National Institutes of Health

Data and Informatics Working Group

Draft Report to
The Advisory Committee to the Director

June 15, 2012

Working Group Members

David DeMets, Ph.D., Professor, Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison; co-chair

Lawrence Tabak, D.D.S., Ph.D., Principal Deputy Director, National Institutes of Health; co-chair

Russ Altman, M.D., Ph.D., Professor and Chair, Department of Bioengineering, Stanford University

David Botstein, Ph.D., Director, Lewis-Sigler Institute, Princeton University

Andrea Califano, Ph.D., Professor and Chief, Division of Biomedical Informatics, Columbia University

David Ginsburg, M.D., Professor, Department of Internal Medicine, University of Michigan; Howard Hughes Medical Institute; Chair, National Center for Biotechnology Information (NCBI) Needs-Assessment Panel

Patricia Hurn, Ph.D., Associate Vice Chancellor for Health Science Research, The University of Texas System

Daniel Masys, M.D., Affiliate Professor, Department of Biomedical Informatics and Medical Education, University of Washington

Jill P. Mesirov, Ph.D., Associate Director and Chief Informatics Officer, Broad Institute; Ad Hoc Member, NCBI Needs-Assessment Panel

Shawn Murphy, M.D., Ph.D., Associate Director, Laboratory of Computer Science, and Associate Professor, Department of Neurology, Harvard University

Lucila Ohno-Machado, M.D., Ph.D., Associate Dean for Informatics, Professor of Medicine, and Chief, Division of Biomedical Informatics, University of California, San Diego

Ad-hoc Members

David Avrin, M.D., Ph.D., Professor and Vice Chairman, Department of Radiology, University of California at San Francisco

Paul Chang, M.D., Professor and Vice-Chairman, Department of Radiology, University of Chicago

Christopher Chute, M.D., Dr.P.H, Professor, Department of Health Sciences Research, Mayo Clinic College of Medicine

Ted Hanss, M.B.A., Chief Information Officer, University of Michigan Medical School

Paul Harris, Ph.D., Director, Office of Research Informatics, Vanderbilt University

Marc Overcash, Deputy Chief Information Officer, Emory University School of Medicine

James Thrall, M.D., Radiologist-in-Chief and Professor of Radiology, Massachusetts General Hospital, Harvard Medical School

A. Jerome York, M.B.A., Vice President and Chief Information Officer, The University of Texas Health Science Center at San Antonio

Acknowledgements

We are most grateful to the members of the Data and Informatics Working Group for their considerable efforts. We acknowledge David Bluemke, Jim Cimino, John Gallin, John J. McGowan, Jon McKeeby, Andrea Norris, and George Santangelo for providing background information and expertise on the National Institutes of Health (NIH) for the Working Group members. Great appreciation is extended to members of the NIH Office of Extramural Research team that gathered the training data that appear in this draft report and the trans-NIH BioMedical Informatics Coordinating Committee for their additional contributions to this data. We also thank members of the Biomedical Information Science and Technology Initiative project team, external review panel, and community for their permission to reference and publish the National Centers for Biomedical Computing mid-course review report. Input from a number of Institute and Center Directors not directly involved with the project is gratefully acknowledged.

Finally, we acknowledge with our deepest thanks the truly outstanding efforts of our team: Jennifer Weisman, Steve Thornton, Kevin Wright, and Justin Hentges.

Dr. David DeMets, Co-Chair, Data and Informatics Working Group of the Advisory Committee to the NIH Director

Dr. Lawrence Tabak, Co-Chair, Data and Informatics Working Group of the Advisory Committee to the NIH Director

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

1.1 Committee Charge and Approach

In response to the accelerating growth of biomedical research datasets, the Director of the National Institutes of Health (NIH) charged the Advisory Committee to the Director (ACD) to form a special Data and Informatics Working Group (DIWG). The DIWG was asked to provide the ACD and the NIH Director with expert advice on the management, integration, and analysis of large biomedical research datasets. The DIWG was charged to address the following areas:

- research data spanning basic science through clinical and population research
- administrative data related to grant applications, reviews, and management
- management of information technology (IT) at the NIH

The DIWG met nine times in 2011 and 2012, including two in-person meetings and seven teleconferences, toward the goal of providing a set of consensus recommendations to the ACD at its June 2012 meeting. In addition, the DIWG published a Request for Information (RFI) as part of their deliberations (see Appendix, Section 6.1 for a summary and analysis of the input received).

The overall goals of the DIWG's work are at once simple and compelling:

- to advance basic and translational science by facilitating and enhancing the sharing of research-generated data
- to promote the development of new analytical methods and software for this emerging data
- to increase the workforce in quantitative science toward maximizing the return on the NIH's public investment in biomedical research

The DIWG believes that achieving these goals in an era of "Big Data" requires innovations in technical infrastructure and policy. Thus, its deliberations and recommendations address technology and policy as complementary areas in which NIH initiatives can catalyze research productivity on a national, if not global, scale.

1.2 DIWG Vision Statement

Research in the life sciences has undergone a dramatic transformation in the past two decades. Colossal changes in biomedical research technologies and methods have shifted the bottleneck in scientific productivity from data production to data management, communication, and interpretation. Given the current and emerging needs of the biomedical research community, the NIH has a number of key opportunities to encourage and better support a research ecosystem that leverages data and tools, and to strengthen the workforce of people doing this research. The need for advances in cultivating this ecosystem is particularly evident considering the current and growing deluge of data originating from next-generation sequencing, molecular profiling, imaging, and quantitative phenotyping efforts.

The DIWG recommends that the NIH should invest in technology and tools needed to enable researchers to easily find, access, analyze, and curate research data. NIH funding for methods and equipment to adequately represent, store, analyze, and disseminate data throughout their useful lifespan should be coupled to NIH funding toward generating those original data. The NIH should also increase the capacity of the workforce (both for experts and non-experts in the quantitative disciplines), and employ strategic planning to leverage IT advances for the entire NIH community. The NIH should continue to develop a collaborative network of centers to implement this expanded vision of sharing data and developing and disseminating methods and tools. These centers will provide a means to make these resources available to the biomedical research community and to the general public, and will provide training on and support of the tools and their proper use.

1.3 Overview of Recommendations

A brief description of the DIWG's recommendations appears below. More detail can be found in Sections 2-4.

Recommendation 1: Promote Data Sharing Through Central and Federated Catalogues

Recommendation 1a. Establish a Minimal Metadata Framework for Data Sharing

The NIH should establish a truly minimal set of relevant data descriptions, or metadata, for biomedically relevant types of data. Doing so will facilitate data sharing among NIH-funded researchers. This resource will allow broad adoption of standards for data dissemination and retrieval. The NIH should convene a workshop of experts from the user community to provide advice on creating a metadata framework.

Recommendation 1b. Create Catalogues and Tools to Facilitate Data Sharing

The NIH should create and maintain a centralized catalogue for data sharing. The catalogue should include data appendices to facilitate searches, be linked to the published literature from NIH-funded research, and include the associated minimal metadata as defined in the metadata framework to be established (described above).

Recommendation 1c. Enhance and Incentivize a Data Sharing Policy for NIH-Funded Data

The NIH should update its 2003 data sharing policy to require additional data accessibility requirements. The NIH should also incentivize data sharing by making available the number of accesses or downloads of datasets shared through the centralized resource to be established (described above). Finally, the NIH should create and provide model data-use agreements to facilitate appropriate data sharing.

Recommendation 2: Support the Development, Implementation, Evaluation, Maintenance, and Dissemination of Informatics Methods and Applications

Recommendation 2a. Fund All Phases of Scientific Software Development via Appropriate Mechanisms

The development and distribution of analytical methods and software tools valuable to the research community occurs through a series of stages: prototyping, engineering/hardening, dissemination, and maintenance/support. The NIH should devote resources to target funding for each of these four stages.

Recommendation 2b. Assess How to Leverage the Lessons Learned from the National Centers for Biomedical Computing

The National Centers for Biomedical Computing (NCBCs) have been an engine of valuable collaboration between researchers conducting experimental and computational science, and each center has typically prompted dozens of additional funded efforts. The NIH should consider the natural evolution of the NCBCs into a more focused activity.

Recommendation 3: Build Capacity by Training the Workforce in the Relevant Quantitative Sciences such as Bioinformatics, Biomathematics, Biostatistics, and Clinical Informatics

Recommendation 3a. Increase Funding for Quantitative Training and Fellowship Awards

NIH-funded training of computational and quantitative experts should grow to help meet the increasing demand for professionals in this field. To determine the appropriate level of funding increase, the NIH should perform a supply-and-demand analysis of the population of computational and quantitative experts, as well as develop a strategy to target and reduce identified gaps. The NCBCs should also continue to play an important educational role toward informing and fulfilling this endeavor.

Recommendation 3b. Enhance Review of Quantitative Training Applications

The NIH should investigate options to enhance the review of specialized quantitative training grants that are typically not reviewed by those with the most relevant experience in this field. Potential approaches include the formation of a dedicated study section for the review of training grants for quantitative science (*e.g.,* bioinformatics, clinical informatics, biostatistics, and statistical genetics).

