



## Assessing the Long-Term Viability of Data Platforms for Research

**Niall Brennan, Angela Pupino, Katie Martin, Daniel Kurowski, and Sally Rodriguez**

Health care in the United States is notoriously fragmented. A patient may receive care from a variety of health systems, physicians, and other providers, and that care may or may not be recorded in a patchwork of administrative claims data systems and electronic health record (EHR) systems.

At the same time, health data has become an increasingly valuable commodity. Billions of health data points are generated every day by payers, providers, clearinghouses, and software products, but they are rarely integrated or easy to access. Researchers and policymakers want to use as much health data as possible to generate insights into how our health care system is performing, but they face significant barriers to access. This proliferation of health care data is frequently difficult for researchers to access. Data sources can be expensive, of variable quality, or both. In addition, even if researchers are successful in acquiring certain datasets, combining those datasets with other information is often impossible for reasons of cost, feasibility, or privacy.

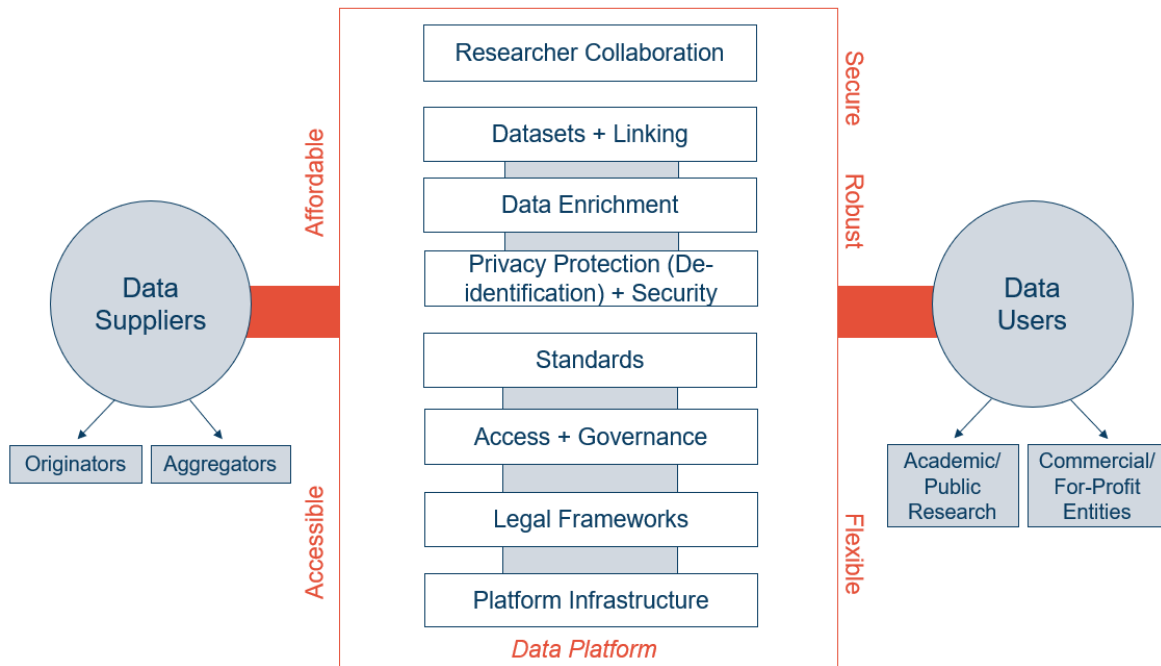
The COVID-19 pandemic exposed these gaps in our national health data infrastructure. We found that public health agencies still generally rely on manual data collection and sharing processes ill-suited to a fast-moving pandemic. The absence of a national all-payer claims database means that researchers hoping to track changes to the health system must settle for a patchwork of system-level EHR data and claims datasets with data from one or a few insurers.

### **What Is a Data Platform, and How Can It Help?**

One potential solution to the challenges researchers face when accessing health care data is the creation of a dedicated data “platform” for hosting, linking, and analyzing datasets to support non-commercial academic research endeavors.

While there is no single definition, for the purposes of this project we define a data platform as the tool that allows a data user to access a database or set of datasets. Data platforms act as a conduit between data suppliers and data users, allowing data to be accessed for research and analysis. Regardless of the platform’s shape or size, it needs to include features like technical infrastructure, data governance and legal frameworks, datasets, and data enrichment materials. The platform interface also ought to enable collaboration between researchers and teams, as well as privacy safeguards and data standards. Ideally, a data platform also would be accessible, secure, robust, and flexible. Figure 1 presents a conceptual model for how a data platform could function.

