The National Academies of SCIENCES ENGINEERING MEDICINE

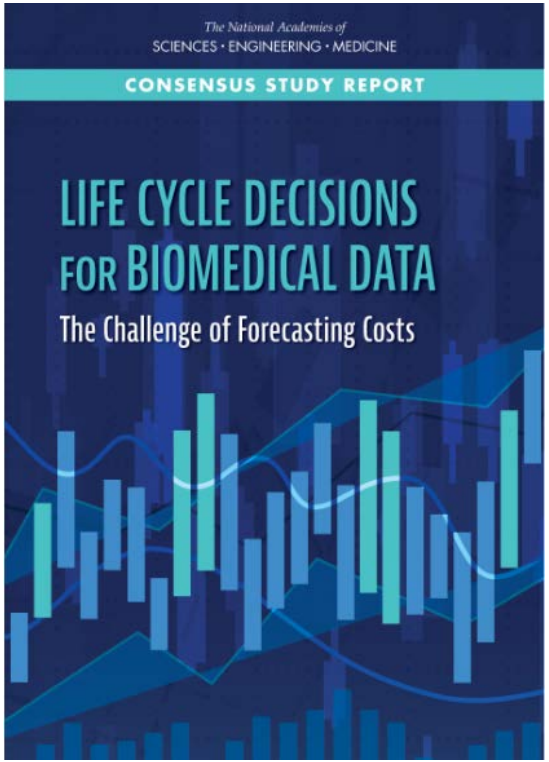
Board on
Mathematical Sciences & Analytics

LIFE CYCLE DECISIONS FOR BIOMEDICAL DATA

The Challenge of Forecasting Costs

The National Academies of SCIENCES ENGINEERING MEDICINE

BOARD ON MATHEMATICAL SCIENCES AND ANALYTICS



Forecasting Data Costs for Funders

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

Presented to the Public
August 20, 2020

Forecasting Data Costs for Researchers, Funders, and Storage Providers

August 2020 weekly webinar series, 12-1pm ET

August 13: *Forecasting Data Costs for Researchers*

Recording available at https://vimeo.com/showcase/7444639

August 20: *Forecasting Data Costs for Funding Institutions*

August 27: *Forecasting Data Costs for Storage Providers*

This webinar series is sponsored by the
National Library of Medicine of the National Institutes of Health

Watch webinar videos and learn more about BMSA at https://biomed-data-costs.eventbrite.com/

Forecasting Data Costs
for Funding Organizations

Margaret Levenstein (Moderator)
University of Michigan

Alexa McCray
Harvard University

Amy Friedlander
National Science Foundation

Michael Lauer
National Institutes of Health

Forecasting Data Costs
for Funding Organizations

Alexa McCray

Professor of Medicine,
Harvard Medical School

Forecasting Data Costs: Highlights for Funders

Context

- Biomedical researchers generate, collect, and store more research data than ever.
- Sustained data access and preservation generate costs that are difficult to predict.
- The current funding system does not accommodate forecasting costs for sustained data access and preservation.

Data Value

- The perceived value of data influences decisions regarding their life cycle.
- Data value does not necessarily correlate with the financial investment made to collect those data.
- When new data are integrated with existing data, the value of the entire pool is increased.

Statement of Task

The 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.