Recommendation 3c. Create a Required Quantitative Component for All NIH Training and Fellowship Awards

The NIH should include a required computational or quantitative component in all training and fellowship grants. This action would contribute to substantiating a workforce of clinical and biological scientists trained to have some basic proficiency in the understanding and use of quantitative tools in order to fully harness the power of the data they generate. The NIH should draw on the experience and expertise of the Clinical and Translational Science Awards (CTSAs) in developing the curricula for this core competency.

Recommendation 4: Develop an NIH-Wide "On-Campus" IT Strategic Plan

Recommendation 4a. For NIH Administrative Data:

The NIH should update its inventory of existing analytic and reporting tools and make this resource more widely available. The NIH should also enhance the sharing and coordination of resources and tools to benefit all NIH staff as well as the extramural community.

Recommendation 4b. For the NIH Clinical Center:

The NIH Clinical Center (CC) should enhance the coordination of common services that span the Institutes and Centers (ICs), to reduce redundancy and promote efficiency. In addition, the CC should create an informatics laboratory devoted to the development of implementation of new solutions and strategies to address its unique concerns. Finally, the CC should strengthen relationships with other NIH translational activities including the National Center for Advancing Translational Sciences (NCATS) and the CTSA centers.

Recommendation 4c. For the NIH IT and Informatics Environment:

The NIH should employ a strategic planning process for trans-agency IT design that includes considerations of the management of Big Data and strategies to implement models for high-value IT initiatives. The first step in this process should be an NIH-wide IT assessment of current services and capabilities. Next, the NIH should continue to refine and expand IT governance. Finally, the NIH should recruit a Chief Science Information Officer (CSIO) and establish an external advisory group to serve the needs of/guide the plans and actions of the NIH Chief Information Officer (CIO) and CSIO.

Recommendation 5: Provide a Serious, Substantial, and Sustained Funding Commitment to Enable Recommendations 1-4

The current level of NIH funding for IT-related methodology and training has not kept pace with the ever-accelerating demands and challenges of the Big Data environment. The NIH must provide a serious, substantial, and sustained increase in funding IT efforts in order to enable the implementation of the DIWG's recommendations 1-4. Without a systematic and increased investment to advance computation and informatics support at the trans-NIH level and at every IC, the biomedical research community will not be able to make efficient and productive use of the massive amount of data that are currently being generated with NIH funding.

1.4 Report Overview

This report is organized into the following sections following the executive summary to provide a more in-depth view into the background and the DIWG's recommendations:

Section 2 provides a detailed account of the DIWG's recommendations related to research data spanning basic science through clinical and population research, including workforce considerations (Recommendations 1-3).

Section 3 provides a detailed explanation of the DIWG's recommendations concerning NIH "on campus" data and informatics issues, including those relevant to grants administrative data, NIH CC informatics, and the NIH-wide IT and informatics environment (Recommendation 4).

Section 4 provides details about the DIWG's recommendation regarding the need for a funding commitment (Recommendation 5).

Section 5 provides acknowledgements.

Section 6 includes references cited in the report.

Section 7 includes appendices.

2 RESEARCH DATA SPANNING BASIC SCIENCE THROUGH CLINICAL AND POPULATION RESEARCH

2.1 Background

Research in the life sciences has undergone a dramatic transformation in the past two decades. Fueled by high-throughput laboratory technologies for assessing the properties and activities of genes, proteins and other biomolecules, the "omics" era is one in which a single experiment performed in a few hours generates terabytes (trillions of bytes) of data. Moreover, this extensive amount of data requires both quantitative biostatistical analysis and semantic interpretation to fully decipher observed patterns. Translational and clinical research has experienced similar growth in data volume, in which gigabyte-scale digital images are common, and complex phenotypes derived from clinical data involve data extracted from millions of records with billions of observable attributes. The growth of biomedical research data is evident in many ways: in the deposit of molecular data into public databanks such as GenBank (which as of this writing contains more than 140 billion DNA bases from more than 150 million reported sequences[1]), and within the published PubMed literature that comprises over 21 million citations and is growing at a rate of more than 700,000 new publications per year[2].

Significant and influential changes in biomedical research technologies and methods have shifted the bottleneck in scientific productivity from data production to data management, communication — and most importantly — interpretation. The biomedical research community is within a few years of the

[1] ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt
[2] http://www.nlm.nih.gov/pubs/factsheets/medline.html

"thousand-dollar human genome needing a million-dollar interpretation." Thus, the observations of the ACD Working Group on Biomedical Computing as delivered 13 years ago, in their June 1999 report to the ACD on the Biomedical Information Science and Technology Initiative (BISTI)[3] are especially timely and relevant:

> *Increasingly, researchers spend less time in their "wet labs" gathering data and more time on computation. As a consequence, more researchers find themselves working in teams to harness the new technologies. A broad segment of the biomedical research community perceives a shortfall of suitably educated people who are competent to support those teams. The problem is not just a shortage of computationally sophisticated associates, however. What is needed is a higher level of competence in mathematics and computer science among biologists themselves. While that trend will surely come of its own, it is the interest of the NIH to accelerate the process. Digital methodologies — not just digital technology — are the hallmark of tomorrow's biomedicine.*

It is clear that modern interdisciplinary team science requires an infrastructure and a set of policies and incentives to promote data sharing, and it needs an environment that fosters the development, dissemination, and effective use of computational tools for the analysis of datasets whose size and complexity have grown by orders of magnitude in recent years. Achieving a vision of seamless integration of biomedical data and computational tools is made necessarily more complex by the need to address unique requirements of clinical research IT. Confidentiality issues, as well as fundamental differences between basic science and clinical investigation, create real challenges for the successful integration of molecular and clinical datasets. The sections below identify a common set of principles and desirable outcomes that apply to biomedical data of all types, but also include special considerations for specific classes of data that are important to the life sciences and to the NIH mission.

2.2 Findings

The biomedical research community needs increased NIH-wide programmatic support for bioinformatics and computational biology, both in terms of the research itself and in the resulting software. This need is particularly evident considering the growing deluge of data stemming from next-generation sequencing, molecular profiling, imaging, and quantitative phenotyping efforts. Particular attention should be devoted to the support of a data-analysis framework, both with respect to the dissemination of data models that allow effective integration, as well as to the design, implementation, and maintenance of data analysis algorithms and tools.

Currently, data sharing among biomedical researchers is lacking, due to multiple factors. Among these is the fact that there is no technical infrastructure for NIH-funded researchers to easily submit datasets associated with their work, nor is there a simple way to make those datasets available to other researchers. Second, there is little motivation to share data, since the most common current unit of academic credit is co-authorship in the peer-reviewed literature. Moreover, promotion and tenure in academic health centers seldom includes specific recognition of data sharing outside of the construct of co-authorship on scientific publications. The NIH has a unique opportunity — as research sponsor, as steward of the peer-review process for awarding research funding, and as the major public library for access to research results. The elements of this opportunity are outlined below in brief; noting the DIWG's awareness that actual implementation by the NIH may be affected by resource availability and Federal policy.

Google and the National Security Agency process significantly more data every day than does *the entire biomedical research community*.[4] These entities facilitate access to and searchability of vast amounts of

[3] http://www.bisti.nih.gov/library/june_1999_Rpt.asp

[4] In 2011, it was estimated that NSA processed every six hours an amount of data equivalent to all of the knowledge housed at the Library of Congress (Calvert, 2011). In 2012, it was estimated that Google processed about 24PB (petabytes) of data per day (Roe, 2012).

data to non-expert users, by generating applications that create new knowledge from the data with no *a priori* restrictions on its format. These exemplars provide evidence that the Big Data challenge as related to biomedical research can be addressed in a similar fashion, although not at present. The development of minimal standards would reduce dramatically the amount of effort required to successfully complete such a task within the biomedical research universe. In the case of Google, the HTML format represented such a minimal standard[5].

Experience has shown that given easy and unencumbered access to data, biomedical scientists will develop the necessary analytical tools to "clean up" the data and use it for discovery and confirmation. For example, the Nucleic Acids Research database inventory alone comprises more than 1,380 databases in support of molecular biology (Galperin & Fernandez-Suarez, 2012). In other spheres, data organization is based primarily on the creation and search of large data stores. A similar approach may work well for biomedicine, adjusting for the special privacy needs required for human subjects data.

Biomedical datasets are usually structured and in most cases, that structure is not self-documenting. For this reason, a key unmet need for biomedical research data sharing and re-use is the development of a minimal set of metadata (literally, "data about data") that describes the content and structure of a dataset, the conditions under which it was produced, and any other characteristics of the data that need to be understood in order to analyze it or combine it with other related datasets. As described in the DIWG's recommendations, the NIH should create a metadata framework to facilitate data sharing among NIH-funded researchers. NIH should convene a workshop of experts from the user community to provide advice on the creation of the metadata framework.

