The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Forecasting Costs of Biomedical Data Preservation

A User Guide for Biomedical Researchers

Summary

Biomedical researchers are generating, collecting, and storing more research data than ever. Preserving those data in discoverable and accessible ways is increasingly important, though doing so generates costs that may be difficult to predict. Allocating responsibility for such costs may further complicate a research endeavor. This guide will help researchers identify and think through the major decisions in forecasting life cycle costs for preserving, archiving, and promoting access to biomedical data. The recommendations presented here reflect the in-depth analysis of the following report from the National Academies of Science, Engineering, and Medicine: *Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs*.[1]

Background

The costs of constructing, maintaining, and accessing biomedical data can vary widely. The National Library of Medicine of the National Institutes of Health tasked the National Academies of Sciences, Engineering, and Medicine with developing a framework for forecasting long-term costs for preserving, archiving, and accessing various types of biomedical data and estimating potential future benefits to research. The resulting National Academies report highlights major cost drivers for biomedical research information resources and puts forth steps for individuals and organizations to consider the life cycle costs associated with the data. This user guide summarizes several ways in which biomedical information resources may vary and how each variation is likely to affect costs or utility.

The life cycle of digital data typically involves the following three major states:

- **State 1: The Primary Research and Data Management Environment**
  In State 1, data are actively captured as they are created, and then analyzed. Those managing or using a State 1 data environment should be focused on standardizing, documenting, sharing, and preserving data and algorithms.

[1] National Academies of Sciences, Engineering, and Medicine. 2020. *Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25639.

- **State 2: An Active Repository and Platform**
  In State 2, data may be acquired, curated, aggregated, accessed, and analyzed. This is an active information system that usually provides services to a wide range of users. Data are acquired from the primary research environment, from another active repository, or may be revived from archival storage for active use.

- **State 3: A Long-term Preservation Platform**
  In State 3, content is preserved across changes in governance, assessment of data value, and technology. The platform may include an extract of data from a single data set, multiple data sets, or an information system in a system-agnostic format. In this state, data are neither directly analyzable nor easily accessible. Content (e.g., data and code) are preserved in a long-term preservation platform when it is anticipated that the data will not be actively used for the foreseeable future, or if the resources are not available to maintain an active repository.

Data take different forms in each state, and each state includes different activities with different personnel, hardware, and management requirements. It is important to note that the labor and computation needed to transform data from one state to another can require significant resources and data may not transition through the three states sequentially because of the unique needs of the research endeavor or repository.

Value of Data

The perceived value of data influences preservation, access, and archiving decisions as well as decisions made regarding transition of data from state to state. However, assessing the value of biomedical data is challenging and needs to extend beyond monetary costs.

The value of a single data set reflects factors such as its uniqueness, the number of times it is used, the cost per use, and the impact of reuse. The number of different tasks or decisions that the data support may be a good indicator of their value. Data valuation can also depend on the data being findable, accessible, interoperable, and reusable (FAIR), and data standardization and documentation play an important role.

Cost Forecasting

The cost of preserving and providing access to data depends on choices made throughout the data life cycle and on the presence of tools, institutional support, and incentives that affect those choices. These choices often predate the launch of an individual research project in which data are generated. Funder requirements, data management mandates, institutional review board specifications, federal regulations, and journal requirements can all influence costs across the data life cycle. Data management plans that incorporate costs and value across the data life cycle may reduce the cost and time required for later data deposit and sharing.

Cost forecasters will likely need to consult with multiple individuals with varied expertise to minimize uncertainty in the forecast. In most cases, the cost of long-term data preservation will not be accrued by a single individual or institution; the cost burden can shift over the course of the data lifecycle. Understanding where costs will be accrued and who has managerial responsibility for them will inform decision-makers for all data states.

Forecasting Framework

The framework presented here should be considered the basis of a cost forecast rather than a one-size-fits-all analytical tool for all applications. How it is applied in any situation depends on the circumstances, needs, and resources available to those involved. The activities, decisions, and cost drivers will be situationally dependent, and the framework will need to be modified to suit the specific purpose. In whatever application, however, the forecaster is encouraged to think beyond the costs associated with the specific data state being developed or managed. In the long term, it is more efficient to think early about how decisions may affect the costs of data management and access in future data states, the transitions to those states, and to the future value of data to the scientific enterprise.