Framework Foundation: Three Data States

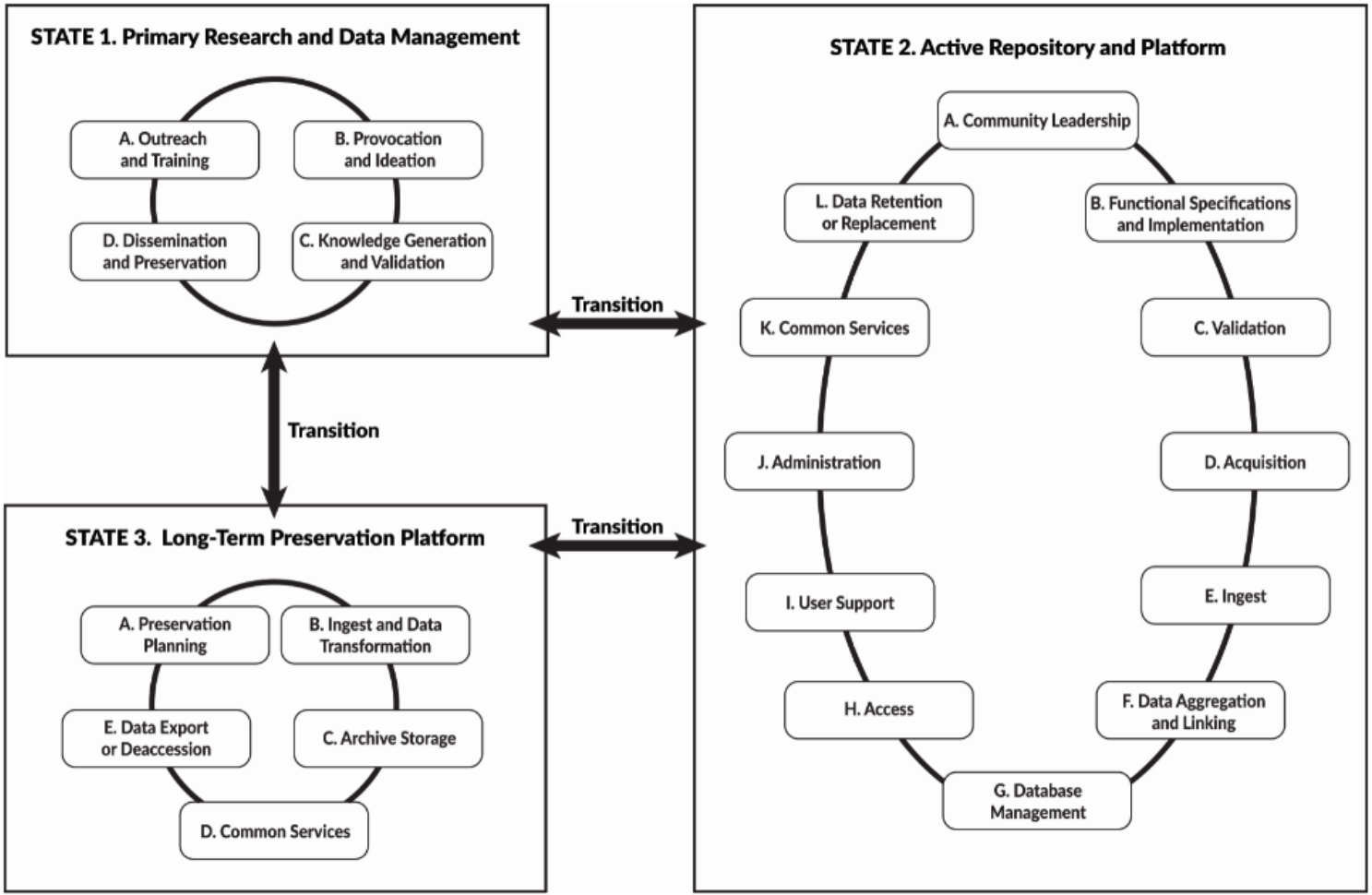
State 1: **Primary research**/data management environment; data are captured and analyzed

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)



Framework Foundation: Data States and Activities

Cost Forecasting Framework: Cost Drivers

Data properties that affect the costs of data access and preservation

- Content
- Capabilities
- Control
- External Context
- Data Life Cycle
- Contributors and Users
- Availability
- Confidentiality
- Maintenance and Operations
- Standards, Regulatory, and Governance concerns

Appendix E contains a template that includes questions to help the cost forecaster identify key decision points for each cost driver.

Cost Forecasting Framework: Cost Components

- *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)

Cost Forecasting Framework: Considerations

- Helps forecaster identify major cost drivers
- Helps 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

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

- Create data environments that foster discoverability and interpretability through long-term *planning* and *investment* throughout the data life cycle.
- 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.

Actions

- Support *standardization* efforts (including *tools/methodologies* to estimate the cost of standards development)
- Provide explicit support for *metadata preparation*
- Structure cost forecasts for State 2 resources around *communities* and *research programs rather than individual research efforts.*
  - State 2 resources serve research communities—it may be inappropriate to allocate costs of managing data back to individual researchers.

