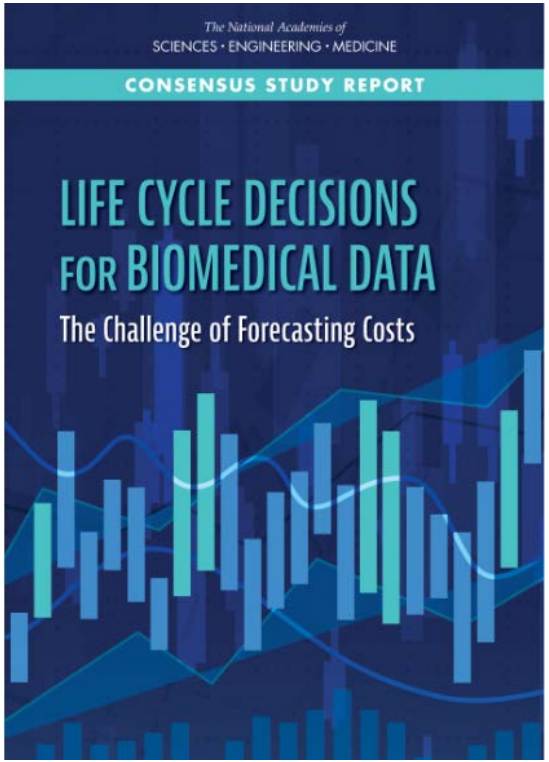
The National Academies of SCIENCES ENGINEERING MEDICINE

BOARD ON MATHEMATICAL SCIENCES AND ANALYTICS



Report Dissemination

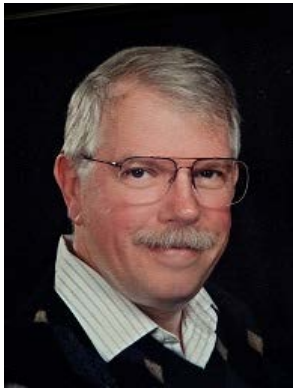
Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs

Presented to the Public
July 20, 2020

Committee Members Speaking Today

David Chu (Chair)
Institute for Defense Analysis

Alexa McCray
Harvard University



David Maier
Portland State University

Maryann Martone
University of California, San Diego

Statement of Task

National Library of Medicine of the National Institutes of Health asked for a framework for forecasting long-term costs for preserving, archiving, and accessing biomedical data.

Statement of Task (cont.)

- Economic factors to consider when examining the life cycle cost for data sets;
- Cost consequences for accessioning and deaccessioning data
- Economic factors in designating data sets as high value
- Assumptions built into the collection and modeling processes
- Anticipated disruptors in next 5-10 years; and
- Critical factors for successful adoption of forecasting approaches

Workshop: Planning for Long-Term Use of Biomedical Data (July 11-12, 2019)

NLM also requested a workshop to consider

- Tools/practices to integrate risk management practices into data-related decisions
- Methods to encourage life time data cost tracking
- Burdens on community to implement practices