The table on page 6 provides an overview of steps to direct a cost forecaster's efforts throughout the process of forecasting data costs for a biomedical information resource. This framework is meant to be a starting point for cost forecasters to begin analyzing the costs of their biomedical data resource. Even so, it is often helpful to see an example first. The committee put together a video think-aloud that walks a user through the thought process and mechanics of the framework.

To identify data characteristics, data contributors, and data users (step 2 in the table), the cost forecaster will need to work with her institution, project funders, and perhaps the broader research community to identify or develop appropriate metrics to better understand and manage costs. Consulting with information technology professionals, metadata librarians, software engineers, and many others may be necessary to compile the information necessary to identify the major cost drivers (step 5 in the table).

The primary cost drivers often relate to the following:

- **Content** – the amount, kinds, and qualities of data that a biomedical information resource is expected to host. Generally, the larger and more complex a data set, the more costly it will be. Costs can be lowered by greater compressibility and replaceability.
- **Capabilities** – what information resource users are able to do with the data therein. More functionality and capabilities for a data resource typically means greater costs.
- **Control** – aspects of a biomedical information resource that deal with control and oversight of the resource (e.g., quality control measures). Increased controls on the data or the repository result in higher costs.
- **External Context** – relationships between the biomedical information resource and other, external resources. Although cost relationships can vary, costs typically increase if the resource is replicated and if its content is relatively distinct.
- **Data Life Cycle** – aspects of a biomedical information resource's expected evolution over time. Longer-term costs will be incurred if the resource is anticipated to be updated or grow in size. However, outlining a useful life span and moving the resource to offline or deep storage can reduce costs.
- **Contributors and Users** – a biomedical information resource's users and their characteristics. The wider the audience for a biomedical data resource, the more costly it will be.
- **Availability** – expectations about the availability of the data in a biomedical information resource. This can encompass the reliability of the resource hosting the data, how quickly new data appear, how fast requests for data are serviced, and from where the data can be accessed. A resource which offers greater accessibility (both to the data and to user assistance) will have greater associated costs.
- **Confidentiality, Ownership, and Security** – data protection and the rights of those associated with the data. Taking measures to ensure higher confidentiality and security will increase costs. Costs may also be increased if multiple parties have ownership or rights to the data.
- **Maintenance and Operations** – obligations for maintenance and operation of the biomedical information resource. Frequent maintenance and more extensive risk mitigation efforts will drive up costs. Costs may or may not be offset by the possibility of charging for use of the resource.
- **Standards and Regulatory Compliance, and Other Governance Concerns** – community conventions, rules, policies, laws, and stakeholder concerns with which the operators of a biomedical information resource may have or want to comply. Greater oversight will incur greater costs, as will using more modern applicable standards.

Once the investigator has considered where the data are coming from, and how they will be used, he can begin to quantify the costs. To do so, the researcher can use the data set characteristics, activities, and cost drivers described above and in the cost driver workbook. Many of the activities and cost drivers in the template may not be directly applicable to a primary research and data management environment, but the forecaster needs to remain aware of potential future cost drivers so that decisions might be made that could keep life cycle costs low. In most circumstances, labor costs will be the largest single element of the cost forecast.

The reliability of a cost forecast is also an important consideration. Quantitatively assessing uncertainties may be difficult, but the cost forecaster should communicate concerns with decision-makers, even if they cannot be precisely characterized.

It is also important for cost forecasters to keep an eye towards emerging disruptors, which could radically change how research is conducted and data are collected, used, archived, or preserved. The uncertainty of the scale and speed of the adoption of emerging technologies (particularly for computational work and data storage and reuse) in the next 5 to 10 years is one example, as are changes in legislation and policy related to data. Disruptors may be positive, negative, or mixed, and could raise or lower the cost of data management and preservation. There is no way to anticipate the impacts of potential disruptors, but building flexibility into data planning can help to mitigate their effects.

