside deCODE" [41]. These developments also were aided and supported by those in Iceland who donated their blood samples. Encouragement to participate was seen by some as patriotic in building a biological powerhouse in the North Atlantic, while others viewed the high levels of participation as indicative of "coercion" because many Icelandic citizens had heavily invested in deCODE shares [4].

During the mid-2000s, the company sought to become a full-fledged biotech company. Continuing to build its downstream capabilities, deCODE partnered with both Merck and Bayer. The company also invested heavily in its technological capabilities. It partnered with IBM and strengthened its computing and data mining technologies.

In the late 2000s, however, investors in deCODE grew impatient. The firm filed for bankruptcy in 2009, selling data access and technological assets to another private entity, while much of the company's management team remained intact. The startup focused on selling direct-to-consumer genetic testing kits to accelerate the collection of data samples, but it encountered significant pushback from the medical community and regulators. In 2010, deCODE downsized from 750 people to 125 people. All downstream activities were sold off. The company focused on basic scientific research.

In 2012, Amgen acquired deCODE. With Amgen came independence, stability, and the financial resources needed to focus on fundamental research that would leverage access to the population-based data of about 120,000 Icelanders. According to the editor of *Nature Genetics*, some 5% of the journal's cumulative articles since 2000 have been authored by deCODE researchers. deCODE itself claims to have published more than 400 articles across various outlets [1].

In 2013, data and research activities were separated from proprietary technologies, and the latter were incorporated into a venture called NextCODE. The new platform venture was legally and commercially separate from the access to data. The data were owned by the Icelandic government and had to remain in Iceland.

3.2. WuXiNextCODE: Global Platform of Open Data with Proprietary Infrastructures

In 2015 NextCODE merged with a division of WuXi,[3] a Chinese contract research company, to form WuXiNextCODE. WuXi provided access both to large farms of sequencing machines and to the Chinese market.

In the two years since the merger, WuXiNextCODE has built the leading large-scale, integrated, global genetics platform, with strong cloud-based computing,

[3] WuXi is used here to correspond to WuXiPharma and WuXiApptech.

deep learning, and elastic relational database infrastructures. The platform is positioned as a bridge between research and clinical care, and the goal is to expand the overlap between the two. To turn genetic data into actionable knowledge for a person's treatment plans can require linking millions of different individuals' genomes. Genome sequencing and research feed directly into clinical care, and clinical care feeds comprehensive patient health information back into research.

WuXiNextCODE's proprietary technology includes a relational database architecture and an artificial intelligence engine. It is based on streaming data and is unrivaled in its efficiency in the storage and processing of genomic joins. The technology is versatile in accessing data in varied formats, including from web pages. The platform offers a workbench and interactive query tools for researchers and clinicians.

Although the platform's infrastructure technologies are proprietary, the platform promotes global sharing through open data access to any registered user. The platform manages the largest genome cohort database in the world. Hence, WuXiNextCODE has already accomplished what its rivals, including the FDA and the National Institute of Standards and Technology, have been trying to achieve for years.

The WuXiNextCODE platform has enabled new research collaborations and expedited clinical trials across projects and countries. For example, the platform collaborates with Huawei and its "China precision medicine cloud."[4] The platform is now in use in population genomics projects, precision medicine applications, and clinical diagnosis and wellness in China, England, Ireland, the United States, Qatar, and Singapore. In the United States, the platform is in use at Boston's Children's Hospital and Cincinnati Children's Hospital. In its partnership with Shanghai Children's Hospital, the WuXiNextCODE platform sequences some two million genomes annually. The company offers several products in Chinese markets, including a whole-genome wellness service for Chinese consumers. The company currently is going through an IPO filing in China.

