

---

# Big Data Analytics and the Cancer Moonshot

*Recommendations for  
the White House Cancer  
Moonshot Task Force*

## AUTHORS

Kayla Benker

Todd Harris

Katie Malone

Angelo Mancini

Ola Topczewska

Dan Wagner

---

# Summary

*The White House Cancer Moonshot Task Force asked Civis Analytics to provide an objective perspective on the problems and opportunities in the cancer research space—specifically regarding data, analytics, and technology. This document summarizes the challenges and provides a set of policy recommendations for the Task Force.*

We talked to representatives from over 40 institutions, including pharmaceutical companies, research groups, medical professionals, legal teams, commercial and nonprofit groups working on data systems, and cancer survivors.<sup>1</sup> Our goal was to help answer a few big questions:

- What is the current and potential role of data and analytics in cancer research?
- Where is the system under-delivering, and what are the barriers to getting it right?
- What recommendations can we provide to the administration to accelerate the development of effective therapies?

This year, the cancer community will spend over \$100 billion on research, treatment, and other associated medical costs. Institutions that fund and carry out research—including the National Cancer Institute (NCI), pharmaceutical companies, and independent research labs—will employ tens of thousands of doctors, researchers, and technicians in the search for new treatments. There is a consensus within the community, however, that the system is under-delivering on its potential. Its challenges range from complex technical problems, like how to efficiently store and analyze vast amounts of genetic sequencing data, to bureaucratic barriers that slow innovation.

Despite these challenges, the industry has never been more hopeful about its future. Our conversations suggest that there is enormous promise in bringing together individual-level genomic and clinical data, and using data science techniques to uncover patterns. This work is not meant to supplant existing basic and clinical research, but rather to complement it and accelerate the development of new treatments.

While it's tempting to endorse a specific solution, like the newest subfield of research or a given institution, we recommend that the Moonshot effort focus on systemic reform.

Advanced, data-driven cancer research has three foundational requirements: technology, data, and people. Researchers need the right data infrastructure to store and analyze large cancer research datasets, the right data sharing permissions and standards to build those datasets, and trained experts to build the infrastructure and carry out the analysis. Right now, however, there are major systemic barriers that prevent the U.S. cancer research system from meeting these foundational requirements. To fully deliver on the potential and promise of advanced data analytics in cancer research, we recommend reforms in each of these three areas: data infrastructure, data sharing, and people and skills.

**“[Researchers reach a point where the] bureaucracy becomes so dense that they start walking away from interesting things, because...there is a [significant bureaucratic] cost.”**

COMPUTATIONAL BIOLOGIST  
AT A MAJOR UNIVERSITY



## DATA INFRASTRUCTURE

Applying cutting-edge data science techniques to cancer research requires datasets of immense complexity and size, which require server space and computing capacity far beyond what most academic research labs can handle. For example, The Cancer Genome Atlas (TCGA) is a large dataset of genetic sequences (currently 2.5 petabytes) stored on a handful of research servers, and it requires vast resources and specialized infrastructure to download and use. If an institution wanted to create their own copy of the dataset, it would cost at least \$20,000 per month just to store this data in a cloud-based system and even more to store it on local servers.<sup>2</sup>

While pharmaceutical companies and major research institutions are able to keep up with the demands of data this size, most researchers can't. In efforts to manage this data, universities and academic consortia set up their own data infrastructures, resulting in duplication of effort and great expense. With a central infrastructure in which to combine this data, researchers could investigate the relationship between a patient's genetic code, the genetics of their cancer, and the best way to treat their individual disease at a scale that is not currently possible.

A large independent research organization told us that storing the necessary data required such intensive resources that they had to invest in new cooling systems to prevent the building from overheating, rather than spending money on new research.

In addition to the technological demands of data storage, there is also a need for adequate analytics infrastructure to allow researchers in the field to perform computations on the data. Currently, this capacity is under-supplied.

Finally, a major barrier to breakthrough data analytics in the cancer space is that critical data sources are not structured in a consistent way that enables easy analysis. This is a problem with genomic data, but the issue is more pronounced with patient health data, which differs significantly from hospital to hospital and requires laborious data processing. Currently, in order to do the kinds of analyses discussed in this document, researchers would need to act as data engineers, combining and cleaning different types of EMR data. This is a very high barrier to cross and—in our view—not an effective use of researchers' time.

“Clinical data is like crude oil—great but not going to power a car unless you refine it.”

AMY ABERNETHY  
CHIEF MEDICAL OFFICER/  
CHIEF SCIENTIFIC OFFICER,  
FLATIRON HEALTH

***We recommend a government-funded network capable of housing these large datasets in a way that makes them accessible to many more researchers. We envision developing this repository in stages. This central infrastructure should build upon the NCI Genomic Data Commons (GDC).<sup>3</sup> Additionally, the Department of Health and Human Services (DHHS) should establish a Center of Excellence (COE) to help researchers with data formatting and processing. As a complement to these two efforts, the NCI should invest in a cloud-based analytics platform, which will give researchers the computational capacity to perform this work.<sup>4</sup>***

## DATA SHARING

In order to use data science to identify relationships between a patient's genetics, the genetics of their cancer, and clinical outcomes, the cancer field will need to unify vast amounts of patient-level clinical and genomic data. Currently, patient-level data sharing is impeded by three underlying problems.

First, there isn't a consistent standard for EMR data, leading different hospitals to collect different types of information.

Second, the exchange of patient health information is tightly controlled by privacy regulation, chiefly the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and state laws. These laws prevent sharing of identifiable health information by healthcare providers for research purposes without individual patient consent. The principles of privacy, control over one's own health information, and informed consent for research are incredibly important and should not be undermined to further the development of new treatments. However, we see high variability in the way these principles are interpreted by research institutions and considerable ambiguity about their application. At times, the resulting data sharing policies slow down research without necessarily improving patient security or privacy. There is room to create clarity and standardization around these regulations.

Under HIPAA, patients are allowed to access and share their own medical records. We see potential for a grassroots solution to the problems of data sharing. Under this model—one that has been prototyped by initiatives like Sync for Science and the Blue Button Initiative—patients would be able to export their health information and donate it to science. Currently, there is not enough awareness of existing efforts or supporting infrastructure to facilitate the data hand-off. We envision a future in which the government partners with outside organizations to support this process.

Third, academic researchers face a different set of data sharing problems. Right now, researchers generate many types of data using different methods. While some data sharing requirements exist, including the NIH Genomic Data Sharing Policy, there isn't currently a centralized place to store research data, and data sharing is onerous for researchers.

**It took 18 months for member institutions to agree on conventions for measuring improvement in patient outcomes that could be generalized to the entire network.**

FORMER LEADER OF A MAJOR  
CANCER DATA SHARING NETWORK

**The principles of privacy, control over one's own health information, and informed consent for research are incredibly important.**

**"These rules [HIPAA] were not designed to promote big data analysis—they were developed in the 90s. Data analysis was happening, but not in the way that we're talking about today."**

TEAM OF LEGAL EXPERTS SPECIALIZING  
IN PRIVACY REGULATION



*We recommend that the government partner with industry leaders to establish interoperability standards for EMRs and incentivize vendors to comply with them. To address misinterpretation and uncertainty around privacy regulations, the government should provide guidance to researchers on forms of data sharing that are compliant with existing regulation, and—in the long run—begin a conversation about reforming the regulations themselves. In partnership with research organizations and the cancer advocacy community, the government should increase publicity and funding for patient-driven data donation programs. Finally, the government should expand requirements for sharing federally funded research data into a centralized repository.*

A researcher at a major university was prohibited from using email to move any research data at all because of privacy concerns, so his team needed to exchange bulky external hard drives to share information.

## PEOPLE AND SKILLS

New data science technologies and techniques have the potential to revolutionize cancer research, but the field has not kept pace with the changing data landscape. To address this, the government needs to help close skill gaps in two areas of the cancer space. First, in order to build a centralized data infrastructure and the software to access it, the government needs to employ data engineers, systems engineers, and designers to create a system that is scalable, extensible, and accessible to people at different skill levels.

Second, in order to analyze data at the rate it is generated, there need to be more skilled bioinformaticians and data scientists in the cancer research field. Currently, there are too few viable career paths for full-time data scientists within academia, and the positions that do exist compete with the private sector for skilled workers.