Toward enhancing the utility and efficiency of biomedical research datasets and IT needs, in general, the NIH must be careful to keep a pragmatic, biomedically motivated perspective. Establishing universal frameworks for data integration and analysis has been attempted in the past with suboptimal results. It is likely that these efforts were not as successful as they could have been because they were based on abstract, theoretical objectives, rather than on tangible, community and biomedical research-driven problems. Specifically, no single solution will support all future investigations: Data should not be integrated for the sake of integration, but rather as a means to ask and answer specific biomedical questions and needs. In addition to the generalizable principles affecting all classes of research data, there are special considerations for the acquisition, management, communication and analysis of specific types, as enumerated below.

Special Considerations for Molecular Profiling Data

The increasing need to connect genotype and phenotype findings — as well as the increasing pace of data production from molecular and clinical sources (including images) — have exposed important gaps in the way the scientific community has been approaching the problem of data harmonization, integration, analysis, and dissemination.

Tens of thousands of subjects may be required to obtain reliable evidence relating disease and outcome phenotypes to the weak and rare effects typically reported from genetic variants. The costs of assembling, phenotyping, and studying these large populations are substantial — recently estimated at $3 billion for the analyses from 500,000 individuals. Automation in phenotypic data collection and presentation, especially from the clinical environments from which these data are commonly collected, could facilitate the use of electronic health record data from hundreds of millions of patients (Kohane, 2011).

The most explosive growth in molecular data is currently being driven by high-throughput, next-generation, or "NextGen," DNA-sequencing technologies. These laboratory methods and associated instrumentation generate "raw sequence reads" comprising terabytes of data, which are then reduced to consensus DNA-sequence outputs representing complete genomes of model organisms and humans.

[5] The current HTML standard can be found at w3c.org (World Wide Web Consortium (W3C), 2002).

Moreover, as technology improves and costs decline, more types of data (e.g., expression and epigenetic) are being acquired via sequencing. The gigabyte-scale datasets that result from these technologies overwhelm the communications bandwidth of the current global Internet, and as a result the most common data transport from sequencing centers to users is via a hard drive or other computer media sent in a physical package.

Compressing such data efficiently and in a lossless fashion could be achieved, considering the fundamental observation that within and across eukaryotic species, genomes are much more alike than they are different. That there are 1 to 4 million differences between individuals with a 3-billion nucleotide genome can alternatively be stated that 99 percent of the data is identical and thus unnecessary to transmit repetitively in the process of sharing data. This evolutionary reality presents an opportunity for the NIH to sponsor approaches toward developing reference standard genomes. Such genomic tools are in essence a set of ordered characters that can be used as a common infrastructure for digital subtraction. The process can be likened to “dehydrating” genomic and other common molecular sequence data for the purpose of communication across bandwidth-limited infrastructures such as the open Internet, then “rehydrated” by the end user without loss of fidelity to the original observations (Masys, et al., 2012).

Special Considerations for Phenotypic Data

Phenotype (from Greek *phainein*, “to show” plus *typos*, “type”) can be defined as “the composite of an organism's observable characteristics or traits” according to Wikipedia. Although the term was originally linked to the concept of genotype, it is now often more loosely construed as groupings of observations and measurements that define a biological state of interest. In the realm of NIH-funded research, phenotypes may be defined for objects ranging from unicellular organisms to humans, and they may include components from almost any subdomain of the life sciences, including chemistry, molecular and cell biology, tissues, organ systems, as well as human clinical data such as signs, symptoms, and laboratory measurements. Unlike specific data types familiar in computer science (such as text, integers, binary large objects), phenotypes are less well-defined and usually composed of a variable number of elements closely aligned to a particular research projects aims (*e.g.*, the phenotype of a person with type II diabetes who has received a specific drug and experienced a particular adverse event).

For purposes of this report, phenotypic data are categorized as either sensitive or nonsensitive. Sensitive phenotype data is that which is normally derived from or associated with humans in such a way that raises concerns about privacy, confidentiality, and/or yields cultural implications (*e.g.*, stigmatizing behaviors that may be associated with a social or ethnic group). Big Data phenotypes are becoming more common, such as the many phenotypes that may be exhibited by individuals with multiple diseases over the course of their lifetimes and recorded in electronic medical records (Ritchie, et al., 2010). In this context, phenotypic observations are becoming a substrate for discovery research, (Denny, et al., 2010) as well as remaining essential for traditional forms of translational and clinical research. The focus of this segment of the report is on sensitive phenotypes that are derived from human data: much of that from clinical research and/or healthcare operations.

There exist a set of critical issues to resolve in order to share phenotypic data. Some specific, short term goals include the need to:

- provide transparency regarding current policies
- develop a common language for permitted and inappropriate use of data
- establish an appropriate forum to draft the policies

Data Governance

Access to and analysis of phenotypic data is challenging and involves trade-offs when the data is de-identified. Data de-identification itself is an emerging, *bona fide* scientific sub-discipline of informatics, and methods for establishing quantitatively residual re-identification risk are an important and evolving area of information science research. Since the handling of potentially sensitive phenotypic data requires a combination of technology and policy components, establishing clear acceptable-use policies — including penalties for mishandling disclosed data — is an important facet of establishing networks of trusted parties. For example, the data could be governed in ways that limit data exposure to only those investigators with well-established motives and who can be monitored to assure the public that their data are not misused (Murphy, Gainer, Mendis, Churchill, & Kohane, 2011).

Methods for Data Sharing

A responsible infrastructure for sharing subject phenotypes must respect subject privacy concerns and concerns regarding data ownership. Regarding data from an individual's electronic health record, the enterprise that makes that data available may expose itself to risks, and thus there is enormous reluctance to release fully identified electronic health records outside of the originating organization. Countering such risk prescribes sharing solutions that are more complicated than other types of simple data transmission to a public research database, and various solutions have been developed to address these real-world concerns. For example, distributed queries against data repositories may allow extracts of data to be released that are subsets of full medical records (Weber, et al., 2009), and other models for distributed computation also contribute to preservation of individual privacy (Wu, Jian, Kim, & Ohno-Machado, 2012). Institutions often find it reassuring if they know that these data will be used for a specific purpose and then destroyed. Conceptually, queries could also be performed within an institution by some distributed system such that no data at all would need to be released; however, there is some re-identification risk even when this type of solution is employed (Vinterbo, Sarwate, & Boxwala, 2012).

Distributed query systems require thoughtful decisions about data ownership. At one extreme, a central agency such as the NIH could control the queries. On the other hand, so-called peer-to-peer distributed query systems could be employed, which negotiate independently every link of one phenotype data owner to another. The NIH should convene a workshop of experts to provide advice on the merits of various types of query systems.

Data Characterization

Researchers use human phenotype data derived from electronic health records and that is a byproduct of care delivery, data collected during research studies, and data acquired from the environment to define phenotypes that can be shared among researchers at different institutions with variable health care delivery systems. The management and publication of such "common" phenotype definitions will be an important aspect of progress in discovery research going forward, and thus it is vital to derive a workable solution for maintaining these definitions.

Phenotypic data have limitations in accuracy and completeness. There are no easy solutions to this phenomenon; however, descriptions that help document accuracy and completeness that can be transferred among institutions will promote a greater understanding of the inherent limitations. Provenance must be considered: When data are taken out of the context in which they were collected, some features can be lost. Means to retain context include requiring standard annotations describing the precise conditions under which data are collected.

The simple act of naming an attribute of a patient or research subject is usually a local process. The extreme diversity and richness of humans, in general, creates this variability. It is not possible to pre-determine every attribute of a human phenotype that a researcher collects and assign it a standard name. To address this challenge, naming is usually standardized after data collection, and local naming conventions are then mapped to agreed-upon standards. Effective and widely available tools for mapping local nomenclature to standard nomenclature would be a critical resource.

Special Considerations for Imaging Data

Remarkable advances in medical imaging enable researchers to glean important phenotypic evidence of disease, providing a representation of pathology that can be identified, described, quantified, and monitored. Increasing sophistication and precision of imaging tools has generated a concomitant increase in IT needs, based on large file sizes and intensive computing power required for image processing and analyses. Thus, any discussion of informatics and IT infrastructure to support current and near-future requirements for the management, integration, and analysis of large biomedical digital datasets must also include imaging.

The DIWG recognizes that the fields of radiology and medical imaging have been pioneers in the creation and adoption of national and international standards supporting digital imaging and interoperability. The adoption of these standards to achieve scalable interoperability among imaging modalities, archives, and viewing workstations is now routine in the clinical imaging world. Indeed, many of the work flow and integration methods developed within radiology now serve as a model for information system interoperability throughout the healthcare enterprise via the Integrating the Healthcare Enterprise initiative, which is used to improve the way computer systems in healthcare share information by promoting the coordinated use of established standards. One such standard, DICOM (Digital Imaging and Communications in Medicine) details the handling, storing, printing, and transmitting of medical imaging information such that DICOM files can be exchanged between two entities.

Unfortunately, however, translating this success in universal adoption and leveraging of accepted standards for digital medical imaging in the clinic has not occurred to a significant extent with regard to research applications. Significant barriers to scalable, seamless, and efficient inter-institutional digital-image dataset discovery and consumption for research still exist, as described below.

“Impedance Mismatch” Between Clinical Imaging Archives and Research Archives

While DICOM has been a critical enabling standard for clinical digital image management, standard DICOM server/client data transfer methods have not served inter-institutional digital image dataset exchange for research applications. Making matters worse, no research inter-institutional digital image exchange methods are natively supported by clinical image management vendors.