Advances for Practice

- Identify and support *incentives, tools, and training* for adopting good data management practices, including cost-forecasting practices.
- Recognize explicitly that *scientific data constitute an asset* and that *data stewardship requires support*.
- *Systematically collect data* on costs associated with the biomedical research data enterprise.
  - A clear *locus of responsibility* for compiling information systematically *is needed*.

Forecasting Data Costs
for Funding Organizations

Amy Friedlander

(Acting) Office Director;
Office of Advanced Cyberinfrastructure,
Directorate for Computer and Information Science and Engineering

Forecasting Data Costs
for Funding Organizations

Michael Lauer

Deputy Director for Extramural Research,
National Institutes of Health

NIH National Institutes of Health
Office of Extramural Research

Forecasting Data Costs for Funders: Webinar on Biomedical Data Preservation

Michael S. Lauer MD
Deputy Director for Extramural Research, National Institutes of Health (NIH)

Thursday, August 20, 2020
NASEM Virtual Meeting

Longstanding Commitment

Examples

NIH Data Sharing Policy (2003)

Establishes expectation that research data from large awards (≥ $500K) will be shared

NIH Public Access Policy (2008)

Ensures public access to published results of NIH –funded research

NIH Genome-Wide Association Studies Policy (2008) & NIH Genomic Data Sharing Policy (2015)

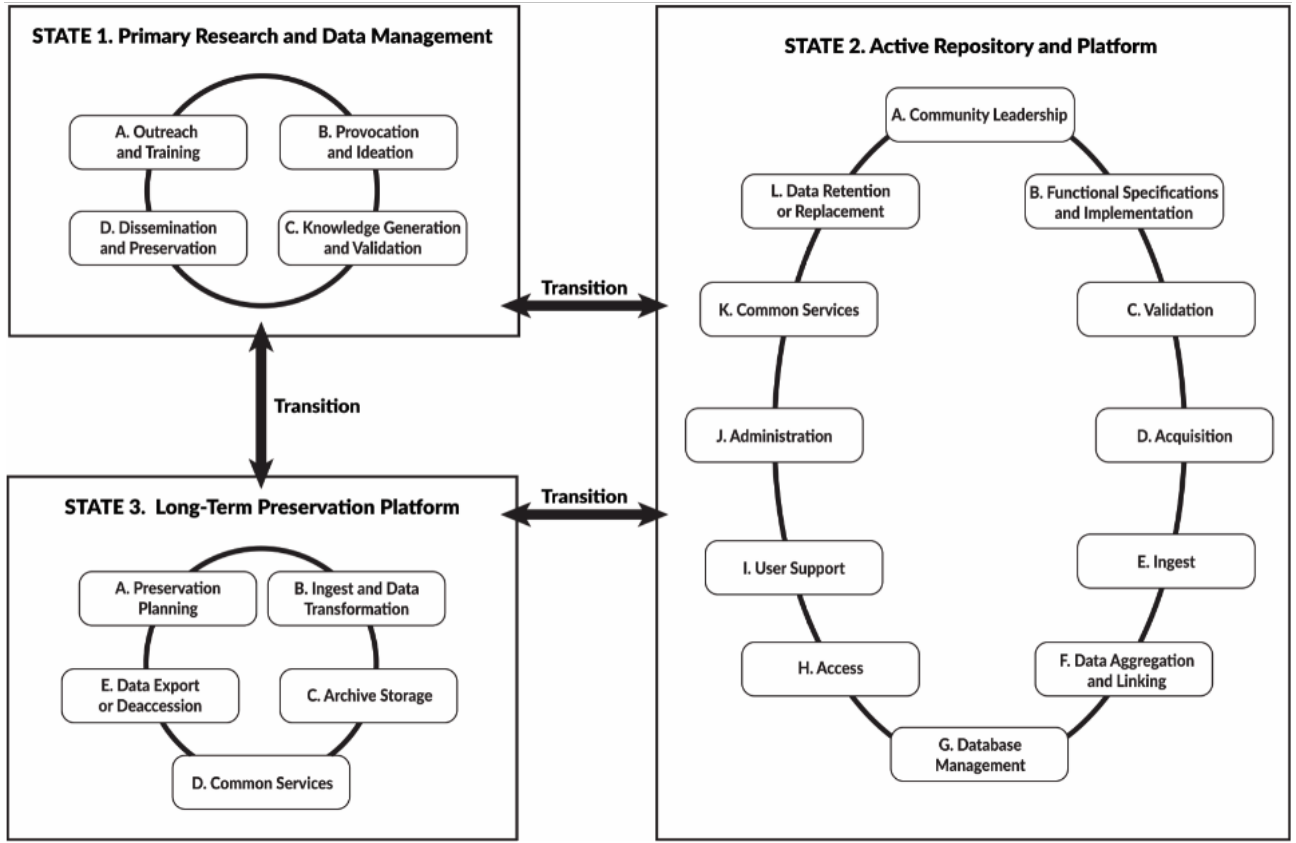
Establishes expectations for sharing large-scale genomic data

NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information (2017)

Establishes expectation for the timely registration and submission of results information for all NIH-funded clinical trials

https://osp.od.nih.gov/wp-content/uploads/Draft_DMS_Policy_Webinar_Dec2019.pdf

Data States



https://www.nap.edu/resource/25639/biomed_public_webinar_0717.pdf

Does Data Preservation Work?

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

Use of the National Heart, Lung, and Blood Institute Data Repository

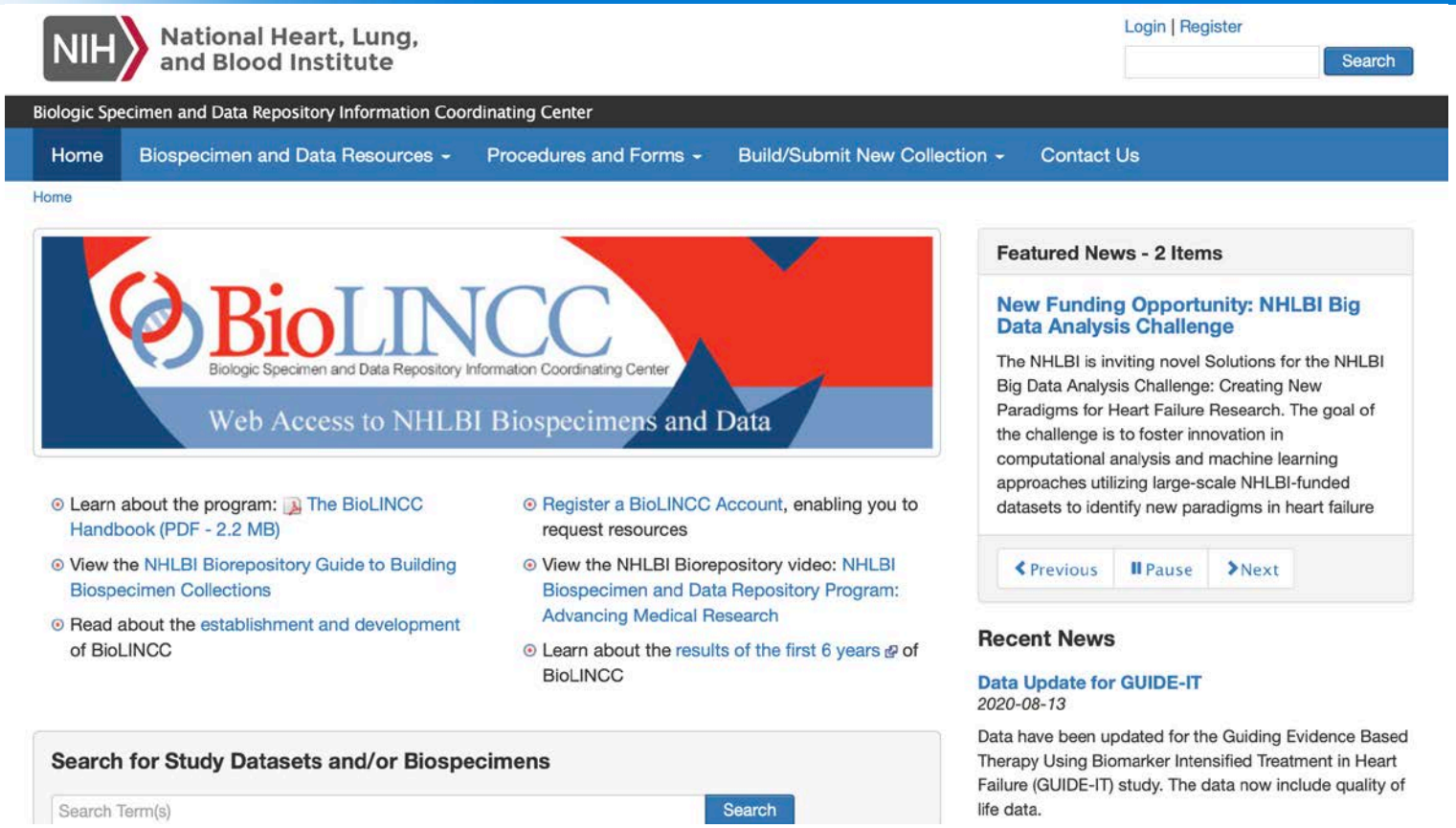
Sean A. Coady, M.S., M.A., George A. Mensah, M.D., Elizabeth L. Wagner, M.P.H., Miriam E. Goldfarb, B.S., A.S.N., Denise M. Hitchcock, B.S., and Carol A. Giffen, D.V.M.