Proceedings available:
https://www.nap.edu/catalog/25707/

The National Academies of SCIENCES ENGINEERING MEDICINE

Committee Membership

DAVID S.C. CHU, Institute for Defense Analyses, *Chair*

✵ ILKAY ALTINTAS, University of California, San Diego

G. SAYEED CHOUDHURY, Johns Hopkins University

MARGARET C. LEVENSTEIN, University of Michigan

✵ CLIFFORD A. LYNCH, Coalition for Networked Information

DAVID MAIER, Portland State University

CHARLES F. MANSKI, NAS, Northwestern University

MARYANN MARTONE, University of California, San Diego

ALEXA T. McCRAY, NAM, Harvard Medical School

✵ MICHELLE N. MEYER, Geisinger

WILLIAM W. STEAD, NAM, Vanderbilt University Medical Center

✵ LARS VILHUBER, Cornell University

Framework Foundation: Three Data States

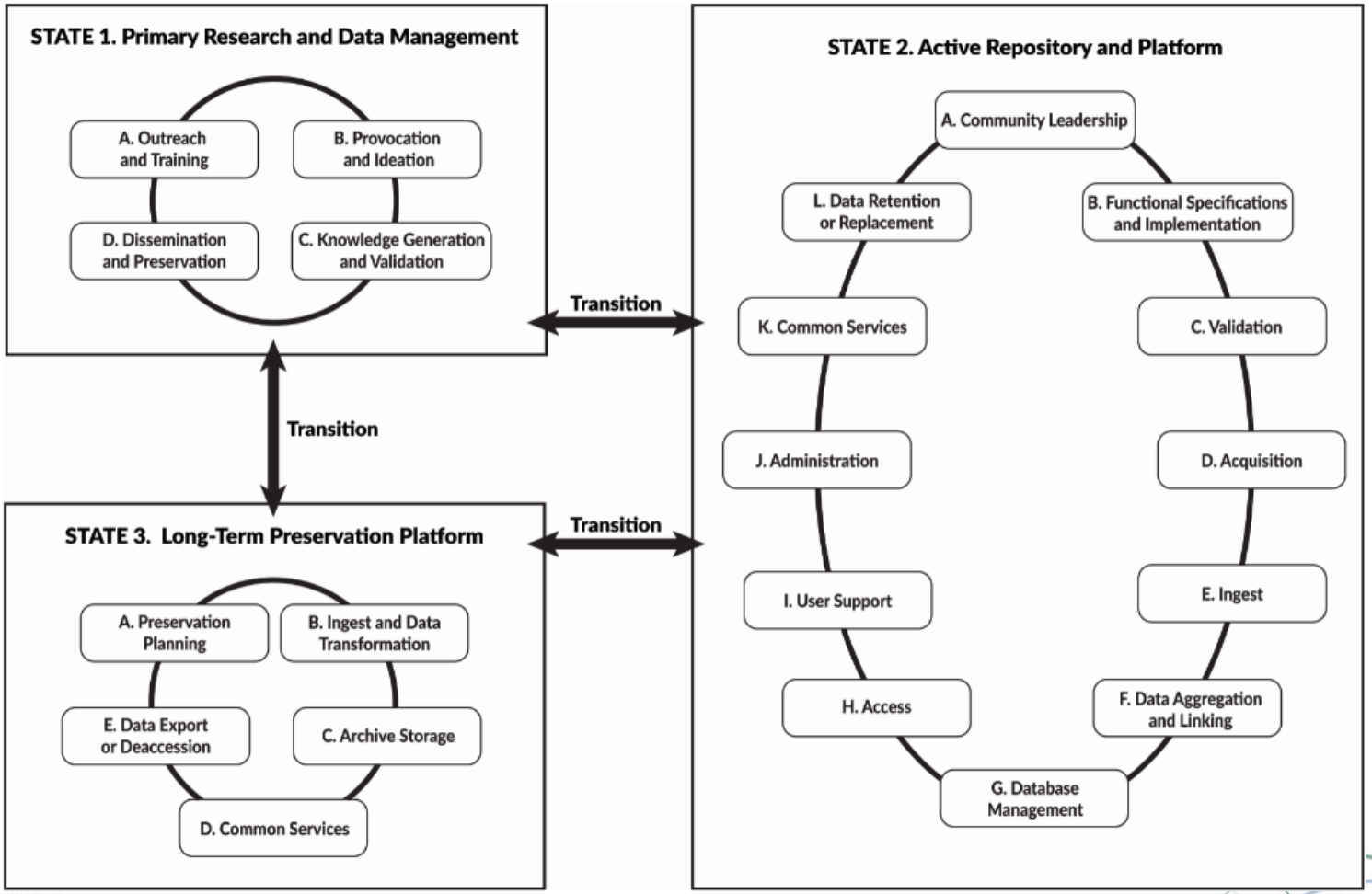
State 1: **Primary research**/data management environment; data are captured and analyzed

State 2: **Active repository** and platform; data may be acquired, curated, aggregated, accessed, and analyzed