Federal Privacy and Security Regulations

Protecting individually identifiable health information while retaining the research utility of an image dataset (*e.g.*, associating image data objects with other patient-specific phenotypic evidence) is not a trivial endeavor.

Highly Constrained, “Siloed” Research Imaging Archives/Architecture

Existing research imaging archives, which are typically designed to test a specific hypothesis and are often used by a predefined group of investigators, may lack flexibility with respect to data schema and data discovery, accessibility, and consumption — especially for future, unanticipated use cases. This “optimized for one hypothesis” approach can result in a proliferation of siloed image archives that may lack interoperability and utility for future hypotheses/use cases.

“Central” Inter-Institutional Resources

Central registries that contain “reference pointers” to image datasets that reside within various institutional archives have been used; however, use of these resources requires that the transmission of image datasets must be serviced by each institution. This approach has frequently exposed real-world operational performance inefficiencies and security risks. Moreover, it is unlikely that such models can sustain a usefully persistent resource beyond the lifespan of the original research.

2.3 Recommendation 1: Promote Data Sharing Through Central and Federated Repositories

The NIH should act decisively to enable a comprehensive, long-term effort to support the creation, dissemination, integration, and analysis of the many types of data relevant to biomedical research. To achieve this goal, the NIH should focus on achievable and highly valuable initiatives to create an ecosystem of data and tools, as well as to promote the training of people proficient in using them in pursuing biomedical research. Doing so will require computational resources, data, expertise, and the dedication to producing tools that allow the research community to extract information easily and usefully.

Recommendation 1a. Establish a Minimal Metadata Framework for Data Sharing

A critical first step in integrating data relevant to a particular research project is enabling the larger community access to the data in existing repositories, as well as ensuring the data's interoperability. The NIH should create a centralized, searchable resource containing a truly minimal set of relevant metadata for biomedically relevant types of data. Such a resource will allow the research community to broadly adopt data dissemination and retrieval standards.

To ensure broad community compliance, the NIH should set a low barrier for annotating and posting metadata. Furthermore, to incentivize investigators, the agency should mandate the use of data annotation using these standards by tying compliance to funding. Post-publication, public availability of data could also be required, but not necessarily via a centralized database. For example, posted data sets could declare their format, using extant community standards, when such are available. It is important to recognize in this context that as technology and science change, data producers need flexibility. Likewise, searchable resources should keep to an absolute minimum any unnecessary mandates about formats and metadata, and be prepared to rapidly accept new formats as technology shifts.

Special considerations exist for the development of metadata related to molecular profiling, phenotype, and imaging. For example, critical elements of metadata frameworks involving these types of data include the use of standard terminologies to refer to basic biological entities (*e.g.*, genes, proteins, splicing variants, drugs, diseases, noncoding RNAs, and cell types). In addition, establishing ontologies of biological relationships (*e.g.*, binding, inhibition, and sequencing) would help to characterize relationships within the dataset(s). Such choices will be best made at the data implementation stage. For clinically derived phenotype data, a standards development process for representing this complex research data in computer interpretable formats would facilitate data sharing.

Several successful precedents exist for data sharing standards from the transcriptomics, proteomics, and metabolomics communities, as well as from older efforts in DNA sequencing, protein three-dimensional structure determination, and several others. Community-driven efforts have proposed useful checklists for appropriate metadata annotation, some of which have been widely adopted. These include: MIAME (Minimum Information about a Microarray Experiment (FGED Society, 2010)), MIAPE (Minimum Information about a Proteomic Experiment (HUPO Proteomics Standards Initiative, 2011)) and MIBBI (Minimum Information for Biological and Biomedical Investigation (MIBBI Project, 2012)). Importantly, the Biositemap effort, a product of the NIH Roadmap National Centers of Biomedical Computing, has created a minimal standard already implemented by the eight NCBCs for representing, locating, querying, and composing information about computational resources (http://biositemap.org). The underlying feature in all these projects is that they provide information sufficient for a motivated researcher to understand what was measured, how it was measured, and what the measurement means.

In setting its own standards, the NIH should learn from similar efforts related to interoperability standards, such as TCP/IP or the PC architecture, to avoid obvious conflicts of interest. It is crucial that the community that sets standards is not permitted to constrain the design of or benefit from the software that uses them, either economically or by gaining an unfair advantage to their tools. For instance, an academic group that supports analytical tools relying on a specific standard may propose that framework for adoption, but the group should not be in a position to mandate its adoption to the community. The DIWG has learned that Nature Publishing Group is "developing a product idea around data descriptors," which is very similar to the metadata repository idea above (Nature Publishing Group, 2012). Thus, there is a pressing need for the NIH to move quickly in its plans and implementation of setting metadata standards.

Standards and methods that facilitate cloud-based image sharing could be advantageous, but existing clinical production system image vendors have been slow to adopt them. Accordingly, the NIH should encourage the development of a readily available free/open source (ideally virtualizable) software-based edge-appliance/data gateway that would facilitate interoperability between the clinical image archive and the proposed cloud-based research image archive. The NIH could consider as models a number of existing “edge appliance”/state aggregators: the National Institute of Bioimaging and Bioengineering (NIBIB)/Radiological Society of North America (RSNA) Image Share Network edge appliance, the ACR TRIAD (American College of Radiology Transfer of Images and Data) application and services, as well as NCI-developed caGRID/Globus Toolkit offerings. Given the decreasing cost of cloud-based persistent storage, a permanent persistent cloud-based imaging research archive may be possible. This accessibility should be lower the "hassle barrier" with respect to data object recruitment, discovery, extraction, normalization, and consumption. The NIH should also encourage leveraging well-established standards and methods for cloud-based/internet data representation and exchange (such as XML, Simple Object Access Protocol or SOAP, and Representational State Transfer or REST).

Since future uses of data objects cannot be reliably predicted, the NIH should consider:

- adopting a "sparse (or minimal) metadata indexing" approach (like Google), which indexes image data objects with a minimal metadata schema
- adopting a "cast a wide net and cull at the edge/client" strategy (like Google)

Although hits from a query that are based on use of a minimal metadata schema will result in a high proportion of false-positive image data object candidates, current and near-future local-client computing capabilities should allow investigators to select locally the desired subset of data objects in a relatively efficient manner (especially if image dataset metadata can be consumed granularly. Candidate image data objects should include associated “rich” and granular image object metadata via XML for subsequent refined/granular culling. Such a "sparse metadata indexing" model will hopefully improve compliance from investigators who will be expected to contribute to such an image repository.

The DIWG recognizes the need to efficiently extract metadata and specific image subsets that exist within DICOM image datasets without transmitting and consuming the entire DICOM object. The NIH should consider as preliminary models evolving standards such as Medical Imaging Network Transport (MINT), Annotation and Image Markup (AIM), and extensions to Web Access to DICOM Persistent Objects (WADO), along with access to associated narrative interpretation reports via DICOM SR (Structured Reporting). Ideally, image data object metadata schema should allow the association with other patient/subject-specific phenotypic data objects (*e.g.*, anatomic pathology, laboratory pathology) using appropriate electronic honest broker/aliasing, HIPAA/HITECH[6]-compliant approaches.

A bold vision for biomedical data and computing becomes significantly more complex due to the needs of the clinical research community and for those investigators dealing with human genotypes. The confidentiality issues, as well as the differences between basic science and clinical investigations, create special requirements for integrating molecular and clinical data sets. For example, while providing access to minimal metadata may reduce the risk of the future re-identification of next-generation sequencing samples, the value of those data is lower than that of the actual sequences. Current interpretation of the HIPAA Privacy Rule[7] with respect to the use of protected health information for research purposes restricts which identifiers may be associated with the data in order meet the de-identification standard beyond what many researchers would consider useful for high-impact sharing of clinical data.

[6] Health Insurance Portability and Accountability Act of 1996/ Health Information Technology for Economic and Clinical Health Act

[7] 45 CFR Part 160 and Subparts A and E of Part 164

Recommendation 1b. Create Catalogues and Tools to Facilitate Data Sharing

Another challenge to biomedical data access is that investigators often rely on an informal network of colleagues to know which data are even available, limiting the data's potential use by the broader scientific community. When the NIH created ClinicalTrials.gov in collaboration with the Food and Drug Administration (FDA) and medical journals, the resource enabled clinical research investigators to track ongoing or completed trials. Subsequent requirements to enter outcome data have added to its value. The DIWG believes that establishing an analogous repository of molecular, phenotype, imaging, and other biomedical research data would be of genuine value to the biomedical research community.

Thus, the NIH should create a resource that enables NIH-funded researchers to easily upload data appendices and their associated minimal metadata to a resource linked to PubMed[8]. In this scenario, a PubMed search would reveal not only the relevant published literature, but also hyperlinks to datasets associated with the publication(s) of interest. Those links (and perhaps archiving of the data itself) could be maintained by NCBI as part of the PubMed system and linked not only to the literature but also to the searchable metadata framework described above.