**“There is a skill gap at every level that begins with the inability to download and process a dataset.”**

RESEARCHER AT A PROMINENT  
CANCER CENTER

In addition to dedicated personnel who have experience with data science, the government should also support training programs for researchers who are interested in exploring the new types of work they could do with large scale data.

**“I do not think that we need to compete with Silicon Valley but instead work with them. Tell kids that if they want to change the world, they can write apps for research to improve patient outcomes and patient care.”**

CANCER RESEARCHER AT EMORY  
UNIVERSITY

Finally, the government should support partnerships with the tech industry to facilitate the exchange of information and skills between the cancer research field and the private sector.

***The government should invest in the data science capabilities of the cancer research community, fund long-term career paths for skilled people, and facilitate public-private skill exchanges with the tech industry.***

# Summary of Recommendations



## DATA INFRASTRUCTURE

Page 10

1. Continue to develop a centralized repository to hold diverse cancer data, building on the NCI Genomic Data Commons (GDC).
2. Launch a DHHS Center of Excellence (COE) on cancer data that will provide guidance on data integration and standardization.
3. Build a cloud-based cancer analytics platform that researchers can use to do large scale analyses.



## DATA SHARING

Page 16

1. Develop and enforce Electronic Medical Record (EMR) formatting standards.
2. Provide guidance on HIPAA and other privacy regulations to research institutions.
3. Encourage a patient-driven data donation model by supporting infrastructure to move patient data from an EMR to a central data repository and publicizing existing efforts.
4. Expand sharing requirements for academic data, and incentivize sharing into a centralized location for easy access.
5. Work with privacy experts, legal experts, patient advocates, researchers, and others to discuss reforms to HIPAA and other privacy regulations.



## PEOPLE AND SKILLS

Page 20

1. Invest in data science training programs for academic researchers.
2. Staff the centralized repository, cloud platform, and COE with the engineers and designers needed for success.
3. Support public-private partnerships with tech industry.



---

# Why it Matters

*The recommendations outlined in this document require significant effort and funding. At a time of limited federal resources, every dollar spent on our recommendations is a dollar not spent on other types of work. But it's worth it. There is ground-breaking work being done, but by and large, the field lags behind the private sector in joining the big data revolution. The recommendations in this document outline the conditions for success in the cancer space; they allow researchers to collaborate and exchange information more effectively, giving them space to focus on what they do best—developing new therapies.*

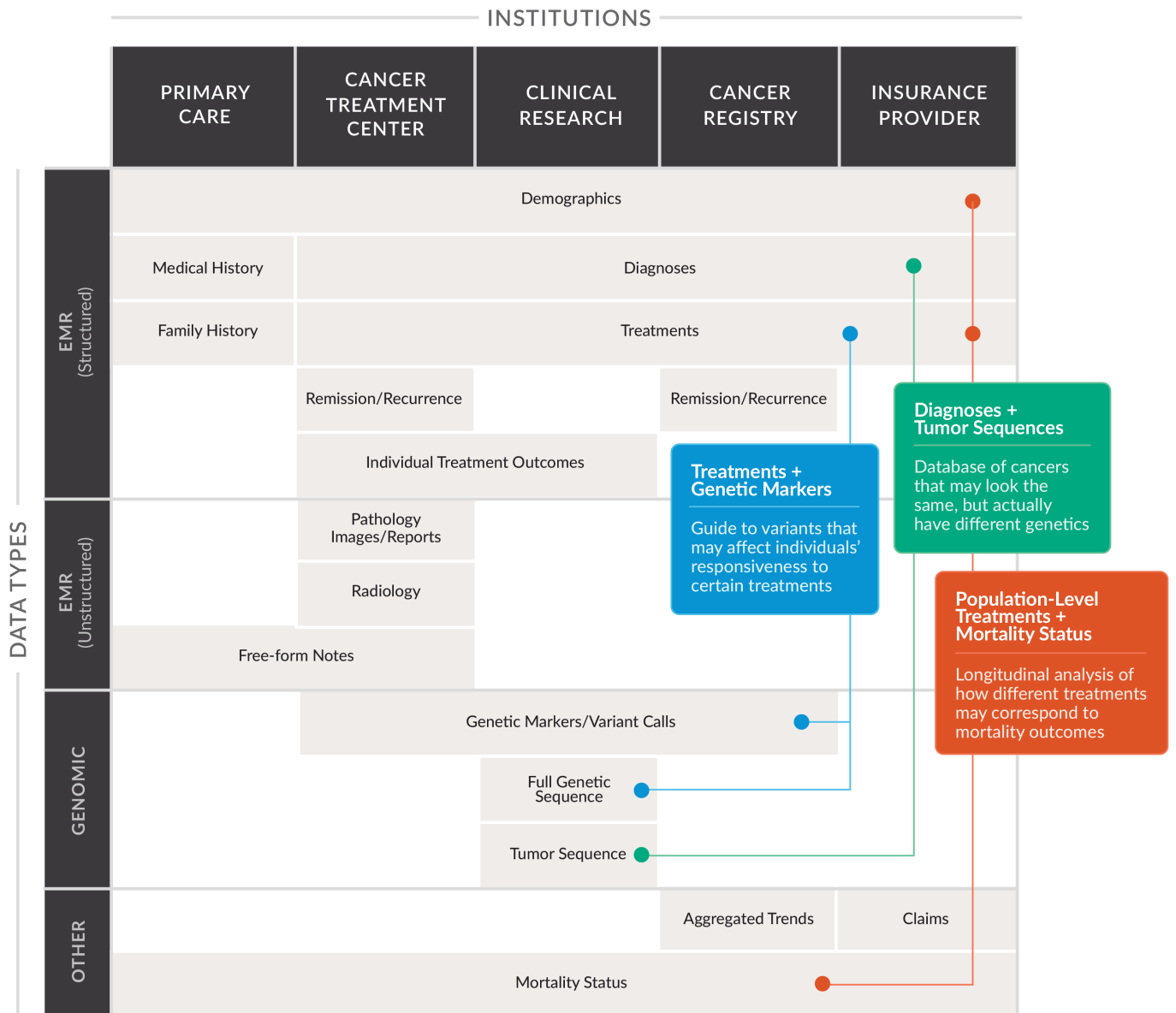
In the same way that investing in federal highways enables commerce nationwide, investing in infrastructure for cancer data facilitates treatment development. A dataset with millions of treatment outcomes joined with tens of thousands of genomic and tumor sequences would be a game-changer. It would allow researchers to study relationships between specific genetic variations and responsiveness to different treatments, moving us closer to truly individualized medicine. It would also enable new kinds of research, leading to better clinical treatments. But solving the data problem alone is not enough; the non-data components are just as important.

**In the same way that investing in federal highways enables commerce nationwide, investing in infrastructure for cancer data facilitates treatment development.**

Clarification of privacy regulation would reduce ambiguity and bureaucratic costs in the research process while continuing to safeguard patient rights.

Data standardization would cut down on laborious data processing. With a more nimble and centralized data infrastructure, better ability to share data, and the people and skills to do the work, the capacity of the cancer field will grow alongside the data and ultimately produce better therapies.





The cancer data landscape consists of a variety of institutions using and generating many different types of data. Much of this data is currently siloed. Combining these data sources would allow researchers to answer existing questions in new ways and open the door to entirely new methods of inquiry.