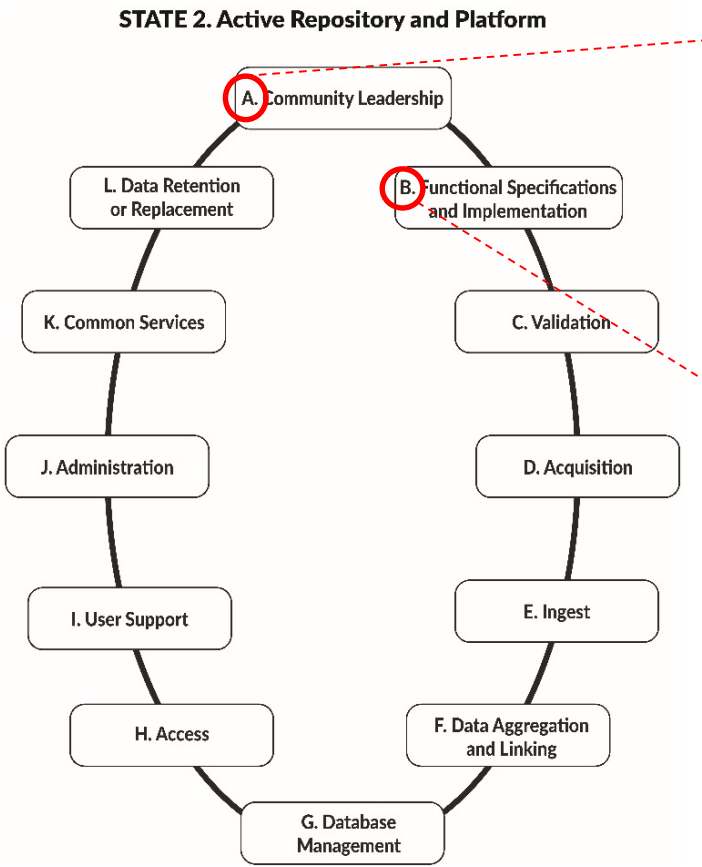
State 3: **Long-term preservation** platform

(Box 2.1 in text)

(Fig 2.1 in text)



Conceptualization of data states and activities



The National Academies of SCIENCES ENGINEERING MEDICINE

State 2: Active Repository and Platform

Activity	Subactivities	Personnel
A. Community Leadership Engagement with the broader community in the development of tools, standards, and best practices	1. Develop community data standards and best practices and policies. 2. Share lessons from development of repository systems and tools. 3. Identify community needs through community outreach.	Researcher, informatician, records management specialist, data librarian, communication specialist
B. Functional Specifications and Implementation Processes involved in designing or modifying and implementing the system for access and use	1. Design or modify and implement the repository infrastructure. 2. Consult with stakeholders on proposed design. 3. Design or modify and implement analytic tools. 4. Design or modify and implement search capabilities. 5. Design or modify and	Senior staff, software engineer, informatician, research domain project manager, IT project manager, IT security specialist

(Excerpt of Table 2.2)

Cost Forecasting Framework

- *Helps forecaster identify major cost drivers*
- *Basis for a cost forecast (not a one-size-fits-all analysis tool)*
- *Will help forecaster identify decisions that impact short- and long-term costs and data value*
- *Forecaster necessarily focuses on costs of current funding period, but must be aware that early decisions affect long-term costs of data curation and use*

"Steps" of the framework may occur concurrently or iteratively as new information is gathered

Cost Driver Categories

A. Content
B. Capabilities
C. Control
D. External Context
E. Data Life Cycle
F. Contributors and Users
G. Availability
H. Confidentiality
I. Maintenance and Operations
J. Standards, Regulatory, and Governance Concerns

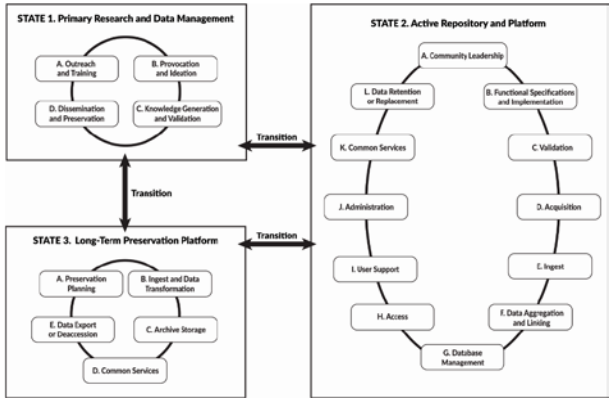
Decision points related to each cost driver are identified

Category	Cost Driver	Decision Points/Issues
A. Content		
A.1	Size (volume and number of items)	1. How many files will be in a single data submission? . . . 5. In what kind of medium will data be captured in the short and long terms?
A.2	Complexity and Diversity of Data	1. How complex is the underlying structure of the data? . . . 6. What are the relationships among these data types (e.g., are the data correlated)?
A.3	Metadata Requirements	1. How much metadata must be stored with each data object to make them FAIR? . . . 5. How much metadata can be extracted computationally?
A.4	Depth Versus Breadth	Will the repository be restricted to certain data classes or types that the repository must support?
A.5	Processing Level and Fidelity	1. Do the raw data need to be stored? . . .