Recommendation 1c. Enhance and Incentivize a Data Sharing Policy for NIH-Funded Data

Most of the data generated by investigators in the course of government agency-funded research is never published and, as a result, never shared. The NIH should help set reasonable standards for data dissemination from government-funded research that extend the current requirements for its 2003 data sharing policy (NIH-OD-03-0320)[9]. In practice, there is little specific guidance to investigators and reviewers on data sharing other than the requirement that there be a data sharing plan for grants over $500,000. Moreover, there is little critical review of the data sharing component of grant applications. The DIWG believes that a more proactive NIH policy, combined with an available data sharing infrastructure such as that outlined above, would give more substance to the NIH data sharing requirement. For instance, the NIH should consider whether it is reasonable to require that any data included in NIH-funded grant progress reports is made available to the community following a reasonable embargo time (*e.g.*, 2 to 3 years), within applicable HIPAA regulations and respect for individual consent. Additionally, any NIH-funded research should require, for example, digital image data objects to be "usefully persistent" beyond the lifespan of the original research and to be accessible by others.

The DIWG suggests that use of the data sharing resource described above would be voluntary but incentivized. As registered users select and access or download data, a record of the data access or download (and perhaps the results of follow-up automated inquiries regarding the outcome of the data use) would be maintained by NCBI and forwarded to the electronic research administration (eRA) Commons infrastructure so that it could be used in at least two ways:

- as a report to study sections evaluating current grant proposals of previously funded investigators showing whether and how much of their previous research data has been used by others
- as a summary of the numbers of accesses or downloads and any associated usage data made available to the investigators themselves, which could be downloaded from eRA Commons and included in academic dossiers for promotion and tenure actions

The NIH should also work to make sure that data sources for both published and unpublished studies are appropriately referenced in publications and that data dissemination does not constitute a precedent for rejection by journals.

[8] http://www.ncbi.nlm.nih.gov/pubmed/

[9] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

The National Science Foundation's (NSF's) published data sharing policy[10] stipulates that all proposals submitted from January 2011 onwards must include a data management plan. The policy reads:

> *Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.*

As noted earlier, the NIH also has a data sharing policy published in 2003. The NIH should further refine its policy to include mandatory deposit of metadata for any new data type generated by the research community it funds in a framework such as the one described above. Metadata deposits should be monitored to avoid proliferation of data "dialects" that provide virtually the same information but rely on different metadata standards.

Finally, many institutions have concerns regarding liabilities for inappropriate use of certain kinds of shared data that may prevent them from participating in data sharing with other institutions. The situation is particularly evident for clinically related phenotype and imaging data. The NIH should develop simplified model data use agreements that explicitly identify permitted and restricted uses of shared data, and it should disseminate these model agreements widely to the research community. Doing so will help the NIH ensure that sharing of this type of data does not have such high barriers that often limit or inhibit data sharing.

2.4 Recommendation 2: Support the Development, Implementation, Evaluation, Maintenance, and Dissemination of Informatics Methods and Applications

Biomedical research analytical methods and software are often in the early stages of development, since the emerging data are several scales larger and more complex than previously produced and require new approaches. The DIWG recognizes that such development will need to continue for the decade(s) ahead.

Recommendation 2a. Fund All Phases of Scientific Software Development

The development and implementation of analytical methods and software tools valuable to the research community generally follow a four-stage process. The NIH should devote resources to target funding for each of these stages:

- **prototyping** within the context of targeted scientific research projects
- **engineering and hardening** within robust software tools that provide appropriate user interfaces and data input/output features for effective community adoption and utilization
- **dissemination** to the research community — this process that may require the availability of appropriate data storage and computational resources
- **maintenance and support** is required to address users' questions, community-driven requests for bug fixes, usability improvements, and new features

This approach should be institutionalized across the NIH with appropriate funding tools for each one of the four stages, as a way for the informatics and computing community to create, validate, disseminate, and support useful analytical tools.

Among the areas for potential investment in software development are federated query systems that answer practical, real-world questions involving phenotype and associated molecular data including genomic, proteomic and metabolomics data. Since there is unlikely to be a single standard nomenclature for data representation across all research institutions, there exists a real need for software tools that map local nomenclature to a standard naming and coding system.

[10] http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Recommendation 2b. Assess How to Leverage the Lessons Learned from the NCBCs

In their June 1999 Biomedical Information Science and Technology Initiative (BISTI)[11] report, the ACD Working Group on Biomedical Computing recommended that:

> *The NIH should establish between five and twenty National Programs of Excellence in Biomedical Computing devoted to all facets of this emerging discipline, from the basic research to the tools to do the work. It is the expectation that those National Programs will play a major role in educating biomedical-computation researchers.*

This recommendation resulted in the use of the NIH Common Fund to establish approximately eight large centers: the National Centers for Biomedical Computing (NCBCs). Since these centers were established eight years ago, many lessons have been learned. Multiple members of the DIWG have participated in the NCBC program, and thus the DIWG's recommendations are based in part on direct experience with this initiative. Additionally, the NIH convened an independent, external panel in 2007 to perform a mid-course program review of the NCBCs. While the NIH intends to perform and publish additional assessments of the NCBC initiative, the draft mid-course review report is included as an Appendix of this report (see Appendix, Section 6.2), to provide a preliminary external view.

The NCBCs have been an engine of valuable collaboration between researchers conducting experimental and computational science, and each center has typically prompted dozens of additional funded efforts. One drawback to the program, however, has been that the small number of active centers has not covered effectively all relevant areas of need for biomedical computation or for all of the active contributing groups. For example, the mid-course review report highlighted a number of grand challenges in biomedical computing that have not been currently addressed by the NCBCs:

- establishing a large-scale concerted effort in quantitative multi-scale modeling
- developing methods for “active” computational scientific inquiry
- creating comprehensive, integrated computational modeling/statistical/information systems

Moreover, due to the limited funding of the NCBC program and to the size of the overall research area, there is virtually no overlap of focus among the current centers. As a result, there has been less opportunity for synergy and complementary approaches of the type that have universally benefited the research community in the past.

NIH should consider the natural evolution of the NCBCs into a more focused activity, whose implementation is critical to the long-term success of initial efforts to integrate the experimental and the computational sciences. A large body of collaborating R01s would provide additional national participation, complementary skills and expertise for collaboration. The complexity, scope, and size of the individual centers should be reduced while increasing the number of centers. More targeted foci or areas of expertise would enable a more nimble and flexible center structure. The NIH should also encourage and enable more overlap between centers, to facilitate collaboration.

2.5 Recommendation 3: Build Capacity by Training the Work Force in the Relevant Quantitative Sciences such as Bioinformatics, Biomathematics, Biostatistics, and Clinical Informatics

Biomedical data integration must be linked not only to the creation of algorithms for representing, storing, and analyzing these data, but also to the larger issue of establishing and maintaining a novel workforce of career biomedical computational and informatics professionals.

[11] http://www.bisti.nih.gov/library/june_1999_Rpt.asp

Recommendation 3a. Increase Funding for Quantitative Training and Fellowship Awards

Rough estimates of the NIH training and fellowship grants in these domains over the past several years show that the financial commitment has been relatively constant (see Appendix, Section 6.3). The DIWG believes instead that NIH-funded training of computational and quantitative experts should grow to help meet the increasing demand for professionals in this field. To determine the appropriate level of funding increase, the NIH should perform a supply-and-demand analysis of the population of computational and quantitative experts, as well as develop a strategy to target and reduce identified gaps.

The NCBCs should also continue to play an important educational role toward informing and fulfilling this endeavor. In addition, since the 60 NIH-funded CTSA sites have already established mechanisms to create training programs for clinical informatics and biomedical informatics, they play another important educational function. To that end, curricula for the CTSA programs are in various stages of development, and an organized CTSA training consortium meets periodically to share efforts in clinical research methods, biostatistics, and biomedical informatics.

Recommendation 3b. Enhance Review of Quantitative Training Applications

The NIH should investigate options to enhance the review of specialized quantitative training grants that are typically not reviewed by those with the most relevant experience in this field. Potential approaches include the formation of a dedicated study section for the review of training grants for quantitative science (*e.g.,* bioinformatics, clinical informatics, biostatistics, and statistical genetics).

While the CTSA sites and the National Library of Medicine (NLM) fund most of the training of clinically oriented informatics expertise, funding for bioinformatics, biostatistics, or statistical genomics expertise is often scattered across the ICs. Study sections that review some training grants (*e.g.*, T32s) are typically populated by basic researchers who are not directly involved in either training bioinformaticians or clinical informaticians, and thus these individuals are not optimal peer reviewers for the task at hand. Furthermore, although some reviews of training applications are conducted by various disease-oriented ICs, informatics methods often apply uniformly across diseases.

Recommendation 3c. Create a Required Quantitative Component for All Training and Fellowship Awards

Including a dedicated computational or quantitative component in all NIH-funded training and fellowship grants would contribute to substantiating a workforce of clinical and biological scientists trained to have some basic proficiency in the understanding and use of quantitative tools in order to fully harness the power of the data they generate. The NIH should draw on the experience and expertise of the CTSAs in developing the curricula for this core competency.