# Build Infrastructure to Store and Analyze Shared Data

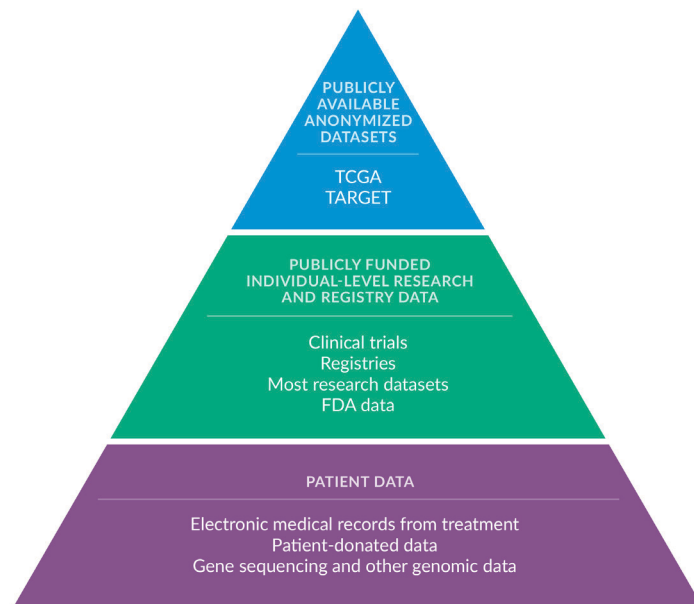
*The government should invest in a centralized data repository which integrates diverse data sources. This data repository should be modeled on the GDC, and should integrate data in stages, starting with (1) large publicly funded genomic database projects (e.g., TCGA, TARGET), then (2) federally funded cancer research studies and cancer registries, and finally, (3) patient-donated data. In order to resolve the problem of data formatting inconsistencies, the government should establish a Center of Excellence to help researchers standardize and work with data. Additionally, the government should invest in a cloud-based analytics platform modeled on the NCI Genomic Cloud pilot program.*

There's power in unifying genomic and clinical outcome data.

## Why it's important

There's power in unifying genomic<sup>5</sup> and clinical outcome data.<sup>6</sup> A central infrastructure enables new types of analysis. By tying genomic data to clinical data and harnessing data science techniques, researchers can begin to answer questions like the following:

- Are there genetic variations that are common across different types of cancers?
- Can the effects of these variations be targeted with tailored treatments?
- Based on someone's genomic profile and the genome of their cancer, what combination of treatments will be most effective, and what side effects or complications are most likely?
- Are there genetic patterns in tumors that make them more or less resistant to treatment, and can those patterns be identified to develop new therapies?



**Stages of Data Integration into Centralized Data Infrastructure**  
 We envision the process of loading cancer data into the GDC or successor centralized repository proceeding in these three stages. In Stage 1, existing large-scale public data sources, including TCGA and TARGET would be moved into the system.



## Where we stand

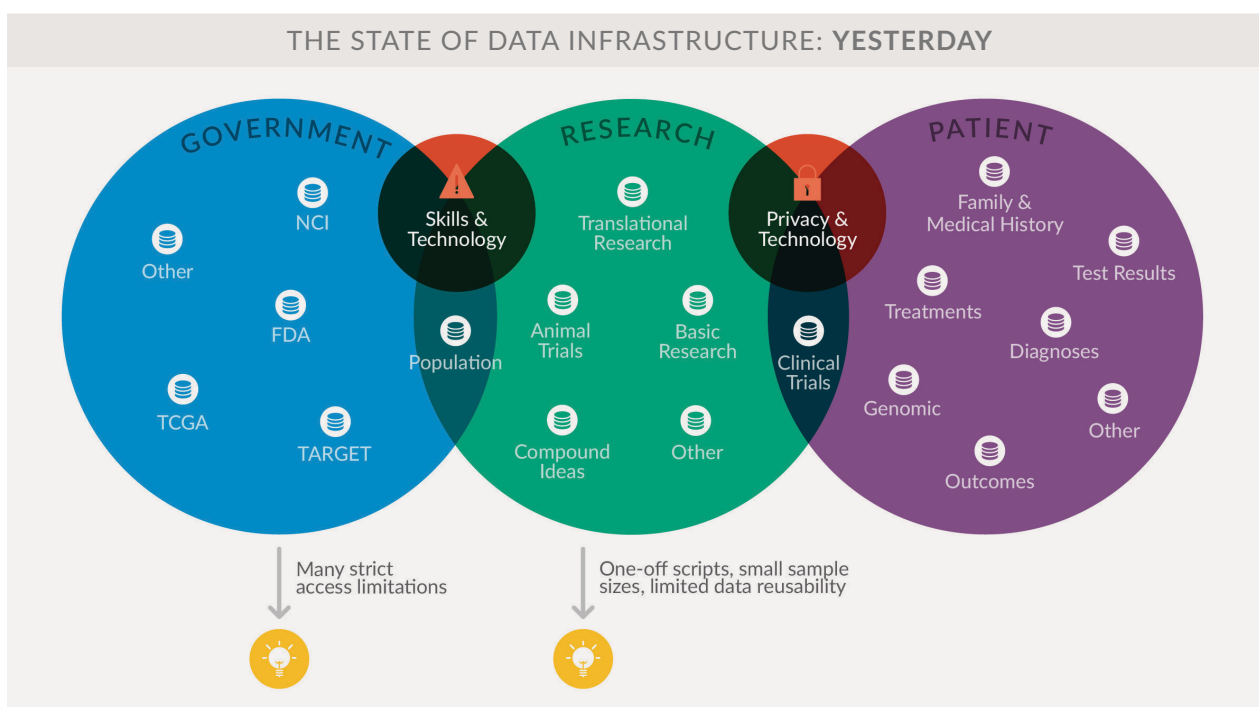
The cancer research space is not set up to apply data science methods to cancer data at scale. There are several promising government-funded computing initiatives under development that are good prototypes for the Moonshot:

- **The NCI's Genomic Data Commons:** This recently launched initiative to centrally house diverse cancer data is a promising step in the direction of creating a central data repository, but it is still in its early stages. In the short term, the government should continue to support the GDC and work to expand its capabilities and the types of data that it houses.
- **The NCI Genomic Cloud Pilots:** These programs provide cloud-based analytics resources.<sup>7</sup> They give researchers the capacity to work with large, complex genomic data at scale. However, they are still in early phases and are primarily focused on data from TCGA.

We see the GDC and the Cloud Pilots as exemplary models for today's cancer data infrastructure and analytics initiatives. Throughout the document and in our recommendations, we refer to a central repository and analytics platform in broad terms; in the short term, we see this work occurring within the scope of the GDC and Cloud Pilot initiatives. In the medium- to long-term, the NCI should evaluate whether these are the best models to aggregate all necessary data, or whether a successor system would be better at delivering on the potential outlined in this document.

## What are the problems?

In order to do large-scale data analytics, a researcher needs robust server capacity to store and analyze data. This is technologically demanding, expensive, and beyond the capacity of most research teams. In our interviews, we found that institutions that could afford these high costs tended to create their own analytics centers and only share resources with a select set of collaborators. This leads to duplicated effort in setting up distinct servers, downloading data, and sharing information.



*Researcher access to government data is complicated by strict access limitations, and a skills and tools gap in the research community prevents efficient analysis of cancer data. Privacy regulations slow the flow of patient-level data into research, and technological barriers prevent the efficient transfer of patient-level data.*

Another major challenge is that data must be standardized before it can be used in models or analyses. Genomic data is highly sensitive to so-called “batch effects” (variations that occur as a result of inconsistent collection and processing), and it needs to be combined in sophisticated ways that require a deep understanding of both biology and data science.<sup>8,9</sup>

Storing and analyzing data to do large-scale data analytics is technologically demanding, expensive, and beyond the capabilities of most research teams.

EMR data presents an even larger challenge for integration. Hospitals use different vendors, collect different types of information and metadata over the course of treatment, and export data from systems in ways that are not necessarily interoperable.<sup>10</sup> Many hospitals don't produce machine-readable output files. There is an entire industry devoted to standardizing EMRs: one nonprofit organization told us that they go through an intensive two to six week “informatic mapping engagement” when working with a new hospital network's system in order to align it with other EMR data. While there's immense value in using EMR data for research, it shouldn't be the job of cancer biologists to build algorithms to clean up data.

While there's immense value in using EMR data for research, it shouldn't be the job of cancer biologists to build algorithms to clean up data.



## Infrastructure Solutions

### SHORT-TERM

**Provide additional funding for the GDC's efforts to incorporate TCGA, TARGET, and other public datasets into a central repository accessible to credentialed researchers.**

- Fund upgrades so that the central repository continues to harness state-of-the-art storage technologies.
- Finalize incorporation of TCGA and TARGET, establish future checkpoints, and ensure resources scale with growing amounts of cancer data.
- Require that researchers working on relevant publicly funded research store study data and the code required to replicate their findings at the central repository after publication.