## Data Platform: Conceptual Framework



**Figure 1: Data Platform Conceptual Model**

Data platforms, sometimes also called data enclaves, are useful because they can host large amounts of data from different sources in a secure and HIPAA-compliant environment, both providing greater access to data and relieving researchers and their organizations from the burden of storing and protecting data. While data platforms are widely used on a siloed basis to host proprietary health-related data, there is an opportunity for a platform containing data from multiple sources that can be accessed quickly and affordably by researchers and policymakers.

The Health Care Cost Institute, in partnership with the Robert Wood Johnson Foundation, conducted this study to research and analyze the feasibility of developing a permanent data platform that would make health care data available for non-commercial use by academic researchers. This project gathered input from both the demand side (i.e., potential platform users) and supply side (i.e., organizations that could potentially provide data to the platform). In addition, we consulted with technical, legal, and data privacy experts to assess the operational and legal challenges associated with such an endeavor. We assembled a Technical Expert Panel to review key findings from our interviews and developed a business plan that outlines the potential funding needed to support such a platform.

This report will outline the challenges that researchers face when using data for health services research, user expectations for platforms, lessons learned from health data platforms, case studies from currently operating non-commercial data platforms, and a potential business model for a non-commercial data platform.

## **Secondary Data Sources Are Critical to Understanding the Function and Impact of the US Health Care System**

Most data used by health services researchers is not generated with research in mind. Instead, the majority of data is generated by organizations such as payers and providers in the course of their daily operations. For example, EHRs generate data through patient-provider or provider-provider interactions, and administrative health care claims data is generated from charges and payments for health care services.

Typically, health care organizations use generated data to help them answer business questions. Although some organizations may have in-house research functions that will create data warehouses or extracts specifically designed for research, operational data warehouses generally are not structured to answer research questions.

### **Getting from Operational Data to Research-Ready Data**

To be useful in supporting broader analytical insights, operational data must be translated into research-ready data. Such translation requires sufficiently documenting the architecture of the relational database when joining schemas and tables from across the enterprise system, establishing the appropriate longitudinal structure to allow research, and converting transaction records into appropriate data models. Moreover, the data must remain relatively static; a researcher pulling data on a cohort using a query on one day should not get dramatically different results on the same cohort using the same query the next day. As such, data requires sufficient run-out or time to allow the data to mature.

Optimizing data from secondary data sources for research also requires an understanding of the data's structure and limitations. Researchers must be able to review and, more importantly, understand the data sources, data elements, and data quality in order to formulate the research question and conduct the analysis. For example, a researcher examining the characteristics of inpatient admissions from a health care claims data source may be interested in utilizing the admission source field on claim form (i.e., the UB-04 form). However, while this field is required by most payers, the results from this field often do not impact payment, and, as with many claims data fields that are not directly linked to payment, can be of varying quality. Similarly, examination of inpatient records for Medicare–Medicaid dually eligible beneficiaries will miss the admission source of the beneficiary admitted to the hospital from a Medicaid nursing facility; the record from the nursing facility rests in a different data system than the inpatient claim, since the payer source for the nursing facility is Medicaid and the payer source for the inpatient stay is Medicare. Good data documentation can help researchers avoid these pitfalls.

A research data resource requires more than just raw data. It also requires transformation of the data so that it is research ready, documentation about data collection and transformation, and subject-matter expertise to help researchers obtain the answers they are looking for.

See **Appendix A** for a use-case highlighting how a research question was answered using two different data resources.

### **Challenges for Non-Commercial Researchers Using Secondary Data Sources**

Once a research-ready dataset is available, researchers and analysts still have several tasks and decision points to complete before beginning a project. This section outlines common challenges that data users face when using secondary data sources for their research.

#### *Data Cleaning*

Because health services researchers are only secondary users of EHR and claims data, data often need substantial cleaning before they can be efficiently utilized for research. Unlike research conducted through primary data collection, where a researcher collects information for the purpose of answering the question at hand, research involving secondary data collection requires an iterative analytic approach to formulate the question and an intimate knowledge of the data so that the question can be aligned with the available data. Data cleaning is a critical first step that represents a time and/or monetary burden for researchers, who can either undertake this process themselves or hire and train data analysts to do it.