3 NIH CAMPUS DATA AND INFORMATICS

3.1 Recommendation 4: Develop an NIH-Wide “On-Campus” IT Strategic Plan

Develop a NIH-wide “on-campus” IT strategic plan to be cost effective by avoiding redundancies, filling gaps, and disseminating successes to the wider NIH community (with particular focus on NIH administrative data, the NIH Clinical Center, and the NIH IT environment).

Recommendation 4a. Administrative Data Related to Grant Applications, Reviews, and Management

Background

Currently, the NIH budget is approximately $31 billion, over 80 percent of which is invested in the biomedical research community spread across U.S. academic and other research institutions. NIH support, administered by the ICs, is in the form of research grants, cooperative agreements, and contracts. Any entity with a budget of this size must review its investments, assess its successes and failures, and plan strategically for the future. In the case of the NIH, the public, Congress, and the biomedical community also want, and deserve, to know the impact of these investments. One challenge for the NIH is to capitalize on new informatics technology to assess the impact of the NIH research investment on both science and improving public health.

Historically, and until recently, NIH administrative staff have had limited tools to retrieve, analyze, and report the results of the NIH collective investment in biomedical research. As a result, such data were accessible only to a few people who were “in the know,” and the analysis was quite limited due to the effort required. Overall evaluation has thus been limited to date, and the ability and potential for strategic planning has not been fully realized. A better way would be a more facile, integrated analysis and reporting tool for use across the NIH by administrative leadership and program staff. This tool (or these tools) would take advantage of recent informatics capabilities.

The NIH and several other Federal agencies currently have access to IT solutions and support for grants administration functions via the eRA systems[12]. Developed, managed, and supported by the NIH’s Office of Extramural Research, eRA offers management solutions for the receipt, processing, review, and award/monitoring of grants.

Most recently, leadership of the National Institute for Allergy and Infectious Diseases (NIAID) developed the “eScientific Portfolio Assistant” software system, which integrates multiple relevant data sources and conducts real-time data analysis (with the ability to drill down to individual research programs or individual researchers) and reporting through user-friendly software interfaces. The system enables retrospective and prospective analyses of funding, policy analysis/strategic planning, and performance monitoring (using a variety of metrics such as publications, citations, patents, drug development, and co-authorships by disease area, region of the country, or institution). The system has enabled more efficient and effective program management and it has also provided an easier way to demonstrate the impact of various programs.

While the NIAID model reporting and analysis system is a major step forward, the DIWG asserts that there should be an NIH-wide coordinated effort to produce or improve such systems. That the ICs are by nature so distributed and functionally separate has led to a fragmented approach that can be inefficient and often redundant, with some areas left unattended. Even currently successful efforts might be even more successful if financial and intellectual capital could be convened. Although the DIWG recognizes and commends ICs, such as the NIAID, with pioneering efforts, a more strategic approach would serve the NIH better.

Specific Administrative Data Recommendations

Update the Inventory of Existing Analytic and Reporting Tools

The DIWG recommends that the inventory of existing efforts and software development be updated and made more widely available across the ICs to be certain of the current status.

[12] http://era.nih.gov/index.cfm

Continue to Share and Coordinate Resources and Tools

The DIWG recommends that the NIH continue to strengthen efforts to identify common and critical needs across the ICs to gain efficiency and avoid redundancy. Although it is clear that the NIH expends great effort to host internal workshops to harmonize efforts and advances in portfolio management and analysis, continued and increased efforts in this area should be brought to bear to benefit both NIH staff and the extramural community.

The DIWG also recommends that the NIH continue efforts to share and coordinate tools across the ICs, with training such that the most expert and experienced as well as newly recruited staff can make effective use of them. In addition, the NIH should make available these query tools, or at least many of them, to the extramural NIH community as well.

Recommendation 4b. NIH Clinical Center

Background

The NIH Clinical Center (CC) is an essential component of the NIH intramural research program, functioning as a research and care delivery site for approximately 20 of the 27 NIH ICs. As noted on its public website (NIH Clinical Center, 2012), the CC is the nation's largest hospital devoted entirely to clinical research. The CC shares the tripartite mission of other academic medical centers: research, patient care, and education. However, the CC is differs by a number of characteristics:

- Every patient is a research study subject.
- The CC has an administrative relationship with other NIH ICs, each one of which has an independent funding appropriation and locally developed procedures and policies.
- The CC has a longstanding history of research and development relationships with academic and private sector developers of diagnostics, devices, and therapies.
- The CC has a vision of outreach, to become a national research resource.
- The CC employs an operations model in which costs are currently borne by NIH-appropriated funds, with recent direction from Congress to investigate the feasibility of external billing where appropriate.

A range of systems and specific computer applications support each of the CC mission areas. These include systems and applications in the following three areas: patient care functions, research, and administrative management.

Findings

Discussions with CC senior staff on the systems infrastructure and information technology issues it confronts led the CC subgroup of the DIWG to the following general observations and findings:

- The highly decentralized NIH management model creates multiple barriers to systems integration.
- Many of the issues of systems integration and applications development to support research and care at the CC are similar to those faced by members of the CTSA consortium.
- The CC is in the position of doing continuous research and development in informatics on an *ad hoc* basis without a dedicated organizational unit to support those activities.
- As a research referral center, the most important CC innovation needed would be a national system of interoperable electronic health records to facilitate internal and external coordination of care. However, this is not a goal not within the CC's purview to achieve.

Specific CC Recommendations

Enhance Coordination of Common Services that Span the ICs

The legislative autonomy and historical independence of the ICs has led to predictable redundancy and variability in technologies, policies, and procedures adopted by the ICs, and the CC pays the price of this high-entropy environment: It must accommodate many different approaches to accomplish identical tasks for each of its IC customers. Although such an "unsupported small-area variation" (Wennberg & Gittelsohn, 1973) was a tolerable cost of doing business in an era of expanding resources, today's fiscal climate does not afford that flexibility. Now, the central focus should be to seek opportunities to achieve economies of scale and adoption of simplified, common technologies and procedures to provide common services to researchers and organizations, both intramural and extramural.

Create an Informatics Laboratory

The DIWG recommends that the NIH create an organizational focal point that functions as an informatics laboratory within the CC, to provide the informatics research and development support needed to achieve its vision of being a national resource and leader. In line with the need to create an expanded workforce with strong quantitative analytical and computational skills (and analogous to the clinical staff fellow and other research training programs of the ICs), this organizational unit should include a training component.

Strengthen Relationships with Translational Activities

The CC should strengthen relationships and communications with NCATS and the CTSA consortium institutions to harvest and share best practices, applications, and lessons learned — particularly for research protocol design and conduct, as well as for research administration.

Recommendation 4c. NIH IT and informatics environment: Design for the future

Background

The DIWG reviewed a high-level overview of the current state of NIH IT systems, including infrastructure and governance. The DIWG considered how best to advance this highly distributed, tactically redundant IT community in ways that it would facilitate the advancement and support of science, as well as retain agility to respond to emerging, high-priority initiatives. In addition, anticipating a more modest NIH funding model, the DIWG discussed key actions to enable a sustainable model for the governance, management, and funding of the NIH IT environment.

Findings

The DIWG's findings are based on a set of interviews with NIH intramural and CC leaders in the IT arena and on various reference materials relevant to IT management at the NIH. Reference materials included examples of strategic plans from different operational perspectives, an overview of NIH major IT systems actively in use, and background documents from the NIH Office of Portfolio Analysis. The DIWG acknowledges the need for a trans-NIH IT strategic plan. This plan should address several components:

- high-performance computing
- bioinformatics capability
- network capacity (wired and wireless)
- data storage and hosting
- alignment of central vs. distributed vs. shared/interoperable cyber-infrastructures
- data integration and accessibility practices
- IT security
- IT funding

Recent and current efforts are underway to assess the major NIH IT enterprise capabilities and services, and this information will be important toward formulating a comprehensive NIH IT strategic plan.

The DIWG also evaluated NIH IT governance at a high level based on information provided by NIH staff. This analysis revealed the existence of multiple structures with varying levels of formality — ranging from the strategic IT Working Group to the more tactical Chief Information Officer (CIO) Advisory Group to domain specific subgroups (such as enterprise architecture and information security). Funding for IT capabilities is for the most part distributed across ICs, although a relatively small portion of funds are allocated for enterprise systems and core enterprise services via a cost-recovery or fee-for-service model.