**Establish a Center of Excellence on cancer data harmonization within the Department of Health and Human Services. This center should have a mandate to provide tools, guidance, and training for harmonization and use of EMR and genomic data.**

- Establish relationships between the Center of Excellence and the GDC/Cloud Pilot programs so that they can build on the strengths of one another.
- Leverage the digital coalition (software engineering, product design, and user experience fellowships within the US Digital Service, Presidential Innovation Fellows, and 18F) to aid the development and maintenance of the central repository and similar programs.
- Use existing resources, including the NIH's Big Data to Knowledge (BD2K) program, to build on this work.

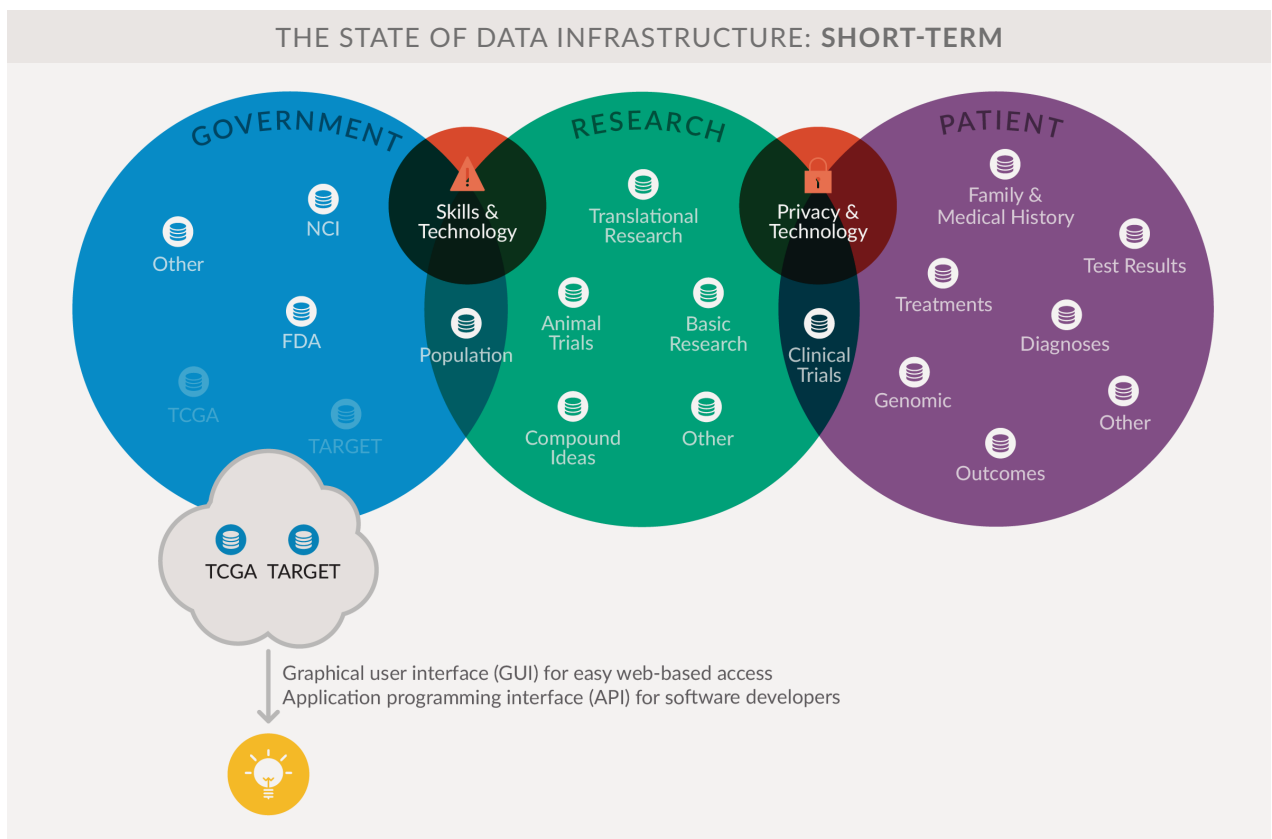


Use publicity around the Cancer Moonshot to build awareness of the central repository in the cancer research field.

- Promote the central repository within the cancer community through talks, demonstrations, and training sessions for potential users.
- Fund research that uses data from the central repository and demonstrates its value.
- Incorporate usage of the central repository into standard biology graduate program curricula.

Establish a pilot program with one hospital to develop a proof-of-concept for how EMR data can be harmonized and brought into a central location.

- The new Center of Excellence should work with one research hospital to build a prototype tool to export EMR data into a centralized repository.
- Focus this EMR pipeline tool on standardizing/harmonizing the EMR data for future research use.



*In the short term, the Moonshot should focus on bringing large publicly funded genomic datasets (such as TCGA and TARGET) into the central data infrastructure. The system should have an accessible GUI and an API for software developers.*

## MEDIUM-TERM

Expand the EMR harmonization pilot program to a larger network of hospitals.

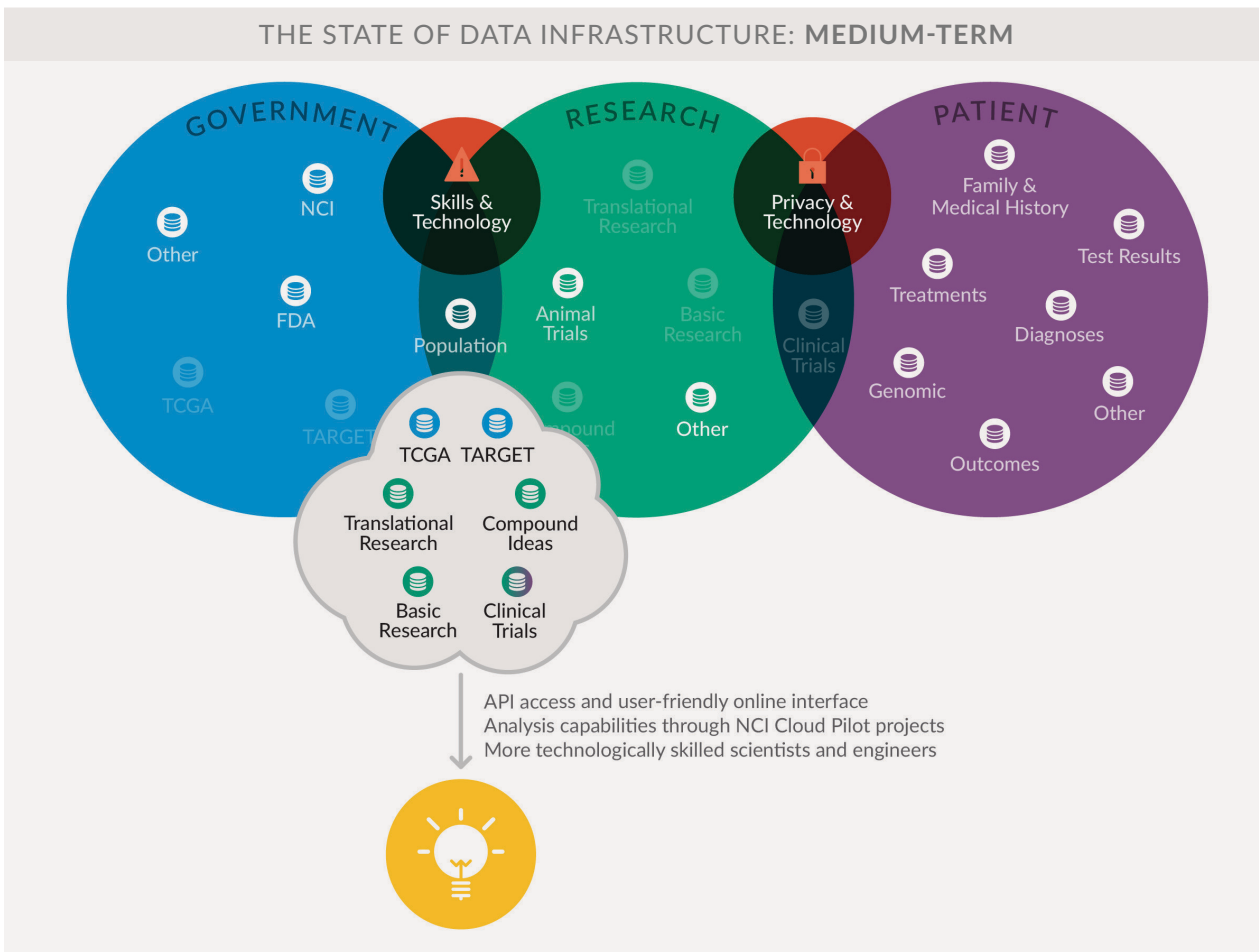
- The Center of Excellence should build upon past work to further develop tools allowing hospitals to export EMR data into a centralized repository.
- Learn best practices from the pilot with a single hospital, and recruit other hospitals to participate.