3.3. START: Open source genetic research database with voluntary resources

In 2007, Anthony Tolcher left the UT Health Science Center in San Antonio to start up South Texas Accelerated Research Therapeutics (START) with two other colleagues. Today, START is one of the largest oncology treatment practices in the San Antonio area and a leading independent (i.e., unaffiliated with an academic medical center) cancer research and drug development center. In fact, START (with centers also in the upper midwest, Europe, and Asia) has arguably become the world leader in Phase I clinical trials for oncology—an area of activity long dominated by academic medical centers.

As an unaffiliated cancer center, START had access to tumor samples only from its own patients. Tissue samples in tumor banks are held by academic medical centers and are available only to researchers at those centers. (The U.S. National Cancer Institute operates open-access tissue banks, but these repositories lack comprehensive healthcare data.)

To increase its access to tumor samples for clinical research, Tolcher and his colleagues at START announced in 2010 the establishment of a San Antonio cancer tumor bank, funded by private donors. Unlike affiliated tumor banks, the START's tumor bank was to be open access—meaning that cancer researchers anywhere in the world could gain access to START's "consented" tissue samples (i.e., samples from patients who had given appropriate written consent). The open access nature of the START tumor bank was intended to encourage cancer researchers around the world to contribute their patients' consented tissue samples, thus accelerating research.

Developing the tumor bank required substantial investment. Communicating the idea to the local oncology community took time. In addition, START had to develop new procedures for obtaining patient consent, obtaining associated medical records data, collecting and transporting samples to the bank, and releasing samples for preclinical research. All these efforts were successful. Since 2010, START's tumor bank has become the world's largest repository of samples of a certain rare cancer; it receives samples weekly from around the world. In addition, START researchers have published widely in the cancer research literature and have received numerous prestigious awards.

The success of START's tumor bank positioned it well for its next major donor-funded initiative in 2012. Known as the San Antonio 1000 Cancer Genome Project (SA1KCGP), it is an open access database of genomic data from 1,000 consenting patients. After an early partnership with Beijing Genomics, Inc. (BGI), START signed on with WuXiNextCODE for low-cost, high-volume genetic sequencing, informatics support, and cloud data storage. In early 2017, START was more than halfway to its goals in terms of the number of samples collected and sequenced and the funds raised for genetic sequencing and analysis.

[4] http://www.bio-itworld.com/2016/5/24/wuxi-nextcode-huawei-launch-precision-medicine-cloud-china.aspx

High-quality medical records data, when linked with genetic samples and data, offer significant advantages for drug discovery—and high-quality electronic health records (EHR) data are scarce. First, the coding of diseases in clinical EHR systems typically is designed for insurance reimbursement purposes and is not of sufficient quality for research needs. Second, the many commercially available EHR systems do not easily connect with one another. To illustrate, Partners Healthcare in Boston has spent billions to implement common EHR systems across its hospitals to support its clinical practice and research on personalized medicine. Even so, it also has had to make considerable investments in bioinformatics analysis to ensure adequate diagnostic coding. NIH genetic data, although open to researchers around the world, offer only limited medical/health/phenotypic data. Thus, if START is able to offer tumor samples, genetic data, *and* high-quality clinical data, it would indeed have a valuable resource.

To get there, START faced the challenge that non-START contributors (e.g., local health care providers whose patients' consented samples are sent to START's tumor bank for research) used many different and incompatible EHR systems. To overcome this challenge, START developed a proprietary software tool, called Clinical Synchrony (trademarked in 2013) for retrieving and standardizing clinical data, including both treatment and survival data. Clinical Synchrony extracts relevant data from providers' systems and loads it into vendor Medidata's Rave (a cloud-based clinical data management system) in a common format, so that it can be searched and analyzed.

The next challenge is to provide researchers with a "data portal" that allows them to easily search and analyze linked genomic and clinical data. START's genomic data currently are curated on the WuXiNextCODE platform; START's clinical data are stored in Medidata's Clinical Cloud. Researchers need an easy way to access the two systems in tandem. Both WuXiNextCODE and Medidata have expressed interest in developing START's data portal.