Specific IT Environment Recommendations

Assess the Current State of IT Services and Capabilities

As the NIH moves to enhance its IT environment in support of Big Data, the DIWG recommends a current-state appraisal that identifies key components and capabilities across the 27 ICs. The NIH is likely unaware of opportunities for greater efficiencies that could be achieved by reducing unnecessary duplication and closing gaps and shortcomings. The current-state appraisal should not only include enterprise IT components, but all decentralized entities such as the CC, and it should provide key data points toward the development of a Strategic Planning Process for IT. The appraisal should not be interpreted as simply an inventory exercise focusing on the details of available hardware, software, and human expertise. As indicated by the recent GAO findings for the FDA (Government Accountability Office, March 2012), this type of appraisal is foundational for the IT future of all Federal agencies, not just the NIH. The current-state appraisal should address:

- computer hardware and software, including attention to mobile applications
- opportunities for NIH-wide procurement, support, and maintenance of hardware that may provide significant financial gains through economies of scale or outright savings
- an IT staff skills inventory, to determine if adequate skills are available to support strategic initiatives and operational needs (This skills inventory can be used to identify training opportunities, provide input for talent management and better align and leverage similar and complementary skills for the NIH.)
- inventory/quantitation of IC IT services to other NIH entities with respect to number of users and discrete services provided to specific audiences (This inventory can provide opportunities to eliminate or consolidate duplicative services, leverage best practices, and help design a pipeline of complementary services.)
- identification of best practices, used to identify and determine which of these practices could be used more widely at the NIH
- broad evaluation of current IT policies, including trans-NIH data standards
- key data repositories and research instrumentation, allowing the NIH to build use cases and user scenarios around the high-impact, high-value data and instrument assets across the agency

Develop a Strategic Planning Process for Trans-NIH IT Design for Big Data

The DIWG recommends that the NIH develop a strategic planning process that establishes a future-state IT environment to facilitate the aggregation, normalization, and integration of data for longitudinal analysis of highly heterogeneous data types, including patient care data, ‘omics data, data from bio-banks and tissue repositories, and data related to clinical trials, quality, and administration. The strategy should incorporate pathways to enable the collection, management, integration, and dissemination of Big Data arising from next-generation sequencing and high resolution, multi-scale imaging studies. Knowledge management components in the plan should include recommended ontologies, terminologies, and metadata, as well as the technologies necessary to support the use and management of these components in trans-NIH and inter-institutional research collaborations in which data could be accessible to individuals with the appropriate consent and compliance approval.

This strategic plan will create a shared vision and a common blueprint toward enabling genotype-to-phenotype based research and translation that will lead to innovative and more targeted and effective patient care. Importantly, the plan should be a process — a continuously evolving program that shapes and provides vision for the IT infrastructure, systems, processes, and personnel necessary to advance NIH intramural research with the appropriate connections to extramural research initiatives. The future state architecture would include:

- a documented business architecture capable of achieving NIH goals through the depiction of business domains and domain-specific functional components
- a documented information architecture clearly showing information that is to be managed by each functional component
- a documented solutions architecture that satisfies the transaction processing, data integration, and business intelligence needs of the business architecture

The process will likely be a federated architecture approach, which will include service-oriented technologies along with object and messaging standards. A key component of the solutions architecture will be to define the role of private and public cloud services.

Develop an Implementation Model for High-Value IT Initiatives

The DIWG recommends that the NIH consider and develop an innovation and implementation model for IT initiatives that highlights centers of excellence or other “bright spots” in a three-phase approach:

- identify individuals or teams who have implemented solutions that can be replicated
- develop a point solution generated by a center of excellence into a proof of concept that may be deployed across multiple ICs
- scale the proof of concept to reach the greater research community, including NIH intramural researchers, NIH extramural researchers, and independently funded industry, academic, non-governmental organizations, and government partners

Continue to Refine and Expand IT Governance

To ensure alignment across all 27 ICs, the NIH should continue to refine and expand its IT governance structure and processes. Currently, the existence of multiple structures at varying levels creates inefficiency as well as potential confusion. For example, the IT Working Group, which is comprised of senior NIH leaders with their charge to view IT strategically, prioritize IT projects and initiatives, and ensure alignment with the NIH mission and objectives, and this may not align with the CIO advisory group, which is more tactical in its efforts and considers deployment of infrastructure and sharing best practices. The NIH IT governance universe also includes a number of domain-specific workgroups, such as those addressing enterprise architecture and information security. The DIWG recommends establishing a stronger, more formalized connection among these governance and advisory groups in order to ensure that tactical efforts support and enable strategic recommendations.

The DIWG also recommends that the NIH establish a data governance committee, charged with establishing policies, processes, and approaches to enable the aggregation, normalization, and integration of data in support of the research objectives of the NIH as detailed in its future-state IT strategic plan. The committee should also focus on standardization of terminologies, metadata, and vocabulary management tools and processes.

Recruit a Chief Science Information Officer for NIH

IT and Big Data challenges cross both scientific program and technical issues. As such, it is crucial to create and recruit a new role of the Chief Science Information Officer (CSIO) for NIH. The CSIO should be a research scientist that can bridge IT policy, infrastructure, and science. The CSIO would work closely with the CIO and serve as the expert programmatic counterpart to the CIO's technical expertise.

Establish an External Advisory Group for the NIH CIO and CSIO

IT is advancing swiftly in the world outside of the NIH. As such, it is more important than ever to create and regularly convene an external advisory group for the NIH CIO and CSIO to help integrate program and technology advances. This advisory body should include external stakeholders in the research community as well as experts in the industry and commercial sector.

4 FUNDING COMMITMENT

4.1 Recommendation 5: Provide a Serious, Substantial, and Sustained Funding Commitment to Enable Recommendations 1-4

NIH funding for methodology and training clearly has not kept pace with the ever-accelerating demands and challenges of the Big Data environment. The NIH must provide a serious and substantial increase in their funding commitment to the recommendations described in this document. Without a systematic and increased investment to advance computation and informatics support at the trans-NIH level and at every IC, the research community served by the NIH will not be able to optimally use the massive amount of data currently being generated with NIH funding.

Moreover, current NIH funding mechanisms for IT-related issues and projects are fragmented among many sources over short temporal periods. This current state poses a significant challenge to upgrading infrastructure for the NIH or for forming multi-year investment strategies. Accordingly, the DIWG recommends that some mechanism be designed and implemented that can provide sustained funding over multiple years in support of unified IT capacity, infrastructure, and human expertise in information sciences and technology.

A final key strategic challenge is to ensure that NIH culture changes commensurate with recognition of the key role of informatics and computation for every IC's mission. Informatics and computation should not be championed by just a few ICs, based on the personal vision of particular leaders. Instead, NIH leadership must accept a distributed commitment to the use of advanced computation and informatics toward supporting the research portfolio of every IC. The DIWG asserts that funding the generation of data must absolutely require concomitant funding for its useful lifespan: the creation of methods and equipment to adequately represent, store, analyze, and disseminate these data.

5 REFERENCES

Phenotype. (2012). Retrieved April 12, 2012, from Wikipedia: http://en.wikipedia.org/wiki/Phenotype

Calvert, S. (2011, March 7). *Md.-based intelligence agencies helped track bin Laden.* Retrieved May 18, 2012, from The Baltimore Sun: http://articles.baltimoresun.com/2011-05-07/news/bs-md-nsa-bin-laden-20110507_1_bin-terrorist-leader-intelligence-agencies

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010, May 1). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics, 26*(9), 1205-10.

FGED Society. (2010). *Minimum information about a nicroarray experiment - MIAME*. Retrieved May 22, 2012, from mged.org: http://www.mged.org/Workgroups/MIAME/miame.html

Galperin, M. Y., & Fernandez-Suarez, X. M. (2012). The 2012 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection. *Nucleic Acids Research (NAR), 40*(Database Issue), D1-D8.

Government Accountability Office. (March 2012). *FDA Needs to Fully Implement Key Management Practices to Lessen Modernization Risks.*

HUPO Proteomics Standards Initiative. (2011, April). *MIAPE: The minimum information about a proteomics experiment*. Retrieved May 22, 2012, from Proteomics Standards Initiative: http://www.psidev.info/groups/miape

Kohane, I. S. (2011). Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 417-428.

Masys, D. R., Jarvik, G. P., Abernethy, N. F., Anderson, N. R., Papanicolaou, G. J., Paltoo, D. N., et al. (2012). Technical desiderata for the integration of genomic data into Electronic Health Records. *Journal of Biomedical Informatics*, In Press.

MIBBI Project. (2012, May 7). Retrieved May 22, 2012, from MIBBI: http://mibbi.sourceforge.net/

Murphy, S. N., Gainer, V., Mendis, M., Churchill, S., & Kohane, I. (2011). Strategies for maintainin patient privacy in i2b2. *Journal of the American Medical Informatics Association*, i103-i108.

National Institutes of Health (NIH). (2011, February 12). *Working Group on Data and Informatics: Mission and Charge.* Retrieved March 14, 2012, from nih.gov: http://acd.od.nih.gov/diwg.htm

National Institutes of Health (NIH). (2012, March 6). *The NIH Almanac - Appropriations.* Retrieved May 22, 2012, from nih.gov: http://www.nih.gov/about/almanac/appropriations/index.htm

National Science Foundation (NSF). (2011, January). *NSF Award and Administration Guide, Chapter VI - Other Post Award Requirements and Considerations.* Retrieved March 14, 2012, from nsf.gov: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp

Nature Publishing Group. (2012, April 4). *Nature Publishing Group releases linked data platform.* Retrieved May 22, 2012, from nature.com: http://www.nature.com/press_releases/linkeddata.html

NIH Clinical Center. (2012). *BTRIS.* Retrieved April 12, 2012, from BTRIS: http://btris.nih.gov/

NIH Clinical Center. (2012). *www.clinicalcenter.nih.gov*. Retrieved April 12, 2012, from NIH Clinical Center: http://www.clinicalcenter.nih.gov

Ritchie, M. D., Denny, J. C., Crawford, D. C., Ramirez, A. H., Weiner, J. B., Pulley, J. M., et al. (2010, April 9). Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *The American Journal of Human Genetics, 86*(4), 560-572.