Evaluate the capacity of the GDC to grow alongside the volume of relevant data. If the GDC is not equipped to continue this work, identify an alternative solution.

- When evaluating the GDC, focus on who is using, what data has been successfully integrated, and how this data is being accessed.
- Determine what resources are necessary to continue using the GDC as a central repository, and study options for upgrading to a successor system that expands upon the capabilities of the GDC.

Incorporate clinical trial data, registry data, and other formatted datasets into the central repository.

- Explore data sharing agreements that would allow information exchange between the central repository and other government sources, including clinicaltrials.gov, clinical trial data submitted to the FDA for drug approval, datasets maintained by the Department of Veterans Affairs or the Centers for Medicare & Medicaid Services, and those associated with Diabetes and Alzheimer’s research.
- Build out capabilities for sharing, storing, and analyzing the genomic and EMR data of healthy individuals (or samples of healthy tissue from patients with cancer).
- Partner with major sequencing centers to incorporate genomic data into the central repository with patient consent.



*In the medium term, the Moonshot should focus on expanding the number of datasets held in the central repository and growing pilot programs for data integration and harmonization to more institutions. Additionally, the Moonshot should hire people with the necessary skills to build out a scalable infrastructure and prepare to provide access to this data to researchers nationwide.*

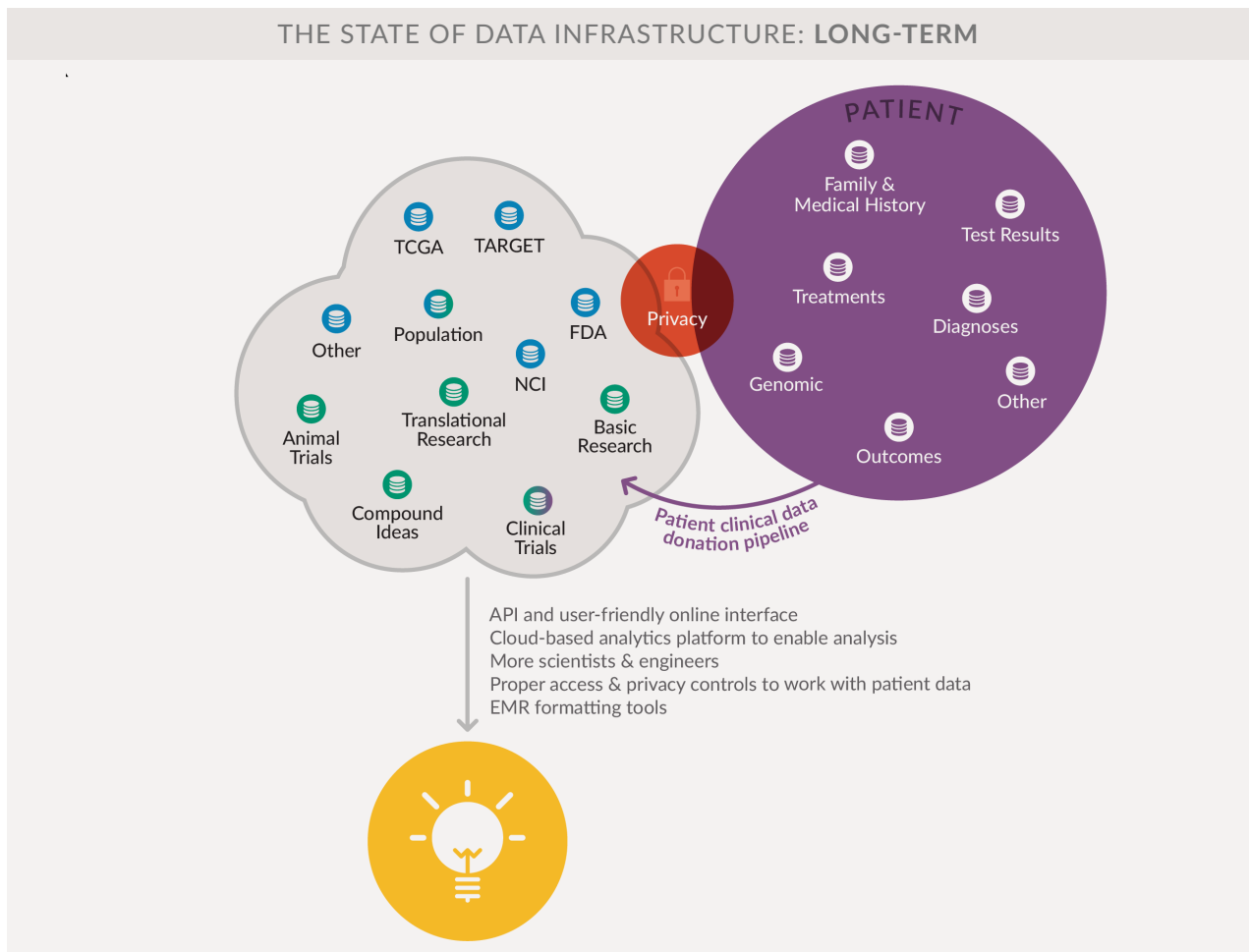
## LONG-TERM

Expand the EMR harmonization pilot program nationwide to bring EMRs and other patient-donated data into the central repository.

- Using best practices from past pilot programs, scale up the pipeline to incorporate data from throughout the U.S.
- Train and credential researchers nationwide to use the system and its datasets.
- Ensure infrastructure will continue to scale as the amount of data grows.

Invest in a cloud-based analytics platform to work in parallel with the central repository.

- Develop infrastructure for researchers to perform analyses (including tools built as part of the NCI Cloud Pilots) on central repository data without downloading the data locally.
- The NCI should identify best practices for what a final cloud system should look like, and evaluate potential new applications of the technology.
- Prioritize private and secure computing environments, collaboration potential, and extensibility by following industry best practices. Ensure that the storage and analysis systems work together through an Application Programming Interface (API),<sup>11</sup> which would provide researchers a standardized way of accessing data stored in the central repository from within the cloud-based analytics platform.



*In the long term, the Moonshot should incorporate public datasets, research datasets, and individual-level patient-donated data into the central infrastructure, and develop a cloud-based analytics platform that enables analysis using the data in the repository.*



# Facilitate Data Sharing

*There are three central barriers that inhibit sharing and will pose challenges to data integration: (1) ambiguity surrounding what constitutes compliant data sharing under privacy regulation, (2) inconsistent data formatting standards, and (3) academic incentives that motivate researchers not to share data. To reduce uncertainty surrounding privacy regulation, the DHHS should provide model protocols for compliant data sharing. In the long run, the government should begin a conversation about reforming HIPAA to make it more transparent, research-friendly, and consistent with other rules, such as state privacy laws and the DHHS Common Rule. To deal with interoperability challenges, the government should create and promote standard data formats. Finally, the government should establish more robust data sharing practices for academia.*

## Why it's important

Data science in cancer research requires large amounts of data, and no single institution is capable of collecting an exhaustive data store of every cancer and every patient. While building infrastructure is an important prerequisite for bringing cancer data together, the government also needs to revisit the regulation and culture around data sharing so researchers can build the datasets they need.

## Where we stand

Protecting patients' privacy and ability to give informed consent for research is an important consideration for any data science project in the field. HIPAA and state privacy regulations are complex, and require separate informed consent for each use of identifiable patient data for research.

EMRs are an important source of data for large scale analytics in the cancer space. Inconsistent data formats make it difficult to aggregate individual-level data from across data silos.

Finally, academic data is not being shared in a consistent manner into a centralized data repository. While some data sharing requirements for researchers exist, processes vary widely and impose a significant burden.