To date, START has no paid staff dedicated to its open source genetic data program. START researchers are participating in the effort as a collateral assignment.

START's scientific contributions are considerable. In 2016 alone, its researchers presented 31 papers and abstracts at a major cancer conference (American Society of Clinical Oncology, 2016). Although START is not affiliated with any university medical school, it sponsors resident visiting scientists.

Table 2 compares the three platforms.

Table 2. Comparison of the three platforms

Company/ Platform	deCODE	WuXiNextCODE	START's SA1kCGP
Key Actors	Founder, Hoffman-LaRoche, venture capitalists	deCODE, Amgen Ventures, WuXi executives	START, donors, WuXiNextCODE, other technology partners
Goals	Scientific research	Infrastructure for precision medicine	Oncology clinical trials
Value proposition	Monetize the data through discovery of new drugs and treatments	Monetize the platform	Create a tissue bank and data resource to support clinical research
Type of data	Genetic, genealogical, and medical data on Icelandic citizens	Genetic, clinical, behavioral and other data brought into the platform by platform customers	Genetic and clinical data on rare cancers contributed by START and regional clinicians
Tensions	Data privacy/security and data access by independent researchers	Scaling and quality of inferences for improved research and clinical care	Open access, data quality, voluntary contributions

4. Discussion

The tales of deCODE, WuXiNextCODE, and START's SA1kCGP show how genetic platforms contribute to "changing models of biomedical research" [7] and clinical care. The platforms facilitate the access to large data sets and the analysis of genetic data combined with detailed phenotypical data from medical records and family history. Increasingly, these platforms would be expected to include behavioral monitoring data from daily activities (e.g., fitness and nutrition data). The entrepreneurial initiatives operate at the edges of traditional health care systems. The platform owners and key architects are neither large university research hospitals nor governments. The initiatives have exhibited considerable flexibility and adaptability to leverage technologies and combine varied data or samples with other datasets. To varying extents, the initiatives also have been able to commercialize their research results in the form of products and therapies in the market. But the initiatives continue to face many concerns and to experience many tensions.

The three genetic platforms have adopted different governance models. deCODE tightly controlled data access and commercialization rights. Its investments were privately funded. In contrast, START's cancer genome project has offered open access to data to encourage voluntary contributions. The investments were funded by financial donations and donations of

time and energy. In the case of the WuXiNextCODE platform, a mixed open–closed data model is followed in which researchers maintain and curate the data they provide but gain access to large volumes of data about human genetic variants and phenotypes provided through the platform. The platform allows researchers to circumvent the nonstandard tools, technical incompatibilities, and data conversions that have previously hampered genetic databanks.

The differences in the governance models and strategies deployed in these platforms were influenced by differing goals and internal and external constraints. deCODE pursued bold scientific discoveries that required improved access to genetic and health data. The firm realized early on that the phenotypic data from medical records and family history were critical to rendering genetic data useful for research. The firm sought to build three linkable databases (i.e., genealogy, health, and genetic) covering the entire population of Iceland. Two of the databases were ultimately built, but the health database was scuttled in the wake of international controversies. However, a change in strategy helped deCODE to reach its goal. deCODE began working directly with physicians to gain access to patients for participation in research studies. As Winickoff [41] reported, "…[deCODE] had found a way to amass large amounts of health information and samples by traditional methods—methods that did not require building the [centralized health system data] architecture for Iceland."

START required high-quality data to carry out its main business of clinical trials for treatments. Its lack of affiliation with medical schools and government agencies created a scarcity of data, which the company resolved by embarking on its own platform initiative. It also relied heavily on existing networks with its cancer clinics and broadened these networks using an open access approach. Unencumbered by the institutional barriers associated with universities and government agencies, START has been able to move fast. Also, its local practitioner networks provided access to samples that represent a spectrum of disease states; thus, START data have not been limited to the most advanced cases of disease that characterize many university repositories. (The most difficult cases often are referred to university hospitals.) START built a technical architecture and open access governance model that encouraged contributions of tumor samples and genetic data. And START's investment in the Clinical Synchrony tool allowed clinical medical record data to be merged into the database. Hence, the different goals and starting points resulted in different governance options. These governance options, in turn, influenced the nature of the regulations and controversies that surrounded the platforms.