#### *Data Storage and Security*

Administrative datasets frequently are large, containing billions of records for millions of patients, which requires relatively sophisticated data-storage capabilities. Additionally, because of the sensitive nature of the data, researchers and their affiliated institutions have to follow strict privacy and security protocols to ensure that the data is accessed only by authorized users. These technical requirements can be cost prohibitive for institutions and researchers who lack funds. Data platforms can alleviate these challenges, democratizing access to data, by providing researchers with secure enclave environments to store, analyze, and link large datasets.

#### *Dataset Linking*

Researchers often seek to link claims and EHR sources with demographic data, geographic data, or provider-level information. When such linkages are possible, connecting disparate datasets can require substantial data cleaning. Provider-level identifiers for physician or billing entities may be inconsistent across datasets, making it particularly difficult to use them to connect separate data sources.

#### *Institutional Barriers*

Culture, practices, and norms at research institutions may create challenges for investigators and analysts looking to pioneer innovative approaches to securing and harnessing data. Researchers may even find it difficult to adopt new data sources at their organizations. For example, government agencies may be slow to adopt commercial data sources, either because of perceptions that private data is worse quality than government sources or because of the cost involved in acquiring private data. Additionally, institutions with less knowledge of data-licensing

processes, such as data use agreements or security and privacy requirements, may be more hesitant to invest in new data sources.

## Demand for Health Care Data

Researchers from academic institutions, nonprofit organizations, and government agencies are all potential users of a data platform that does not allow commercial use. There is significant gray area in the categorization of research teams as either “commercial” or “non-commercial”. For example, platform administrators may choose to allow teams from for-profit companies with government funding to access their data. Conversely, administrators may choose to deny access to academic teams receiving funding from sources such as pharmaceutical companies.

To understand demand-side expectations for health care researcher data platforms, we conducted one-on-one interviews with nine data users representing six universities and one think tank. Additionally, we convened a focus group with seven users of a widely used health data platform.

Although 65% of researchers interviewed expressed high interest in using a data platform, only 20% ever had. The remaining 80% adopted the more traditional approach of physically acquiring and storing data at their institutions. Of those who had used a platform, public platforms such as CMS’s VRDC were the most commonly used.

The researchers interviewed had several expectations of data platforms:

- **Linking Many Types of Data.** The most important feature identified was easy linking between different kinds of datasets. According to one researcher, the ability to link claims and EHR datasets at the patient level alone would be sufficient to encourage them to begin using a data platform. The presence of provider identifiers also was considered an important part of any platform, allowing for records to be linked consistently across datasets at the provider level. Ideally, researchers would be able to follow patients across files (for example, claims data and EHR data or laboratory result data) and different geographic areas, and over time.
- **Flexibility.** Flexibility in research topics and terms of access was another key demand of data users. Because a researcher’s needs and research topics can change quickly, researchers prefer annual and institutional licenses that cover a wide range of research topics over a per-project model. Interviewees expressed difficulty planning for per-project, per-year data access when budgeting years in advance. Researchers also appreciate having flexibility to spend money that they currently have to work on a research project months or years later.
- **Transparency and Quality Documentation.** Researchers also expect that access to a data platform will come with extensive documentation to aid research and data analysis. Understanding a dataset, including populations covered and the meaning and robustness of the data fields, is vital for researchers to shape their analytic questions. Transparency about the details of a dataset helps users know ahead of time what information they can glean in support of their project. Quality documentation facilitates

quicker and easier onboarding of a new dataset. This is especially true for student researchers.

- **Affordability.** As it currently stands, researchers can pay tens of thousands of dollars to access a single dataset for a single project. These costs can be prohibitive for younger researchers and researchers from less-resourced institutions. As a result, interviewees expressed that some institutions prefer to spend less money for worse data over more money for better quality data.

## Supply of Health Care Data

In this section, we explore the supply side of the health data equation. First, we outline how data has emerged as a highly monetizable commodity in recent years and identify the factors that prompt or compel data suppliers to share data for non-commercial reasons. Then, we present the results of interviews with existing commercial and non-commercial data platforms. Finally, we present a series of case studies for some existing non-commercial data platforms.

Health care data has emerged in the last decade as a uniquely valuable commodity. With the emergence of large-scale data analytics, often powered by artificial intelligence or machine learning, there are almost limitless opportunities to harness data to address a wide range of pressing problems facing the health care system today. This has led to an explosion of companies and startups either buying or selling data, frequently for significant amounts of money.

Data is crucial for a host of organizations across the health care ecosystem, ranging from payers wanting to understand their patient or provider populations better to life sciences companies wanting to harness data to evaluate the performance of existing therapeutic regimens and plan for future regimens. Public health agencies need data to better understand the health of populations, and federal and state agencies need data to benchmark program performance and engage in efforts to try and control costs.