This article was published on March 29, 2017, at NEJM.org.

DOI: 10.1056/NEJMsa1603542

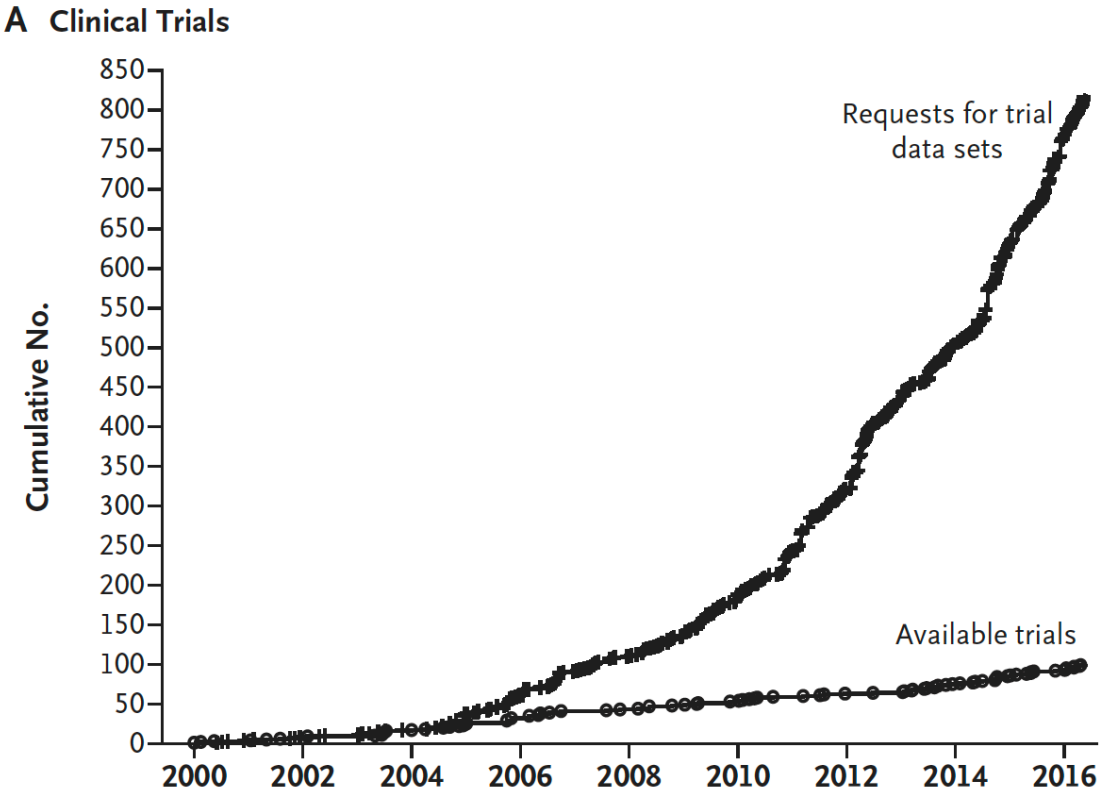
ABSTRACT

Data Repository



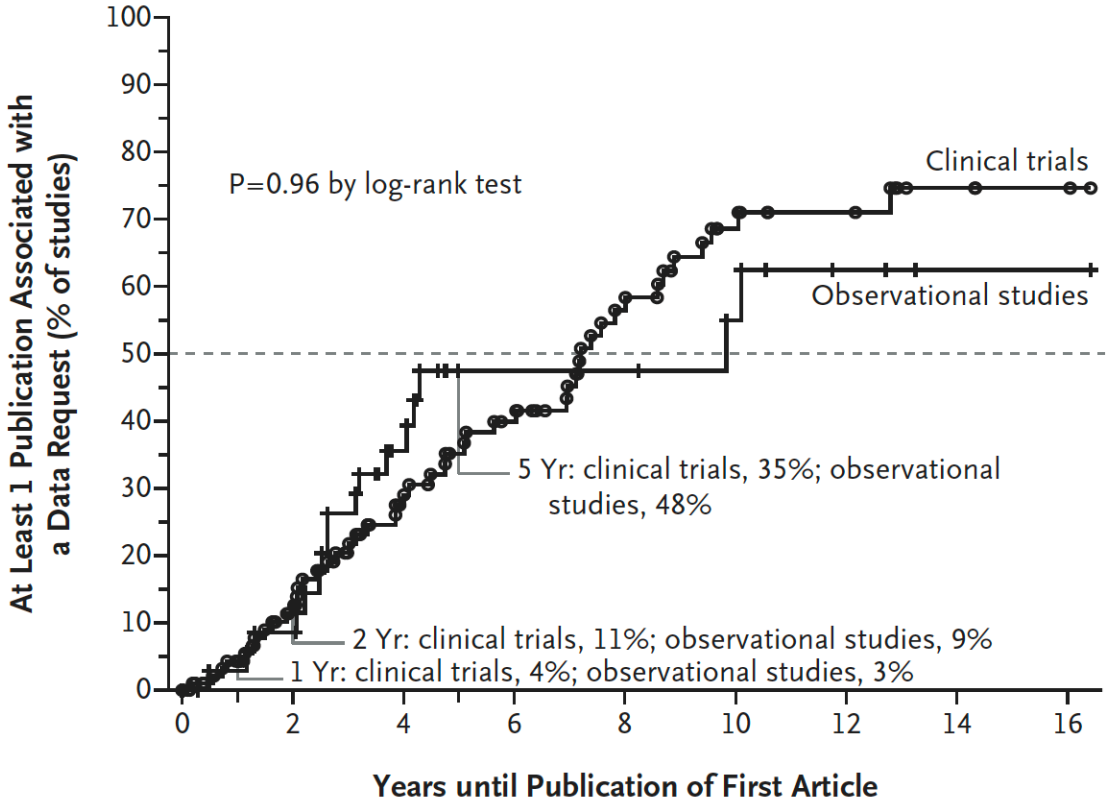
https://biolincc.nhlbi.nih.gov/home/

Data Are Available and Requested



https://www.nejm.org/doi/full/10.1056/NEJMsa1603542

Data Requests Lead to Publications



https://www.nejm.org/doi/full/10.1056/NEJMsa1603542

Authors' Impressions

"The … release of clinical trial data for wide sharing can contribute to the scientific community in multiple ways, including increasing the transparency of findings, examining new hypothesis-generating questions, providing pilot data for grant submissions, testing statistical methods, performing meta-analyses, and developing prediction algorithms."