The evolving regulations and controversies shaped the overall evolution of the platforms. Many governments around the world restricted (or even prohibited) exclusive commercial access to data gathered by public national health programs; the Icelandic government did so as well, although not initially. Independent researchers feared losing access to research data if deCODE retained exclusive access rights. The medical community mounted opposition to deCODE's plans, including its attempts to enter the direct-to-consumer genetic testing market. The reliability of the tests, the value of the tests, and the citizens' ability to understand the long-term ramifications of such tests were particularly questioned. Concerns about data privacy and the security of individuals' health data culminated in a major change from a centralized data initiative to a much more distributed undertaking.

After Amgen acquired deCODE, the digital platform was separated from the access to data. The platform provided a new pathway to commercialization. The technology was much less contested and regulated, compared to the data. deCODE developed a myriad of new technologies involving major patented inventions for the platform. The large scale of its database and the multidimensionality of the genetic and phenotype data rendered traditional data formats and database structures inadequate. deCODE partnered with vendors to develop its AI capabilities for the platform.

The initiatives also highlight how issues related to data quality shaped the evolution of the platforms. Although deCODE and START used different governance models in their platforms, each exhibited tight controls in data gathering and records management to reduce the data quality problems known to threaten both genomic data and phenotypic data. Such challenges have implications not only for the technical design of genetic databanks, but also for the rules governing data contributions and modifications.

The pursuit of high-quality data in genetic platforms created optimism that entrepreneurial ventures might promote data sharing for research and clinical care. The success of emerging approaches to clinical and translational research depends significantly on improved platforms of genetic and health data, which in turn require greater collaboration and data sharing among researchers and clinicians. Still researchers are reluctant to share the data they have collected and analyzed [15]. For example, researchers might comply with norms and rules for rapid publication of genetic sequence data [20] but then exclude the phenotypic data that would make genetic data much more useful for research. Even when

researchers are motivated to share their data, they are constrained by the need to protect patient confidentiality and to follow data security provisions. These concerns suggest the need for access controls to ensure that only appropriate uses of databank resources are allowed. Meanwhile, even modest access restrictions can seriously impede the "open science" goals of genetic databanks and platforms [6]. Designing governance arrangements that promote contributions, ensure data quality, and encourage innovative and responsible data use is a difficult balancing act, and the need for such a design approach represents promising opportunities for future research.

Although platforms certainly need to facilitate data sharing, research also needs to examine how the platforms shape research collaborations and the research questions pursued in such collaborations. How does the composition of research teams shape the platforms? Both deCODE and START show remarkable levels of international collaboration, as well as significant research productivity.

Another avenue of future research would explore how the platforms collaborate, compete, and trigger the formation of new platforms. Generativity needs to be examined at the platform level. START partnered with WuXiNextCODE for genome sequencing and infrastructure services (e.g., cloud storage services), but START is also pursuing its own open platform, including an access portal. How will these initiatives complete or collaborate with public large-scale national and global data-sharing efforts [17, 40]?

We close by returning to the digital platform literature. Although the platform literature has shed much light on the market and technological perspectives, the data-centric perspective is lacking. The room for further developing this perspective is abundant. For example, future studies need to examine meaning-making systems for the data; such systems not only can circumscribe the data but also control how such platforms facilitate new meanings. That is, algorithms and other tools intended to convert data into actionable knowledge become bounded by these semantic meaning-making knowledge structures [37]. Hence, although the platforms we have discussed here offer promising early steps, much research is needed to understand strategies and governance arrangements for digital platforms to promote scientific discoveries and treatments while maintaining the necessary security and privacy.