## What are the problems?

### 1. Privacy policy-compliant data sharing

Patient privacy is as important now as it was when HIPAA, the DHHS Common rule, and state privacy laws were enacted; hospitals, Institutional Review Boards (IRBs), and researchers rightfully take measures to protect patient data. However, the interpretation of HIPAA—especially regarding anonymization or de-identification of data—is unclear.<sup>12,13</sup> This ambiguity leads institutions to default to conservative interpretations of privacy regulations that don't necessarily translate to better privacy or security, but do slow down research.

Broadly speaking, HIPAA limits the extent to which identifiable patient health information can be used for research without patient authorization.





Inconsistencies in de-identification standards can actually lead to worse patient privacy and security outcomes. Many institutions opt to use third parties to certify de-identification, but not all experts use the same methods to ensure the data is sufficiently de-identified.<sup>14</sup>

Institutions often implement restrictive policies when guidelines under HIPAA are unclear because the burden of liability falls on “covered entities” and their “business associates” (e.g., health insurers, health care providers, and health care clearinghouses).

HIPAA and associated regulation play an indirect role in shaping research protocols by providing relevant guidance to IRBs.

- For example, we spoke with an academic researcher who said that because of privacy-related regulations at his institution, colleagues were not allowed to use email attachments to exchange any information at all, and had to rely on cumbersome external hard drives to go about their work.
- Another researcher told us that—due to concerns about re-identification of patients—research proposals that included a patient’s date of birth took much longer to approve than protocols that included a patient’s age.

## 2. Clinical data sharing

A major obstacle to integrating genomic and clinical data at scale is the lack of EMR interoperability. Currently, different hospitals use different EMR software, which do not necessarily output data in a consistent or machine-readable format. Additionally, hospitals vary in the types of information they collect and how they choose to store it. Integrating and making sense of all this information requires vast amounts of time and effort. The Fast Healthcare Interoperability Resources (FHIR) framework, a set of flexible and extensible standards for healthcare data, addresses many of the issues concerning clinical data export and formatting. However, it is not fully implemented and would likely take several years to propagate.

**“[A patient] should be able to view, download, and transmit the [his or her own medical] data. The transmit function there is kind of a road to nowhere—there isn’t a place where this data can be placed.”**

INTERVIEW SUBJECT FAMILIAR WITH MEDICAL DATA POLICY AND PATIENT DATA DONATION OPTIONS

One way to break down clinical data silos without upending HIPAA and related regulations is to empower patients to donate their health data to research. Currently, patients do not have an easy way to consolidate their medical information across providers into a single file, and hospitals and providers have not invested in the technology make this possible. Although there are some isolated attempts to enable data donation,<sup>15</sup> these efforts are underutilized. Someone familiar with medical data policy said of the Sync for Science initiative, “[A patient] should be able to view, download, and transmit the [his or her own medical] data. The transmit function there is kind of a road to nowhere—there isn’t a place where this data can be placed.”

The question of “ownership” over patient data and any downstream products (e.g., patents, new treatments) resulting from that data is a large area of contention in the field. In developing a mechanism for patient-driven data donation, DHHS will need to consider what rights patients should have over any data that is donated and what mechanisms should be in place for revoking consent after donation.

Additionally, many companies offer genetic testing, but the data that they collect is not typically owned by consumers and patients—instead, it is owned by these companies. This further siloes data, and it presents a barrier to both data sharing and patient-driven data donation.

### 3. Academic data sharing

Sharing academic research data is harder than it should be and has a low priority relative to other academic research tasks. Academia rewards publication, not data liquidity. Researchers earn prestige and grant funding by publishing as much as possible, and they are justifiably wary of sharing data they invested a lot of effort into collecting.<sup>16</sup> Further, much of the published work and related data sit behind paywalls. Finally, it is hard to make a dataset available, and researchers would rather spend their grant dollars on additional research than on wrangling datasets for the use of others. Although there are some mechanisms in place that compel researchers to share data—including the NIH Genomic Data Sharing Policy, requirements set by academic journals, and institutional norms—that data is not centralized.

Academia rewards publication,  
not data liquidity.



## Sharing Solutions

### SHORT-TERM

**Create HIPAA—and privacy regulation—compliant research protocols that institutions can use to guide their own work, minimizing ambiguity about legal interpretation.**

- Convene a panel of experts to create a set of HIPAA- and Common Rule-compliant data sharing and usage protocols in order to reduce the uncertainty surrounding privacy regulations, de-identification, and informed consent for patients.
- These protocols would reduce bureaucratic red tape and uncertainty by establishing “safe harbors” for compliant research institutions.
- In addition to focusing on federal regulation, the panel should consider state laws, which often pose a more onerous burden than HIPAA.

**Support the finalization of FHIR and EMR interoperability standards through incentives and regulation.**

- Continue to promote FHIR for hospitals using EMRs and encourage EMR vendors to comply.
- In consultation with researchers, expand and standardize the list of data fields that are required in standard EMR exports under the HITECH Act.
- Mandate that all EMR exports of patient data be machine-readable.

## MEDIUM-TERM

**Make patient data donation for research a reality by creating publicity around donation programs and working with trusted intermediaries—nongovernmental actors who would facilitate this process—to bring patient data into the central platform. This work would build upon the efforts of the White House Precision Medicine Initiative.**

- Support and publicize efforts that allow patients to contribute their EMR and genomic data to research via a “donate my data” button. Sync for Science, the Metastatic Breast Cancer Project’s Count Me In initiative, and the Blue Button initiative are all good examples of this concept, but there is little public awareness about them.
- Establish clear options for which types of data patients could donate through this pipeline and how researchers could use it.
- Develop a set of legal and technical policies that outline the parameters under which patients can revoke their consent for data they have donated.
- Provide funding and guidance to organizations that help patients donate their data and move it into the central data repository.
- As part of the patient data donation pipeline, establish protocols for refreshing periodically for “living” data associated with patients who are alive and still being treated.

**Incentivize researchers to use shared data by providing grant money to those who use and contribute data.**

- Provide federal research grants to studies that use and contribute data from/to the central repository.
- Replicate the NIH Genomic Data Sharing Policy for non-genomic data.
- Require that researchers receiving federal funding share data into the central repository.
- Direct the Center of Excellence to study and build tools that focus on:
  - » Making current datasets more useful and accessible to researchers
  - » Generalizing the prototype of the EMR pipeline for larger-scale use

## LONG-TERM

**Engage the research, medical, patient advocacy, and legal communities in a conversation about necessary reforms to HIPAA, its interplay with state privacy laws, and the Genetic Information Nondiscrimination Act (GINA).**

- Potential reforms include:
  - » Shifting strict liability to a model of gross negligence so that organizations are not driven to create unreasonable barriers to research that do not have added security benefits.
  - » Criminalizing re-identification of patient data.
  - » Making it illegal under GINA to discriminate against someone on the basis of their genetic or health information in fields where it is not already illegal to do so.
  - » Mandating that default ownership of genomic data sequenced by commercial actors remains with the individual whose genome is sequenced, specifying that companies doing the sequencing cannot withhold the data from the patient upon request, and mandating that they must destroy the data if requested to do so by the individual whose DNA it is.



# Invest in People and Skills

*The government should equip the next generation of academic researchers with the data science skills they need to use existing data through training, funding, and incentives. Additionally, the government should fund professional positions that support the functioning of the GDC and cloud-based analytics platform infrastructure, including systems engineers and data engineers.*

## Why it's important

Neither centralized infrastructure nor better data sharing will deliver sufficient value if researchers do not have the training required to analyze extremely large datasets or if the research community lacks professionals that can maintain the infrastructure in the long term.

## Where we stand

Most researchers in the cancer field were trained in the analysis of small datasets generated in laboratory settings. However, analyzing the types and volume of data available today requires different methods of statistical analysis. Graduate programs in the biological sciences have recognized the need for training in these methods, but it will take a generation for current students to make their way fully into the research community.

In addition to equipping researchers with statistical and computational skills, there also needs to be a concerted effort to develop tools so that the existing workforce can do technical tasks more easily. While some tools exist,<sup>17</sup> they are geared at advanced users, and many make the critical assumption that researchers have the computational resources to store and analyze any genomic data they download.