At present, many kinds of data suppliers license data to researchers. Direct sources of health care data, such as health insurance companies, may make versions of their datasets available for commercial or non-commercial use. Medical claims clearinghouses involved in aggregating and processing claims also may license datasets. Government sources, such as the Centers for Medicare and Medicaid Services, provide data for studying Medicaid, Medicare, and other public insurance programs. Data from state and regional all-payer claims databases (APCDs) also may be available.

Each source has distinct benefits and drawbacks in terms of available data, cost, consistency of contributors across years, characteristics of payers, and other factors. Direct sources provide better quality data, but restrict access more carefully. Clearinghouses and other data aggregators, on the other hand, usually provide researchers more flexibility when using their data. However, the data from these sources is often of variable quality.

Because data has become so valuable, data holders have little incentive to make it available for anything less than “top dollar.” Traditionally, the federal government has worked to make data available to researchers, but even this data comes at a cost that may be unaffordable to many non-commercial researchers. As a rough rule of thumb, researchers requesting access to Medicare or Medicaid data should plan on spending between \$50,000 to \$100,000 per year to access the data. These figures do not include the cost of recruiting and retaining people who will analyze the data to provide insights. Among nongovernmental organizations that license health care data, relatively few have the sole mission of providing data for non-commercial use. A prominent organization that does license commercial data solely for non-commercial use is the Health Care Cost Institute (HCCI). A case study on HCCI can be found later in this report.

Other organizations that license data for non-commercial use tend to have a hybrid business model under which they license data for both commercial and non-commercial use. Organizations that adopt this hybrid model include Marketscan, Fair Health, and Optum Labs. These organizations can employ price discrimination models, where higher-paying commercial clients can subsidize lower-cost access for non-commercial researchers. These organizations also hold all pricing information very closely, so it is difficult to ascertain what they are charging different users for access to data.

Historically, data suppliers have hesitated to give away their data unless compelled by court decisions or another powerful externality. One relevant example is the COVID-19 pandemic. As COVID accelerated, researchers struggled to access the complete data they needed to understand the pandemic: from high-risk transmission paths to the most vulnerable populations to the impact of disparate policy measures. Fragmentation hindered research and, as a result, obscured our understanding of COVID and knowledge of how to respond.

Given the scale of the global crisis, multiple national and international efforts emerged to try and fill the data and knowledge gap around the spread, progression, and treatment of COVID-19. Two non-commercial data platforms created during the pandemic—the COVID-19 Research Database and the National COVID Cohort Collaborative—are highlighted in the case studies later in this report.

## Existing Data Platforms

As noted above, a data platform is an interface that allows researchers to access data. But a data platform also has features including legal and data governance frameworks, data security and technical infrastructure, and data enrichment materials. Over the past decade or so, several such platforms have emerged. These include government-run platforms such as CMS’s VRDC or state APCD platforms. In some cases, data suppliers and aggregators themselves have established data platforms rather than licensing individual datasets (i.e. Optum or IBM Truven MarketScan).

In collaboration with professionals from Berkeley Research Group (BRG), HCCI conducted interviews with eleven data platforms.<sup>1</sup> The platforms interviewed differed in data focus and funding model. Some were financially supported by government agencies, nonprofit organizations, or academic research centers, while others operated using commercial data-use models or startups.

Data platforms interviewed identified several key challenges and considerations that emerged through running their platforms:

- **Data Contributor Buy-In.** Data platforms often have to purchase their component datasets at a substantial cost. Others may broker free or low-cost data sharing agreements. Whether they pay for access to contributor datasets, data platforms are understandably concerned about keeping data contributors happy. Data contributors may pull out of the platform if their concerns are not addressed or conditions are violated. One platform admitted to “overcorrecting” in trying to anticipate data contributor concerns, providing access to only partial datasets rather than releasing datasets that researchers would find most useful.
- **Distinguishing Commercial and Non-Commercial Researchers.** The line between “commercial” and “non-commercial” research teams can be surprisingly blurry. Academic researchers may receive funding from commercial interests like pharmaceutical companies, or commercial researchers may team up with academic researchers to receive academic pricing or increased legitimacy. Additionally, researchers working in the commercial sector may have academic affiliations from teaching at universities. This may cause complications for platform governance and compliance, but also may have economic impacts for the platform. For data suppliers using a two-tiered pricing model, academic pricing may eat into the higher-margin commercial pricing model.
- **Researcher Support.** As outlined in the researcher challenges section, a relatively small subset of non-commercial researchers are well-versed in processing, cleaning, and understanding secondary data. Early on in their projects, researchers accessing health data platforms often need a lot of support in learning about the datasets and ramping up their research. Resources like Wiki pages, data dictionaries, and code repositories can be helpful, but substantial resources are still needed to onboard and orient researchers.
- **Data Cleaning and Quality Assurance.** While data cleaning can be a resource-intensive undertaking for data users, the unrefined nature of EHR and claims datasets also poses challenges for those maintaining data platforms. To provide the best experience to users, platforms must train and maintain staff to extract, transform, and load sources into the platform and create resources for users. Platforms also must create processes for open communication to allow users to report errors or variables that do not look right.