Steps for Forecasting Costs of a Biomedical Information Resource

1. Determine the type of data resource environment, its data state(s), and how data might transition between those states during the data life cycle.	• Decide the goals and objectives for the data resource. • Consider how the resource is likely to be used now and in the future. • Identify available guidance that defines the type of resource to be created or managed. • Compare the above with the activities defined for each of the data states and decide which data state(s) best align(s).
2. Identify the characteristics of the data, the data contributors, and users.	• Fill in the cost driver template in order to o Identify the size, complexity, replaceability, and depth versus breadth of the data; metadata requirements; and processing levels and fidelity; o Identify the life cycle issues; and o Identify data contributors and users.
3. Identify the current and potential value of the data and how the data value might be maintained or increased with time.	• Consult with the institution hosting the data resource, the project funders, and the broader research community to develop appropriate metrics for assessing the value of the data. • Identify decisions that affect data value in the shorter and longer terms. • Consider how data generation methodologies affect short- and long-term data value in terms of data contributors and users and the data life cycle.
4. Identify the personnel and infrastructure likely necessary in the short and long terms.	• Identify the major activities and sub-activities associated with the information resource, including activities related to potential transitions between data states. • Identify short- and long-term staffing requirements for the current state and transition between states. • Identify the infrastructure requirements and available resources.
5. Identify the major cost drivers associated with each activity based on the steps above, including how decisions might affect future data use and its cost.	• Identify the major cost drivers and associated uncertainties for each of the activities identified above by completing the cost driver template (download here). • Identify likely relative costs. • Consult with institutional experts and determine available personnel and infrastructure resources. • Work with experts at the host institution to quantify short-term costs and to bound uncertainties in longer-term forecasts.
6. Estimate the costs for relevant cost components based on the characteristics of the data and information resource.	• Identify which cost drivers are important for each cost component of the information resource (e.g., labor, information technology infrastructure and services, media, licenses and subscriptions, facilities and utilities, outside services, travel, and institutional overhead). • Estimate costs for the current funding period. • Estimate costs and cost uncertainties for future funding periods, including costs to transition data to other states.

Community Next Steps

In addition to utilizing this cost-forecasting framework, the following strategies can help enable efficient long-term data management and effective cost forecasting:

- **Create data environments that foster discoverability and interpretability through planning and investment throughout the data life cycle.**
  Making data discoverable, interpretable, and reusable requires a lot of forethought and sustained long-term investment from those who manage the data. However, doing so can help alleviate some of the challenges associated with generating, using, and storing growing volumes of complex data.
- **Incorporate data management activities throughout the data life cycle to strengthen data curation and preservation.**
  Long-term data curation and data management needs are vital throughout the course of the primary research state. Thus, the biomedical research community must expand the focus of data management and curation activities to include the entire life cycle, not just the end of the funding period. Up-front costs may be increased, but data value may also increase, and the overall cost of research may be reduced.
- **Incorporate the expertise and resources needed to create and curate metadata throughout the data life cycle, and in the transition between data states, into the cost forecast.**
  It is up to everyone in the biomedical research community to promote, support, and improve the understanding of the expertise and resources required for proper metadata that facilitates data discoverability and interpretability in all data states. This will likely be more successful if researchers are involved in decision-making and preservation efforts. It will also be more efficient if researchers work with data librarians, who can help ensure adherence to community-accepted data and metadata standards.
- **Weigh the benefits, risks, and costs (both short- and long-term) of data storage and computation options before selecting among them.**
  Decision-makers should give substantial attention to several additional features of the data, regardless of the storage and computation options. Such features include confidentiality, ownership, and security; standards, regulatory, and governance concerns; access control; and the various disruptors listed in the report.

Lastly, the current system for funding research cannot accommodate data life cycle cost forecasting. It is increasingly important to develop the rules governing such a process and to educate the community about the value of implementing them. In doing so, cost forecasting can become an integral part of responsible conduct of research, as opposed to a bureaucratic chore.

Implementing this cost forecasting framework into the research funding system and the broader research community will require a cultural shift, which needs to be driven by community engagement. While oversight entities are in a better position to offer incentives for this change, the process must be driven by researchers so as to better meet their needs and so that they can fully understand and agree to the value returned to them for their efforts. Ultimately, this will benefit the scientific enterprise as a whole, as well as individuals whose well-being biomedical research seeks to advance.