(Excerpt Appendix E)

Steps to Cost Forecasting

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 goals/objectives for resource
- Consider present and future use of resource
- Identify available guidance (e.g., RFAs; community standards; institutional requirements)
- Compare guidance with activities defined for each data state; decide which state(s) best align(s)

Steps to Cost Forecasting (cont)

2. Identify the characteristics of the data, the data contributors, and users.

(Described in Chapter 4)

- Fill in the cost driver template (Appendix E)
  - identify size, complexity, metadata requirements, the depth versus breadth, the processing levels and fidelity, and data replaceability (category A of template)
  - identify life cycle issues (category E)
  - identify data contributors and users (category F)

Steps to Cost Forecasting (cont)

3. **Identify the current and potential value of the data and how the data value might be maintained or increased with time.**

- Consult host institution, funders, and research community to develop metrics to assess data value
- Identify decisions that affect data value in the shorter and longer terms
- Consider how data generation methods affect short- and long-term data value

Steps to Cost Forecasting (cont)

4. Identify the personnel and infrastructure likely necessary in the short and long terms.

- Identify activities and subactivities associated with the information resource, including those to transition between data states (Tables 2.1, 2.2, and 2.3)
- Identify short- and long-term staffing requirements for the current state and for transition between states
- Identify infrastructure requirements and available resources

Steps to Cost Forecasting (cont)

5. Identify the major cost drivers associated with each activity, including how decisions might affect future data use and its cost

- Identify major cost drivers and uncertainties for each activity (answer questions in cost driver template in Appendix E)
- Identify relative costs (Table 4.2)
- Determine available personnel and infrastructure resources(consult host institution, library resources)
- Quantify short-term costs and bound uncertainties in longer-term forecasts (consult experts at your institution)

(Table 4.2)

Correspond with State 1 activities (Table 2.1)

Correspond with State 2 activities (Table 2.2)

Correspond with State 3 activities (Table 2.3)

Rows correspond with cost drivers listed in Ch4 and Appendix E

	State 1: Primary Research and Data Management Environment				State 2: Active Repository and Platform											State 3: Long-Term Preservation Platform				
	I.A Outreach & Training	I.B Provocation & Ideation	I.C Knowledge Generation & Validation	I.D Dissemination & Preservation	II.A Community Leadership	II.B Functional Specifications & Implementation	II.C Validation	II.D Acquisition	II.E Ingest	II.F Data Aggregation & Linking	II.G Database Management	II.H Access	II.I User Support	II.J Administration	II.K Common Services	II.L Data Retention or Replacement	III.A Preservation Planning	III.B Ingest and Data Transformation	III.C Archive Storage	III.D Common Services	III.E Data Export or Deaccession
A. Content																				
A.1. Size			✓	✓		✓			✓	✓	✓	✓			✓	✓		✓	✓		
A.2 Complexity and Diversity of Data Types			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓		✓			
A.3 Metadata Requirements		✓	✓	✓	✓	✓	✓		✓	✓							✓	✓			
A.4 Depth Versus Breadth			✓	✓		✓												✓			
A.5 Processing Level and Fidelity			✓			✓				✓						✓					
A.6 Replaceability of Data			✓					✓								✓					
B. Capabilities																				
B.1 User Annotation						✓				✓			✓			✓					
B.2 Persistent Identifiers			✓	✓		✓			✓	✓						✓		✓			
B.3 Citation		✓				✓			✓		✓					✓					
B.4 Search Capabilities		✓				✓				✓		✓									
B.5 Data Linking and Merging						✓				✓						✓					
B.6 Use Tracking						✓			✓	✓		✓		✓					✓		
B.7 Data Analysis and Visualization			✓			✓						✓									
C. Control																				
C.1 Content Control					✓	✓		✓									✓				
C.2 Quality Control			✓	✓		✓	✓		✓	✓								✓			

Checked items in table indicate the major cost drivers (rows) likely important for a given activity (columns)