https://www.nejm.org/doi/full/10.1056/NEJMsa1603542

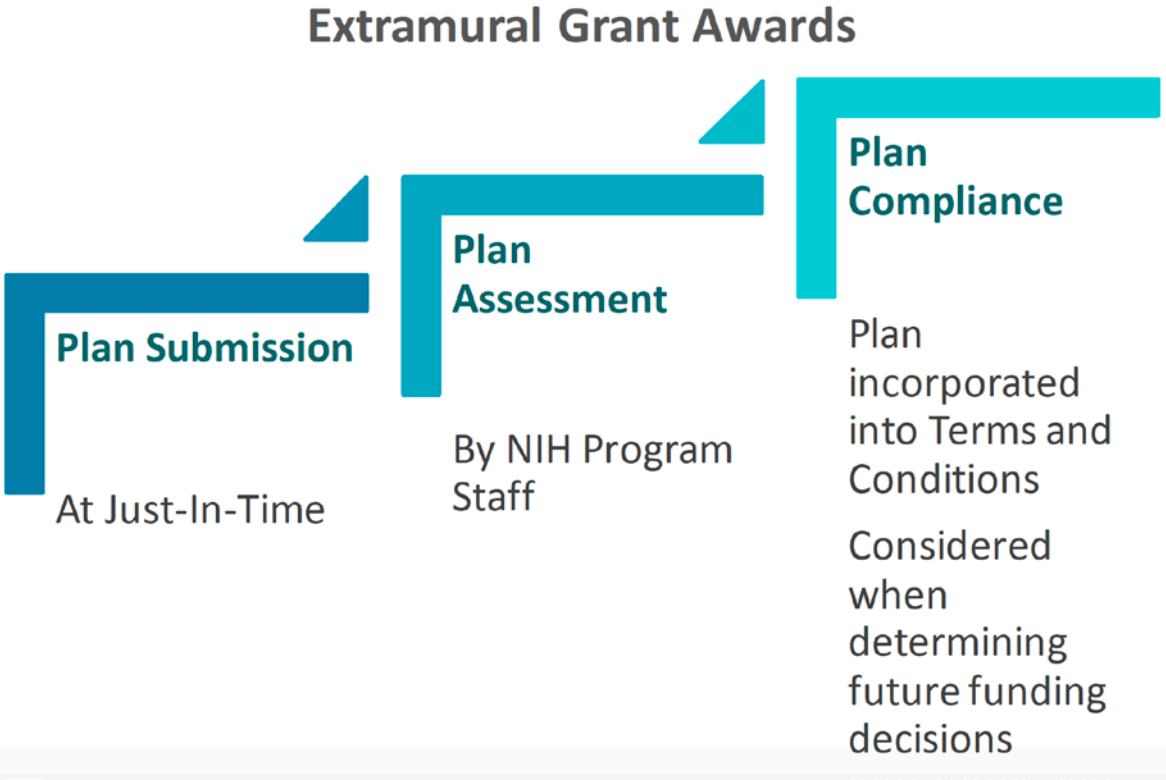
Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental DRAFT Guidance

Notice Number: NOT-OD-20-013

Key Dates

Release Date: November 06, 2019
Response Date: January 10, 2020

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-013.html



https://osp.od.nih.gov/wp-content/uploads/Draft_DMS_Policy_Webinar_Dec2019.pdf

Closing Thoughts

- Longstanding commitment
- Costs are real
- NASEM document: helpful, multi-stakeholder framework
- Value is clear and itself data-driven

Forecasting Data Costs
for Funding Organizations

Please submit questions using the Q&A button in the zoom menu.

Margaret Levenstein (Moderator)
University of Michigan

Alexa McCray
Harvard University

Amy Friedlander
National Science Foundation

Michael Lauer
National Institutes of Health

Forecasting Data Costs for Researchers, Funders, and Storage Providers

August 2020 weekly webinar series, 12-1pm ET

August 13: *Forecasting Data Costs for Researchers*

Recording available at https://vimeo.com/showcase/7444639

August 20: *Forecasting Data Costs for Funding Institutions*

August 27: *Forecasting Data Costs for Storage Providers*

This webinar series is sponsored by the
National Library of Medicine of the National Institutes of Health

Watch webinar videos and learn more about BMSA at https://biomed-data-costs.eventbrite.com/