5. References

[1] 20 years of Decode, Conference. https://www.decode.com/20-years/

[2] G.J. Annas and S. Elias. "23andme and the Fda," New England Journal of Medicine, 2014, (370:11), pp 985-988.

[3] G. Árnason, G. "Icelandic Biobank a Report for Genbenefit ," Manchester: University of Central Lancashire, 2007.

[4] M.A. Austin, S.E. Harding, C.E. McElroy, "Monitoring Ethical, Legal, and Social Issues in Developing Population Genetic Databases," Genetics in medicine, 2003, (5:6), pp 451-457.

[5] D. Budimir. et al. 2011. "Ethical Aspects of Human Biobanks: A Systematic Review." Croatian Medical Journal, 52: 262-79.

[6] T. Caulfield, S. Burningham, Y. Joly et al. " Review of the Key Issues Associated with the Commercialization of Biobanks," Journal of Law and the Biosciences, 2014, 94-110.

[7] W.F. Crowley, Jr. and J.F. Gusella, "Changing Models of Biomedical Research," Science Translational Medicine, 1(1), October 2009.

[8] E. Davidson and E. Vaast, " Digital entrepreneurship and its sociomaterial enactment." Proceedings of the Annual Hawaii International Conference on System Sciences. 2010.

[9] Eggertsson, T. 2011. "The Evolution of Property Rights: The Strange Case of Iceland's Health Records," International Journal of the Commons (5:1).

[10] B. Elger. Ethical Issues of Human Genetic Databases: A challenge to Classical Health Data Research. Ashgate. 2010.

[11] M. Fortun, Promising Genomics: Iceland and deCODE Genetics in a World of Speculation, University of California Press, Berkeley, CA, 2008.

[12] V.S. Gainer, A. Cagan, V.M. Castro et al. "The Biobank Portal for Partners Personalized Medicine: A Query Tool for Working with Consented Biobank Samples, Genotypes, and Phenotypes Using i2b2," J of Personalized Medicine, 2016, 6(11).

[13] A.Gaver, "Bridging Differing Perspectives on Technological Platforms: Toward an Integrative Framework," Research Policy, 2014, 43(7), 1239-1249.

[14] R. Gertz, R. "An Analysis of the Icelandic Supreme Court Judgement on the Health Sector Database Act," *SCRIPTed* (1), 2004, p 241.

[15] D.M. Gitter, "The Challenges of Achieving Open Source Sharing of Biobank Data," in G. Pascuzzi et al., (eds), Comparative Issues in the Governance of Research Biobanks, Springer-Larg, Berlin Germany, 2013.

[16] J. Grisham, "Genomics Company Formed from Framingham Heart Study," Nat Biotech, 2000, (18:8) 08//print, pp 818-819.

[17] E. C. Hayden, “Genetics Push for Global Data-Sharing,” Nature news, June 05, 2013.

[18] J. Huang, O. Henfridsson, M. J. Liu, and S. Newell. “Growing on steroids: Rapidly scaling the user base of digital ventures through digital innovation.” *MIS Quarterly*, 2017, *41*(1), 1–14.

[19] A. Lawler, "Nih Kills Deal to Upgrade Heart Data," in *Science Magazine*, 2001, pp. 27-28.

[20] P. Lee, “Centralization, Fragmentation, and Replication in the Genomic Data Commons,” University of California Davis Legal Studies Research Paper Series, No. 448. August 2015.

[21] M. L. Markus, “Obstacles on the Road to Corporate Data Responsibility,” in Cassidy R. Sugimoto, Hamid R. Ekbia, and Michael Mattioli (Eds.), *Big Data is Not a Monolith: Policies, Practices, and Problems*, Cambridge, MA: The MIT Press, 2016, pp. 143-161.