---

<sup>1</sup> *The views and opinions expressed in this publication are those of the authors and do not necessarily reflect the opinions, position, or policy of Berkeley Research Group, LLC or its other employees and affiliates.*



- **Governance.** Data platforms function best when they have strong governance frameworks at the outset. Some of the platforms interviewed reported that it was harder than they expected to define who should get access to what parts of the data and for what purposes. The balance between rigorous governance policies and low-friction data access for users can be difficult to achieve.

Interviewees also offered recommendations for future data platforms:

- **Tiered Funding Model.** Interviewees mentioned that anyone creating a platform should consider tiered access models that permit some researchers to access the data for a lower access fee than others. Such a model would allow more researchers—especially students and those from universities or departments with smaller budgets or less resources—to access the data. Payment amounts could be based on intensity of data use, number of data users on the team, or some other factor. Many data sources already use tiered funding models in permitting commercial and non-commercial researchers to access the data at different price points.
- **Consider Innovative Incentives for Free Data Contribution.** While many commercial entities are hesitant to give away their data, platforms may entice them to do so by offering services that provide value back to contributors. Possible incentives include:
  - *Offering to clean or validate datasets on behalf of the data contributor.* Since data platform administrators presumably will be cleaning these datasets anyway, this would not be a significant additional lift on the part of the platform.
  - *Citing data contributors in published papers that utilize the platform.* Citations in peer-reviewed journals can increase data contributors' reputations and attract new paying users to their data.
  - *Training students to use secondary data sources.* Some data contributors also noted that they derive value from the opportunity to have students become proficient in using their data. Today's students will be working in health services research in the future, increasing the pool of researchers proficient in using secondary data sources and reducing the onboarding burden for administrators.

### **Non-Commercial Data Platform Case Study: HCCI**

The Health Care Cost Institute was founded in 2011 as a unique partnership among four national health insurance plans as an independent, nonprofit organization to analyze patterns and trends in health care spending in the United States. HCCI houses a claims database that includes more than 55 million covered lives from the employer-sponsored insurance market (i.e., from people who get health insurance through their employer or the employer of a family member). Claims data is the most comprehensive source of real-world evidence available to researchers; it is the gold standard for timely, population-level information about the health care system. HCCI's claims database has information on millions of doctors' visits, health care procedures, prescriptions, and payments by insurers and patients

After aggregating claims data from multiple payers, HCCI statistically deidentifies it to protect patient, payer, and provider identities. HCCI also licenses the dataset to researchers—who

access data remotely via a secure data enclave—to enable even more insights into the drivers of US health care spending. Historically, more than 140 external researchers have licensed HCCI data, with approximately 40 teams conducting work in any given year. HCCI’s dataset is available exclusively to researchers in academia, at government agencies, or in nonprofit organizations. The HCCI dataset cannot be used for commercial purposes and cannot be licensed to anyone other than researchers at organizations described above.

Generally, the dataset is available to academic researchers for a fee of \$45,000 per year. That fee allows two people access to the secure enclave to work with the data. Special pricing is available for students. A \$15,000 fee allows one person to access the data for one year. HCCI charges additional fees for services such as merging external data sources (e.g., American Hospital Association data) with the core HCCI dataset or additional data users.

In recent years, several disruptions have threatened HCCI’s financial sustainability and business model. In 2017, one payer who contributed data and funding to HCCI withdrew both, deciding that it no longer wanted to partner with HCCI. In its place, HCCI contracted with a data aggregator affiliated with a large, national network of health plans. That organization charges HCCI substantial fees for access to their aggregated data. Although the data solved one issue with the founding payer’s departure, it created a new cost center that data access fees alone cannot cover.