	State 2: Active Repository and Platform											
	II.A Community Leadership	II.B Functional Specifications & Implementation	II.C Validation	II.D Acquisition	II.E Ingest	II.F Data Aggregation & Linking	II.G Database Management	II.H Access	II.I User Support	II.J Administration	II.K Common Services	II.L Data Retention or Replacement
A. Content												
A.1. Size		✓			✓	✓	✓	✓			✓	✓
A.2 Complexity and Diversity of Data Types	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓
A.3 Metadata Requirements	✓	✓	✓		✓	✓						
A.4 Depth Versus Breadth		✓										
A.5 Processing Level and Fidelity		✓				✓						✓
A.6 Replaceability of Data				✓								✓
B. Capabilities												
B.1 User Annotation		✓				✓			✓			✓
B.2 Persistent Identifiers		✓			✓	✓						✓
B.3 Citation		✓			✓		✓					✓
B.4 Search Capabilities		✓				✓		✓				
B.5 Data Linking and Merging		✓				✓						✓
B.6 Use Tracking		✓			✓	✓		✓		✓		
B.7 Data Analysis and Visualization		✓						✓				
C. Control												
C.1 Content Control	✓	✓		✓								
C.2 Quality Control		✓	✓		✓	✓						
C.3 Access Control		✓				✓		✓				
C.4 Platform Control		✓									✓	
D. External Context												

Steps to Cost Forecasting (cont)

6. Estimate the costs for relevant cost components based on the characteristics of the data and information resource

- Identify cost drivers important for each cost element (Box 3.2—next slide)
- Estimate costs for the current funding period
- Estimate costs and cost uncertainties for future funding periods, including costs to transition data to other states

Cost Components of a Biomedical Information Resource

- *Labor*—direct salaries and benefits
- *IT infrastructure*—computer purchase, upgrade, and replacement; storage servers; networking equipment; software
- *IT services*—installation, operation, and maintenance of IT infrastructure
- *Media*—consumable storage (e.g., tapes, DVDs)
- *Licenses and subscriptions*—periodic payments for access/use of data, software, services
- *Facilities and utilities*—space for people and IT infrastructure, utilities (might be incorporated into institutional overhead)
- *Outside services*—consultants, external auditors, off-site media storage, training
- *Travel*—costs for outreach activities, to convene governing boards, and so on.
- *Institutional overhead*—indirect costs for administrative and other support (might be allowed in a contract or grant)
- Other "soft" costs (e.g., time users expend to use the data)

(Box 3.2 in text)

Framework Use Cases

- Estimating costs of a new data repository for the BRAIN Initiative

(Chapter 5)

- Estimating costs of new primary research data set

(Chapter 6)

Use Case 1

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
A. Content			
A.1	**Size (volume and number of items)** > size = higher costs	**1. How many files will be in a single data submission?** *Varies, likely from 10 to 10,000.* **2. How large is an average data submission in total?** *Multiple TB.* **3. Are the data sizes likely to stay stable over the life of the resource?** *No, file sizes will likely increase as technologies are developed.* **4. What is the total amount of data expected?** *PBs.* **5. In what kind of medium will data be captured in the short and long terms?** *Data upload into the cloud for short and long term will be captured.*	H
A.2	**Complexity and Diversity of Data** > complexity + diversity = higher cost	**1. How complex is the underlying structure of the data?** *Complex-image data.* **2. How are the included data to be organized?** *To be determined after interviewing funded investigators. Likely individual data sets that include the raw and processed data, but need to determine whether the data should be organized according to studies or projects.* **3. How complex is the experimental paradigm that produced the data?** *Varies—some simple acquisitions; some associated with complex behavioral paradigms.* **4. What sort of additional files might be necessary to upload with the data to properly understand them?** *Experimental protocols, fiducial maps.* **5. How many different data types are being produced?** *Multiple types of imaging data (multimodal data—light and electron microscopy, multiple microscopy types within each; correlated physiology and genomics.* **6. What are the relationships among these data types (e.g., are the data correlated)?** *Some correlated data sets; related data sets will be deposited in the appropriate repository.*	H