Finally, realizing the goals for the central data infrastructure and cloud-based analytics platform outlined in this document will require skilled systems engineers, developers, and designers to ensure that these tools are performing optimally.

## What are the problems?

Working with large datasets requires specialized skills that are not common in the research field. Many researchers rely on a small pool of bioinformaticians to process large data sets for them, which slows down the pace of discovery.<sup>18</sup> This divide makes it difficult to answer important questions.<sup>19</sup>

Many researchers rely on a small pool of bioinformaticians to process large datasets for them, which slows down the pace of discovery.



Despite the importance of skilled technologists, their value is often overlooked by leaders in the field. These leaders often have expertise with smaller datasets gathered in laboratory settings, but lack familiarity with techniques for large-scale analysis, making it difficult for them to recognize the scale of the human capital shortage facing the community. Leaders who do decide to hire individuals with these skills face a competitive marketplace because these skills are highly sought in other data-intensive sectors of the economy. Without the money to attract these professionals, the cancer community will continue to struggle to hire the necessary talent.<sup>20</sup> The cancer community must also retain and nurture existing talent. Non-tenure track technical experts often see no path for advancement in the research field and are instead treated as ancillary technicians.

**The cancer community must also retain and nurture existing talent.**

Additionally, the research community does not have the human capital to develop and maintain the infrastructure and software to handle large-scale datasets. This shortage will directly impact the potential success of the central data infrastructure (e.g., the GDC), and the cloud-based analytics platform outlined in this document. The community needs dedicated data engineers to ensure data quality, systems engineers to design and maintain scalable infrastructure, and user experience experts to ensure that tools are easy to use and accessible.

## People and Skills Solutions

### SHORT-TERM

**Staff the GDC and the cloud-based analytics platform with individuals who have key engineering, data management, and design skills.**

- These people include, but are not limited to:
  - » Systems engineers to manage the central servers.
  - » Software engineers to develop APIs, build software, and integrate with the APIs of other platforms.
  - » User experience designers to ensure that tools are accessible and clear.
  - » Project managers to coordinate development and testing of resulting products.
- Fund software engineering, product design, and user experience fellowships within the NCI to support this work.

**Support programs that give researchers data science skills to perform large-scale analysis.**

- Invest in graduate programs in biology that include data science in their curricula in order to widen the pipeline of future talent.
- Provide funding to universities and research institutions to hire individuals specializing in statistical techniques and software development (e.g., statisticians, bioinformaticians, computer scientists).
- Support training in data science geared at mid-career biologists to equip them to work more independently and comfortably with data.
- Incorporate foundational data science education into standard medical school training.

**Support the development of “staff scientist” roles at academic and nonprofit institutions. These would be designated individuals who would provide guidance on bioinformatics and data science to researchers.**

- These individuals would have defined career trajectories outside tenure, but would not be researchers; they would serve as experts in bioinformatics and be competitively compensated.

## **MEDIUM-TERM**

**Expand the pipeline of students going into data science in the cancer industry through loan forgiveness and training programs for students and mid-career professionals.**

- Extend the federal Public Service Loan Forgiveness Program to include high-need technical specialties (e.g., bioinformatics, statistics) as qualifying public services.
- Support continuing education for researchers in the field regarding working with and analyzing data.

## **LONG-TERM**

**Partner with the private sector to encourage the flow of skills and experience from the tech industry to the cancer research space.**

- Create fellowships and bring technologically skilled individuals into the cancer industry on sabbatical to solve specific problems. These fellowships will build awareness for the challenges faced in cancer research and also harness needed skills to improve the state of the field.
- Create awards that recognize creative applications of data science to the cancer field.



---

# Conclusion

*The problems facing the cancer research community are serious, but they coincide with great opportunities. The research community has access to genomic, clinical, and other data at a scale unimaginable even ten years ago. In order for the field to take full advantage of these resources, the government must play a key role in establishing institutions and norms in the cancer space.*

***In order to solve the problem of inadequate data infrastructure, the government should:***

- 1. Develop a centralized data repository based on the GDC.*
- 2. Create a Center of Excellence for data standardization.*
- 3. Build a large-scale cloud-based analytics platform for researchers to use for computation.*

***To address problems that affect data sharing, the government should:***




- 1. Develop and enforce EMR formatting conventions.*
- 2. Provide guidance to research institutions on compliant data sharing and research protocols under HIPAA and relevant privacy regulation.*
- 3. Support a patient-driven data donation model through funding and publicity for partner organizations that can facilitate the data hand-off from the patient to the central repository.*
- 4. Facilitate academic data sharing.*
- 5. Begin a conversation with relevant stakeholders about necessary reforms to HIPAA and other outdated privacy laws.*

***Finally, to close the skill gap in the cancer space, the government should:***

- 1. Invest in data science training programs for researchers at all career levels.*
- 2. Staff the central data infrastructure with the skilled professionals it needs to succeed.*
- 3. Support partnerships with the tech industry to transfer knowledge and skills.*

***While these changes are not easy, collectively they will establish the conditions for success in the field. The Moonshot represents a once-in-a-generation opportunity to effect systemic change in the cancer research space, prepare the community for future potential, and ultimately save lives.***

# Timeline of Recommendations

	 <b>DATA INFRASTRUCTURE</b>	 <b>DATA SHARING</b>	 <b>PEOPLE AND SKILLS</b>
<b>SHORT-TERM</b>	<p>Provide additional funding for the GDC's efforts to incorporate TCGA, TARGET, and other public datasets into a central repository accessible to credentialed researchers.</p> <p>Establish a Center of Excellence on cancer data harmonization within the Department of Health and Human Services. This center should have a mandate to provide tools, guidance, and training for harmonization and use of EMR and genomic data.</p> <p>Use publicity around the Cancer Moonshot to build awareness of the central repository in the cancer research field.</p> <p>Establish a pilot program with one hospital to develop a proof-of-concept for how EMR data can be harmonized and brought into a central location.</p>	<p>Create HIPAA and privacy regulation compliant research protocols that institutions can use to guide their own work, minimizing ambiguity about legal interpretation.</p> <p>Support the finalization of FHIR and EMR interoperability standards through incentives and regulation.</p>	<p>Staff the GDC and the NCI with individuals who have key engineering, data management, and design skills.</p> <p>Support programs that give researchers data science skills to perform large-scale analysis.</p> <p>Support the development of "staff scientist" roles at academic and nonprofit institutions. These would be designated individuals who would provide guidance on bioinformatics and data science to researchers.</p>
<b>MEDIUM-TERM</b>	<p>Expand the EMR harmonization pilot program to a larger network of hospitals.</p> <p>Evaluate the capacity of the GDC to grow alongside the volume of relevant data. If the GDC is not equipped to continue this work, identify an alternative solution.</p> <p>Incorporate clinical trial data, registry data, and other formatted datasets into the central repository.</p> <p>Incorporate clinical trial data, registry data, and other formatted datasets into the central repository.</p>	<p>Incentivize researchers to use shared data by providing grant money to those who use and contribute data.</p> <p>Make patient data donation for research a reality by creating publicity around donation programs and working with trusted intermediaries—nongovernmental actors who would facilitate this process—to bring patient data into the central platform. This work would build upon the efforts of the White House Precision Medicine Initiative.</p>	<p>Expand the pipeline of students going into data science in the cancer industry through loan forgiveness and training programs for students and mid-career professionals.</p>
<b>LONG-TERM</b>	<p>Expand the EMR harmonization pilot program nationwide to bring EMRs and other patient-donated data into the central repository.</p> <p>Invest in a cloud-based analytics platform to work in parallel with the central repository.</p>	<p>Engage the research, medical, patient advocacy, and legal communities in a conversation about necessary reforms to HIPAA and the Genetic Information Nondiscrimination Act (GINA).</p>	<p>Partner with the private sector to encourage the flow of skills and experience from the tech industry to the cancer research space.</p>





---

# Key Definitions

## **Batch effects**

Genomic data often suffers from so-called “batch effects,” whereby small variations in the processing and sequencing of genetic material result in differences in the resulting output. These differences introduce bias and prevent direct comparison of genomic data. They can also lead to inaccurate conclusions, particularly when batch effects are correlated with a variable of interest. To overcome this problem, large-scale genomic data from multiple sources needs to be normalized before analysis.