[22]J.F. Merz, G.E. McGee, and P. Sankar, "“Iceland Inc.”?: On the Ethics of Commercial Population Genomics," *Social science & medicine, 2004* (58:6), pp 1201-1209.

[23]A.R. Miller and C. Tucker, “Frontiers of Health Policy: Digital Data and Personalized Medicine,” National Bureau of Economic Research, University of Chicago Press, Chicago, USA, 2017.

[24] E. Niiler, "Collapse of Framingham Data Deal Highlights Lack of Cooperative Model," *Nature Biotechnology*, 2001, (19:2), p 103.

[25] F. Pigni, G. Piccoli, and R. Watson, “Digital Data Streams: Creating Value From the Real-Time Flow of Big Data,” California Management Review, 2016, 58(3), pp. 5-25.

[26] T. Ready, "Framingham Data Not for Sale," Nat Med (7:2) 2001, 02//print, pp 137-137.

[27] A. Regalado, A. "23andme Pulls Off Massive Crowdsourced Depression Study," MIT Technology Review:August 1), 2016.

[28] G. Rouleau, G. "Open Science at an Institutional Level: An Interview with Guy Rouleau," Genome Biology, 2017 (18:1), p 14.

[29] J.A. Quelsch and M.L. Rodriquez, 23andMe: Genetic Testing for Consumers (A): Harvard T.H. Chan School of Public Health, 9-514-086, 2014.

[30] H. Rose, “An ethical dilemma. The rise and fall of UmanGenomics - the model biotech company?” Nature. 2003, 425: 123-124.

[31] R. Rosenberg, “Questions still Linger on Heart Study Access: Private Industry’s Right to Use Publicly funded Data for Profit Remains an Issue. Boston Globe, 2001, February 21, D4.

[32] M. Schreieck, M. Wiesche, H. Krcmar, “Design and Governance of Platform Ecosystems – Key Concepts and Issues for Future Research,” Twenty-fourth European Conference on Information Systems (ECIS) Proceedings, Istanbul, Turkey, 2016.

[33] D. Spar and C. Bebenek, “deCODE Genetics: Hunting for Genes to Develop Drugs,” Harvard Business School Case, 9-706-040.

[34] R. Swaminathan, Y. Huang, S. Moosavinasah, R. Buckley, C. W. Bartlett, and S.M Lin, “A Review of Genomics APIs, “Computational and Structural Biotechnology Journal, 2016, 14, 8-15.

[35] L.D.W. Thomas, E. Autio, and D.M. Gann, “ Architectural leverage: Putting platforms in context.” Academy of Management Perspectives, 2014, 28(2), 198–219.

[36] A. Tiwana, B. Konsynski, and A.A. Bush, “ Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics.” Information Systems Research, 2010, *21*(4), 675–687.

[37] I.Tuomi. “Data is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory,” Journal of Management Information Systems, 1999-2000, 16(3), 103-117.

[38] M.W. Van Alstyne, G.G. Parker, and S.P. Choudary, “Pipelines, Platforms, and the New Rules of Strategy,” Harvard Business Review, April 2016, pp. 54-62.

[39] P. Vassilakopoulou, E. Skorve, and M. Aanestad, “A Commons Perspective on Genetic Data Governance: The Case of BCA Data.” Twenty-fourth European Conference on Information Systems (ECIS) Proceedings, Istanbul, Turkey, 2016.

[40] J. Wang, R. Al-Ouran, Y. Hu et al. “Marrvel: Integration of Human and Model Organism Genetic Resources to Facilitate Functional Annotation of the Human Genome.” The American Journal of Human Genetics, 2017.

[41] D.E. Winickoff, “Genome an Nation: Iceland’s Health Sector Database and Its Legacy,” Innovations, 2006 (1:2), pp. 80-105.

[42] S.A. Zahra and S. Nambisan, “Entrepreneurial and Strategic Thinking in Business Ecosystems,” Business Horizons, 2012, 55(3), pp. 219-229.