HCCI’s story highlights several challenges with health data platforms. First, HCCI is dependent on health plans voluntarily sharing granular health care claims data. As discussed elsewhere in this report, plans may be motivated to do so willingly. At other times, however, they are not. Without willing data suppliers, a data platform like HCCI’s cannot facilitate the connection between non-commercial researchers and robust data assets. Second, the costs of aggregating and hosting health care data is substantial even when the organization does not have to pay for the data itself. The additional costs and the loss of private-industry financial support has made HCCI more dependent on philanthropic funding than it had been previously.

### **Non-Commercial Data Platform Case Study: National COVID Cohort Collaborative (NC3)**

One data platform to emerge during the pandemic was the National COVID Cohort Collaborative (NC3), created by the NIH’s National Center for Advancing Translational Sciences. NC3 was created as a repository and secure data enclave for electronic health records from across the country. Thanks to the financial support of the NIH, this data is provided to researchers free of cost. NC3’s three EHR-based datasets are available to researchers from US-based institutions. Researchers from foreign institutions are permitted to access only the deidentified dataset, which algorithmically shifts dates of service and truncates ZIP codes; and the synthetic dataset, which is created to resemble patient information but does not use real patient data. Notably, the synthetic dataset is also accessible to “citizen scientists.”

The platform maintains a dashboard to provide information about the cohort, the prevalence of comorbidities in the data, and publications that have used the platform to potential researchers.<sup>2</sup> As of October 25, 2021, NC3 has researchers working on 272 projects studying records from 67 facilities and 2.8 million COVID patients.<sup>3</sup>

### **Non-Commercial Data Platform Case Study: The COVID-19 Research Database**

The COVID-19 Research Database (RDB) was established in April 2020 to help solve this problem for COVID research. The mission of the RDB is to accelerate the amount of real-world data that researchers can freely access to better understand and mitigate the effects of the COVID-19 pandemic.

To achieve this, the COVID-19 Research Database convened a cross-industry consortium of leading health care organizations and scientists, collaborating across data contribution, governance, privacy review, and analytics to assemble a secure database of HIPAA-compliant, deidentified, and limited patient-level datasets, which includes a wide variety of structured and linked health records, such as claims data, electronic health record data, and consumer data.

The RDB makes all datasets and resources available to public health and policy researchers at no cost to further non-commercial and nonlegal investigations into the direct and indirect impact of COVID-19. Since its inception, the RDB has seen more than 3,300 registrations and supports over 600 active research users. More than 350 study proposals have been processed by the steering committee, and the RDB has led to more than 35 peer-reviewed publications. Research supported by the RDB has been the subject of numerous conferences and presentations, has informed policymaking discussions, and has garnered national attention, informing and improving public understanding of COVID-19.

As laudable as this collaborative effort has been, as the pandemic phase of COVID-19 draws to a close, the reality is that this pro bono data infrastructure was only made available to support research into COVID-19. Already, several data contributors have either withdrawn their data or signaled an intent to do so soon. The collaborative that established the RDB is exploring whether a permanent pro bono research platform can be established on a topic-specific basis (e.g., establishing or maintaining the data resources combined under the COVID-19 database to study health equity or other conditions such as Alzheimer's).

### **Illustrative Data Platform Business Model**

As our qualitative research reveals, creating a data platform requires substantial upfront investment. Initial costs include data storage, processes to implement data security and privacy protections, data-ingestion systems that include standardizing data and data quality checks,

---

<sup>2</sup> <https://covid.cd2h.org/dashboard/>

<sup>3</sup> <https://ncats.nih.gov/n3c>

and, for data that is built for researchers, initial refinements that make the data more user friendly. After the base technical framework and systems are in place, many of those costs recur annually as new data is incorporated. Many sources of health data have millions or even billions of records, and the infrastructure and processes described above need to be scaled accordingly. According to anecdotal evidence and some case studies, building and maintaining a data platform's infrastructure can cost millions of dollars. If a data platform has to pay for data to include in its platform, the costs of becoming operational are even higher.

At the same time, the ability of users to pay for data access varies widely. For academic, government, and some nonprofit policy researchers, data access fees for a given project can take up a substantial portion of their budgets. Data access fees of \$45,000, for example, would be nearly a quarter of a \$200,000 grant project. The constraints mean that non-commercial researchers tend to be more sensitive to price than researchers with fewer limitations; they also may be more willing to pursue lower-price or no-cost alternatives, even if those alternatives compromise the completeness or quality of the data they are using. In contrast, health care companies, such as pharmaceutical firms or device manufacturers, likely have more funding they can dedicate to accessing data and, generally, will pay higher prices for the same data resource.