Use Case 2

Category	Cost Driver	Decision Points/Issues	Relative Cost Potential (Low, Medium, High)
A. Content			
A.1	**Size (volume and number of items)** > size = higher costs	**1. What is the order of magnitude of data that will be produced?** *Gigabytes (GB).* **2. How large is an average data set?** *Per subject ~ 10 GB (multiple scans over time).* **3. Are the data sizes likely to stay stable over the life of the resource?** *Yes.* **4. What is the total amount of data expected?** *~400 GB.* **5. How many individual files in a typical data set?** *Hundreds.* **6. If the data are to be transferred to a repository for long-term management, is there a cost depending on size?** *No. Data will be submitted to OpenNeuro, which currently does not have costs associated with these data.* **7. Are there publicly available data that can be used to augment these data or perform preliminary analyses?** *No relevant data were found.*	L-M
A.2	**Complexity and Diversity of Data** > complexity + diversity = higher cost	**1. How complex is the underlying structure of the data?** *Complex-image data* **2. How complex is the experimental paradigm that produced them?** *Standard fMRI block design.* **3. What sort of additional data are acquired along with the primary data?** *Cognitive assessments, statistical maps, demographic data.* **4. How many different data types are being produced?** *Multiple modalities.* **6. What are the relationships among these data types—for example, are they correlated data?** *Not applicable.*	M

Fostering the Data Management Environment

In lieu of recommendations, the committee suggested

- Strategies (data resource managers, cost forecasters, institutions that support them)
- Actions (data institutions and funding agencies)
- Advances for Practice (agencies and research institutions)

Strategies

Incorporate data management activities throughout the data life cycle to strengthen data curation and preservation.

Up-front costs may increase, but data value may also increase, and overall cost of research may be reduced.

Actions

Identify incentives, tools, and training for adopting good data management practices, including cost-forecasting practices, which facilitate sustainable long-term data preservation, curation, and access

Funding entities need to better understand research-community needs, help the community define desired outcomes, develop metrics and incentives for success.

Actions (cont)

Support

- *standardization efforts (including tools/methodologies to estimate the cost of standards development)*
- *encouraging the use of those tools as part of funding programs*
- *Explicit support of metadata preparation*

Advances for Practice

Recognize explicitly that scientific data constitute an asset and that data stewardship requires support.

Institutions that support or enable research and host data resources benefit from recognition of support.

Advances for Practice (cont)

Systematically collect data on costs associated with the biomedical research data enterprise

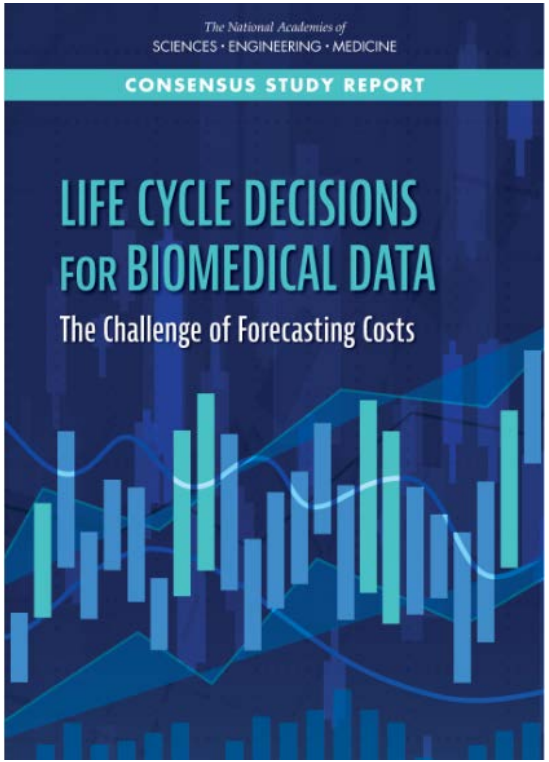
to allow translation of framework into resources and methodologies that benefit researchers and repository institutions.

A clear locus of responsibility for compiling information systematically needed.

Advances for Practice (cont)

Develop easier mechanisms for creating and maintaining data management plans, automatically incorporating data and metadata into resources, and improving citations for data to work together with other research products

BOARD ON MATHEMATICAL SCIENCES AND ANALYTICS



Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs

Report Organization

1. Introduction
2. Framework Foundation: Data States and Associated Activities
3. Cost and Value of Data
4. Cost Forecasting Framework: Identifying Cost Drivers in the Data Life Cycle
5. Applying the Framework to a New State 2 Data Resource
6. Applying the Framework to a New Data Set
7. Potential Disruptors to Forecasting Costs
8. Strategies, actions, and advances to foster the data management environment