## **Cancer registry**

Under federal law, state and local cancer registries are required to collect and manage data on every diagnosed case of cancer. Specific variables collected include patient demographics, such as age, race, and gender, as well information on the type of tumor, its location, treatments used, and patient health outcomes. Registries aggregate the resulting data for research purposes. Cancer registry data is collected at the hospital level and reported up to local, state, and national registries, including the National Program of Cancer Registries, NCI’s Surveillance, Epidemiology, and End Results (SEER) program, and the privately-funded National Cancer Database. Any data shared with the public (to report cancer incidence or mortality statistics, for example) is de-identified and disassociated from the patient.

## **EMR - Electronic Medical Records**

Digital information on a patient’s interactions with a single health care provider; the equivalent of a digitized paper chart.

EMRs differ from electronic health records (EHRs) in that the latter include information from multiple health providers.

## **FHIR - Fast Healthcare Interoperability Resources**

A framework proposed by the American National Standards Institute (ANSI) for electronic health records.

## **GDC - Genomic Data Commons**

Computational facility to be hosted by University of Chicago. Its goal is to harmonize genomic data generated by all NCI funded studies, and provide computational resources for analysis to the scientific community.

## **DHHS Common Rule - U.S. Department of Health and Human Services Common Rule**

Regulation governing data sharing. Under the Common Rule, research protocols that use human subjects must be approved by an IRB, and researchers must obtain informed consent.

Only personally identifiable information falls under the scope of “human subjects research;” however, there is no standard for what constitutes de-identification under the DHHS Common Rule. The Common Rule allows individuals to give consent for subsequent research so long as that research is described in sufficient detail to constitute informed consent.

## **HIPAA - The Health Insurance Portability and Accountability Act of 1996**

HIPAA establishes rules that protect patient health data. It applies to “covered entities,” which are medical providers and others who deal directly with patient data, as well as “business associates” who work with these organizations. Health information is only covered under HIPAA if it is identifiable, and HIPAA provides two ways that data can be certified as de-identified. Organizations may either strip 18 key identifiers from the data, including any “unique identifying number, characteristic, or code,” or they may have the information certified as de-identified by a qualified expert. When information is not de-identified, it may not be disclosed without written consent except as permitted in the law. HIPAA contains an exemption to this written consent rule when information is being shared for the purposes of “treatment, payment, or operations.”

**HITECH Act - Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009**

Enacted as part of the American Recovery and Reinvestment Act of 2009, the HITECH Act supported the digitization of health records in the U.S. It led to the expansion of EMR usage.

**IRBs - Institutional Review Boards**

Internal committees at research institutions charged with reviewing research protocols and ensuring that experiments meet all relevant regulation and ethical guidelines.

**NCI - National Cancer Institute**

A subsidiary of the National Institutes of Health (NIH) responsible for supporting research, education, and public health activity around cancer.

**NCI Cloud Pilots**

The NCI is supporting the development of three cloud-based data analytics platforms which offer access to TCGA and computational resources to researchers. The Cloud Pilots are intended to support large-scale analysis of genomic data and “bring the researchers to the data.” The organizations running the pilots are the Broad Institute, the Institute for Systems Biology, and Seven Bridges Genomics.

**NIH Genomic Data Sharing Policy**

A policy that applies to all NIH-funded research that generates genomic data. Under the policy, researchers must make data available no later than the publication date. Additionally, they are encouraged to obtain broad consent from research participants for use of the data in future research.

**TARGET - Therapeutically Applicable Research to Generate Effective Treatments**

An initiative that produces large-scale genomic data for pediatric cancers. This project is run by NCI's Office of Cancer Genomics and Cancer Therapy Evaluation Program, and the data is being incorporated into the GDC.

**TCGA - The Cancer Genome Atlas**

A project funded by NCI and the National Human Genome Research Institute that aims to bring together genomic data to identify the gene variants that cause cancer. Data in TCGA is available in three tiers, from Level 1, which is raw personally identifiable data, to Level 3, which is aggregate level. TCGA is focused on a few dozen kinds of cancer and has over 10,000 cases of tumor and normal tissue sequences. It can be accessed formally through the TCGA Data Portal and the Cancer Genomics Hub.

**USDS - United States Digital Service**

An initiative at the White House that provides technological support for the federal government.

**White House Precision Medicine Initiative**

White House initiative aimed at advancing the state of individualized medicine through the creation of a large scale research cohort and funding for advancements in genetic research.



---

# Endnotes

1. In total, we had ten conversations with researchers and administrators at academic and governmental institutions, seven with individuals managing cancer data systems, six with pharmaceutical companies, five with medical professionals, three with legal teams, two with policy advocates, two with EMR companies, two with hospital administrators, one with an insurance company, one with a cancer registrar, one with public policy experts, and several patients.
2. Estimate based on Amazon Web Services S3 pricing.
3. The GDC is a cancer data storage project launched in June 2016, sponsored by the NCI, and built by groups at the University of Chicago and the Ontario Institute for Cancer Research. It currently hosts TCGA and TARGET, two major genomic databases, and will be the future home of data from NIH-funded research.
4. In developing the cloud-based analytics platform, the NCI should build upon the lessons learned in the NCI Genomic Cloud Pilots.
5. Genomic data includes whole genomes, gene sequences, gene expression data, and tumor sequences and is created in several ways: (1) large research studies focused on collecting genomic data, like TCGA, (2) targeted research projects that focus on a particular type of cancer or population, (3) clinical settings, if a patient's genome is sequenced as part of diagnosis or treatment decision, and (4) the private sector. While the volume of genomic data is growing rapidly, there is currently an asymmetry of data; most people who are diagnosed with cancer do not undergo genetic sequencing. For further details, see Muir et. al, "The real cost of sequencing: scaling computation to keep pace with data generation", Genome Biology 2016.
6. Clinical data ("outcome data") about whether a person developed cancer, what type, how it was treated, and how they responded to specific treatments exists primarily in a patient's EMRs and—at a lesser level of detail—in cancer registries. The number of clinical offices using EMR systems has dramatically increased following passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, from under 20% in 2001 to more than 80% today.
7. The organizations running the pilots are the Broad Institute, the Institute for Systems Biology, and Seven Bridges Genomics.
8. Interview with Robert Grossman, GDC principal investigator; interview with Tom Summerfelt, VP of Research at Advocate Health Care.
9. A major pharmaceutical company built a pipeline to periodically download the entire TCGA database in order to analyze the data in-house because of inconsistencies in how institutions contributing to TCGA organize and annotate their data.
10. Interview with Tom Summerfelt; interview with the director of analytics at a large research hospital; interview with Julie Johnson, University of Chicago Center for Research Informatics.
11. An API provides programmatic access to data and/or systems, allowing developers to write software extensions.
12. Interviews with several security, compliance, and legal experts.
13. At present, there are two ways that patient data can be considered de-identified under HIPAA: (1) removal of 18 unique identifiers from the data, such as a person's name and date of birth, or (2) certification by a qualified expert, based on statistical and other techniques, that the risk of re-identifying any individual is very small.
14. Additionally, if a single organization combined data on the same patient population that had been de-identified in different ways with different levels of stringency, it could potentially use probabilistic statistical techniques to re-identify individuals who exist in both records.

15. Mandl and Kohane, "Time for a Patient-Driven Health Information Economy?," New England Journal of Medicine, 2016.
16. Interviews with two different researchers working in academia and at a major research hospital, respectively.
17. The most popular tool we encountered is cBioPortal for Cancer Genomics, which was developed by the Memorial Sloan Kettering Cancer Center to give researchers the ability to download and visualize TCGA data and upload their own data for analysis. Another tool is the GeneTorrent client, provided by the NCI's Cancer Genomics Hub and maintained by the University of California Santa Cruz, which also allows researchers to download genomic data from TCGA.
18. Interview with Amy Abernethy, Chief Medical Officer/Chief Scientific Officer at Flatiron Health
19. Interview with an MD/PhD who works in academia now, and formerly worked at the NCI.
20. Separate interviews with two principal data scientists at a major pharmaceutical company; interview with a professor at a major research university.