It can be challenging for data platforms to secure the revenue from a mix of data-access fees from commercial users, data-access fees from non-commercial users, government funding, and philanthropic funding that offset the costs of creating and hosting a usable dataset. A data platform dedicated to non-commercial research (in which licensing to commercial users is not permitted) would require higher fees from non-commercial users, greater government funding, or more philanthropic support to offset the loss of commercial data-access fees.

To illustrate how a health data platform's financial sustainability is affected by costs and demand for data, we commissioned a business model for this project. That model includes the following input costs to building and operating a health data platform:

- Initial build
- Annual maintenance
- Data storage and management
- Data cleaning and linking
- Usage fees
- Data acquisition (where applicable)

A platform also likely would incur legal fees, advertising, business development costs, and staff costs to oversee vendors and interact with users, but we considered those routine operating expenses. It seems unlikely that any of those would be larger than the costs identified above.

For purposes of understanding the cost and revenue dynamics, assume the data platform organization uses a model that costs \$8 million in the first year and \$3 million annually

thereafter (as more of an operations and maintenance phase). The business model then allows us to assess as many user and revenue scenarios as possible. For example:

- If the data platform had 50 data users, it would need to charge almost \$160,000 in data access fees to break even. Assuming the same number of users in future years, the platform would need to charge nearly \$57,000 to each user. If the platform were restricted to non-commercial research, all of those users would need to come from academia, government, or other nonprofit organizations.
- Assuming the platform receives \$1 million from philanthropies in the first year and \$500,000 in subsequent years and has 50 data users, the platform would need to charge nearly \$140,000 in data-access fees per user to break even in the first year. With the same number of users in future years, the platform would need to charge approximately \$47,000 in data-access fees per user.
- If commercial research were permitted, the data platform could charge differential data access fees. With commercial data access fees of \$100,000 per project per year and assuming 50 commercial and 50 non-commercial data users, the platform would need to charge non-commercial users almost \$65,000 in data-access fees per user to break even. There may be circumstances for which there is enough revenue from commercial data users that non-commercial access could be provided at no charge. In our model, it would take 84 commercial users paying \$100,000 each or the same 50 users paying \$163,000 to offer non-commercial users data access for no fee in the first year of operations.

Another way of considering the scenarios above is to assume that non-commercial data access fees were set at \$50,000. To break even, the data platform would need 150 non-commercial data users in the first year and nearly 50 data users to break even in future years. With philanthropic grants or government funding, as well as data-access fees, the platform could be financially sustainable with fewer users.

In any case, the illustrative business model shows that health data platforms need either a substantial commercial business or considerable philanthropic or governmental funding to offset the costs of developing and maintaining a robust data resource for non-commercial researchers to use.

## **Conclusion**

As health data becomes increasingly valuable to data suppliers and researchers, demand for platforms to host, link, and analyze data will continue to rise. However, evidence from currently operating platforms shows that few, if any, research data platforms have established a stable long-term revenue model without permitting commercial users. In the absence of higher charges for commercial users, significant, long-term foundation support or government intervention likely will be necessary to support a data platform.

Powerful externalities, such as the COVID-19 pandemic, have allowed for the creation of data platforms more quickly and cheaply than previously thought possible. However, these externalities are unpredictable and impossible to control. Data philanthropy has taken hold in moments of crisis but does not seem to be taking hold long term. Government action or a true shift toward data philanthropy is the best path to a long-term sustainable research data platform.

Appendices:

## **Appendix A: Quality v. Affordability of Data: Preventive Services Case Study**

Researchers at HCCI used two data sources to examine how the COVID-19 pandemic affected the utilization of preventative health services like childhood vaccinations and cancer screenings. The first data source came from the COVID-19 Research Database, a data repository of claims, EHRs, and other health-adjacent data sources that was made available free to researchers. The second dataset was HCCI's commercial claims dataset, which is available to researchers at cost on a per-project per-year basis.

Data from the COVID-19 Research Database were made available to the researcher both quickly and free of charge—two key benefits of using the platform. HCCI analyzed data from a claims clearinghouse company to measure utilization of certain services over time. However, the structure and quality of the data limited the questions the researchers could answer. Deficiencies in the data included the lack of health insurance enrollment information, poor demographic data, and little available information about the organizations that submitted claims through the clearinghouse. From a researcher perspective, the task of answering a question on service utilization using this data was itself an iterative and difficult process. The project required significant amounts of time to run exploratory data analyses to determine deficiencies in the data and ultimately necessitated a change in the research question. Finally, any intrinsic enrichment of the data, such as mapping geographies, provider characteristics, or service categories, were not included in the data.

While researchers benchmarked results to services that are not expected to change significantly over time (i.e., childbirth), not only was the generalizability of results limited, but the measurement and validity of smaller changes in utilization would have been suspect. Researchers were certain of the results only because they witnessed large declines in service utilization.

The second data source was the HCCI commercial claims dataset. This dataset requires a significant amount of data run out to allow the data to mature; in addition, it is appropriately documented, reviewed for quality, and enriched with other data sources. All of this upfront work allowed researchers to better answer the research question and allowed more flexibility in designing the research model. Additionally, knowing the size and makeup of HCCI's cohort allows researchers to weigh their convenience sample and generalize to the entire employer-sponsored insurance market. However, non-commercial researchers must pay to access HCCI's dataset and must wait weeks or months to access the data as their institutions fill out legal paperwork, including a data use agreement.

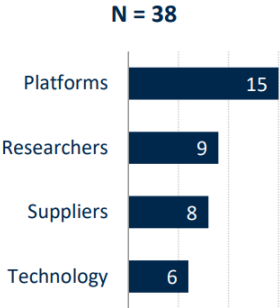
## **Appendix B: Interview Participants and Summary**

### **Organizations Represented in Interviews**

- Agency for Healthcare Research and Quality (AHRQ)
- Aridhia
- Brookings Institution
- CareJourney
- Change Healthcare
- Center for Open Data Enterprise (CODE)
- COVID-19 Research Database
- Data Across Sectors for Health
- Datavant
- Healthjump
- Health Care Cost Institute
- Johns Hopkins University
- Medidata
- Mt. Sinai Medical Center
- National Institute on Aging
- Nightingale
- National COVID Cohort Collaborative (N3C)
- OnPoint
- Open Commons Consortium
- Palantir
- RAND Corporation
- Robert Wood Johnson Foundation/Health Data for Action
- SAS
- Snowflake
- TriNetX
- Truven/IBM MarketScan
- University of California, San Francisco
- University of Maryland School of Public Health
- University of Pittsburgh
- University of Southern California



## Interview Summary

Sources	Methods	Key Issues Explored	Scope of Role										
<ul style="list-style-type: none"> <li>Potential data platform users, including healthcare researchers at academic institutions, corporate organizations and policy think tanks</li> <li>Data suppliers from government and commercial entities</li> <li>Companies and government offices with a data platform</li> </ul>	<ul style="list-style-type: none"> <li>Leveraged relationships from both HCCI and BRG team members, across healthcare, data, tech</li> <li>Conducted in-depth phone interviews focused on key question ranging from 30 to 90 minutes</li> <li>Send a short survey following the interview</li> </ul>	<ul style="list-style-type: none"> <li>Demand for data platform that houses commercial claims data along with other healthcare data</li> <li>Biggest data needs</li> <li>Operational or technical challenges</li> <li>Long-term viability of platform aimed at non-commercial users</li> <li>Likelihood of data philanthropy as means for funding</li> </ul>	<p><b>N = 38</b></p>  <table border="1"> <thead> <tr> <th>Role</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>Platforms</td> <td>15</td> </tr> <tr> <td>Researchers</td> <td>9</td> </tr> <tr> <td>Suppliers</td> <td>8</td> </tr> <tr> <td>Technology</td> <td>6</td> </tr> </tbody> </table> <ul style="list-style-type: none"> <li>Healthcare Researcher</li> <li>Policy Researcher</li> <li>Professor of Pharmacy</li> <li>Research Data Administrator</li> <li>Chief Executive Officer</li> <li>Chief Tech Officer</li> <li>Senior Vice President</li> </ul>	Role	Count	Platforms	15	Researchers	9	Suppliers	8	Technology	6
Role	Count												
Platforms	15												
Researchers	9												
Suppliers	8												
Technology	6												

## Appendix C: TEP Participants

### Technical Expert Panel (TEP) Members

- Shahir Kassam-Adams, Datavant
- Joseph Levy, Johns Hopkins University
- Deven McGraw, Ciitizen
- Stephanie Reisinger, Veradigm
- Kosali Simon, Indiana University
- Erin Trish, University of Southern California Schaeffer Center