Roe, C. (2012, March 15). *he growth of unstructured data: What to do with all those zettabytes?* Retrieved May 18, 2012, from Dataversity: http://www.dataversity.net/the-growth-of-unstructured-data-what-are-we-going-to-do-with-all-those-zettabytes/

Vinterbo, S., Sarwate, A., & Boxwala, A. (2012). Protecting count queries in study design. *Journal of the American Medical Informatics Association*, PMID: 22511018.

Weber, G. M., Murphy, S. N., McMurry, A. J., Macfadden, D., Migrin, D. J., Churchill, S., et al. (2009). The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *Journal of the American Medical Informatics Association*, 624-630.

Wennberg, J. E., & Gittelsohn, A. M. (1973, December). Small area variations in healthcare delivery. *Science, 183*(4117), 1102-1108.

Working Group on Biomedical Computing. (1999, June 3). *The Biomedical Information Science and Technology Initiative.* Retrieved March 14, 2012, from nih.gov: http://acd.od.nih.gov/agendas/060399_Biomed_Computing_WG_RPT.htm

World Wide Web Consortium (W3C). (2002, August 1). *XHTML™ 1.0 the extensible hypertext markup language (second edition).* Retrieved May 22, 2012, from W3C.org: http://www.w3.org/TR/html/

Wu, Y., Jian, X., Kim, J., & Ohno-Machado, L. (2012). Grid Binary LOgistic REgression (GLORE): Building Shared Models without Sharing Data. *Journal of the American Medical Informatics Association*, PMID: 22511014.

6 APPENDICES

6.1 Request for Information

Ripple Effect
COMMUNICATIONS, INC.

INTELLIGENT PROJECT MANAGEMENT™

NIH REQUEST FOR INFORMATION: MANAGEMENT, INTEGRATION, AND ANALYSIS OF LARGE BIOMEDICAL DATASETS

ANALYSIS OF PUBLIC COMMENTS

MAY 10, 2012

Executive Summary

In response to the exponential growth of large biomedical datasets, the National Institutes of Health (NIH) Advisory Committee to the Director (ACD) has formed a Working Group on Data and Informatics.[13] The Working Group was charged with the task of providing expert advice on the management, integration, and analysis of large biomedical datasets. As part of the process, the Working Group gathered input from the extramural community through a Request for Information (RFI): "Input into the Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics" (NOT-OD-12-032).[14] Ripple Effect Communications, Inc. was contracted to provide third party analysis of the comments received through the RFI; this report provides analysis of the 50 responders to the RFI and summarizes the 244 respondent suggestions. The Working Group will make recommendations to the ACD to assist in developing policies regarding the management, integration, and analysis of biomedical datasets.

The Data and Informatics Working Group (DIWG) identified a total of six issues and seventeen sub-issues as important to consider for enhancing data management and informatics. The six issues were:

- Scope of the challenges/issues
- Standards development
- Secondary/future use of data
- Data accessibility
- Incentives for data sharing
- Support needs

Respondents were asked to consider the identified issues as they responded to the following three questions:

1. For any of the areas identified above and any other specific areas you believe are worthy of consideration by the Working Group, please identify the critical issues(s) and impact(s) on institutions, scientists, or both.

2. Please identify and explain which of the issues you identified are, in your opinion, the most important for the Working Group to address and why.

3. Please comment on any specific ways you feel these issues would or should affect NIH policies or processes.

DATA AND METHODS

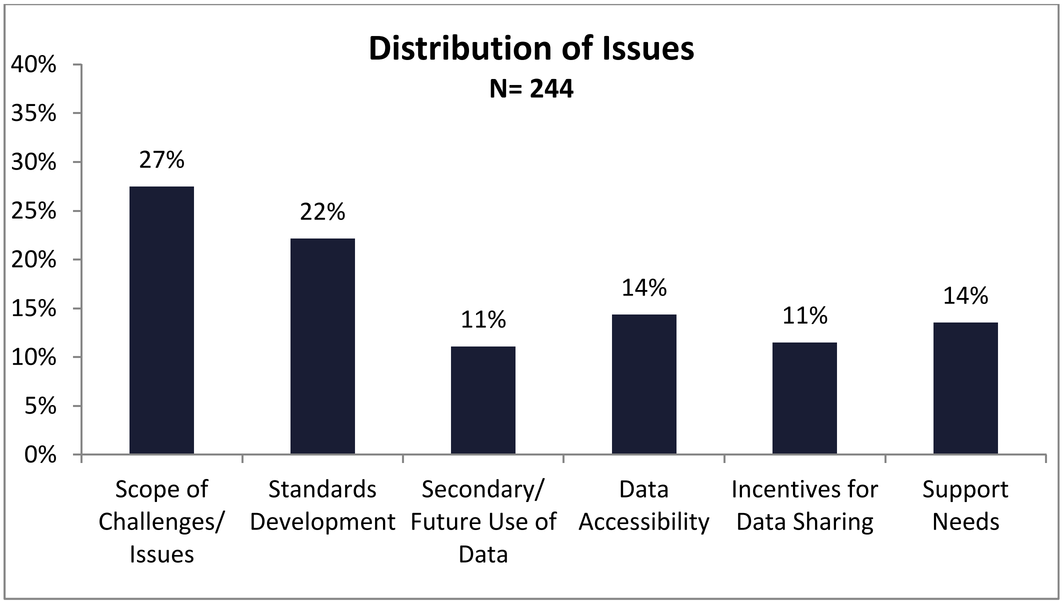
NIH received input from 50 respondents, most of whom provided feedback from a personal perspective (self, 70%; organization, 30%). The 50 respondent submissions were parsed into 244 comments and coded according to the issues identified by the Working Group, as well as by other issues that emerged from the data.

[13] http://acd.od.nih.gov/diwg.htm
[14] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-032.html

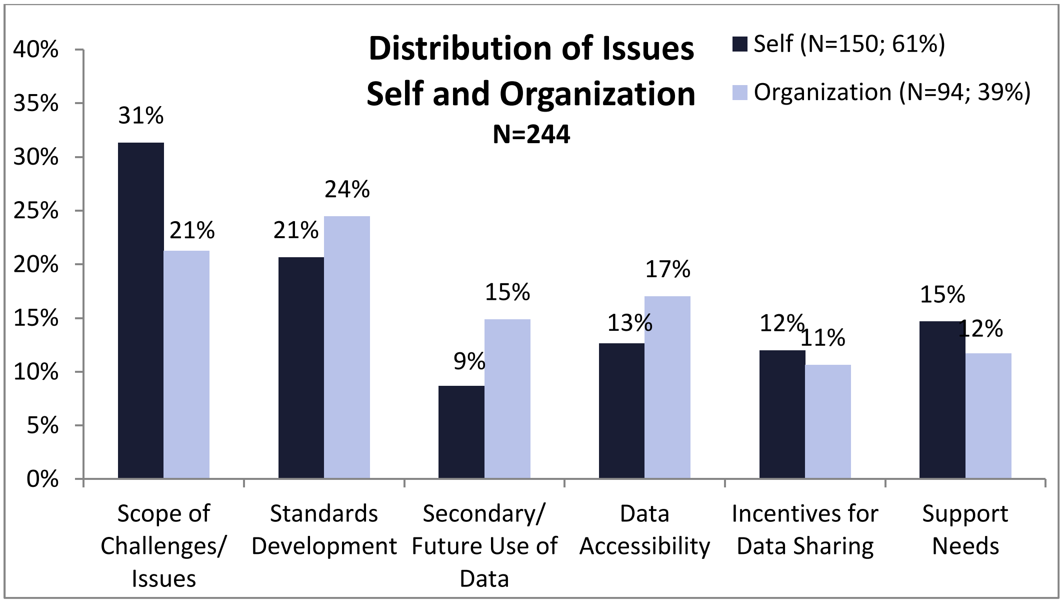
A coding scheme was developed based on six issues and seventeen sub-issues identified by NIH. That structure provided the conceptual foundation, which team members further developed using an iterative, grounded theory approach. The final coding scheme consisted of the six issues and the seventeen sub-issues identified in the RFI, plus three additional sub-issues derived from the data. A total of twenty sub-issues are described in this report. In total, twenty "codes" were applied to the data; these corresponded to the twenty sub-issues.

FREQUENCIES, PRIORITY, AND RECOMMENDATIONS

Of six issues identified by NIH, respondents most frequently commented about the Scope of Challenges/Issues (27%). This issue was followed by Standards Development (22%) and Data Accessibility (14%) to create the top three most frequently-coded issues.



When analyzed by self-reported affiliation, there were slight differences in how the codes were distributed. Those who self-identified as commenting from a personal perspective (self) commented more frequently about Scope of Challenges/Issues, Incentives for Data Sharing, and Support Needs in the review process, compared to those who self-identified as commenting from an organizational perspective (organization).



Priority was assigned to comments when the respondent explicitly stated it was a priority concern. The top three issues when ranked by frequency were the same top three issues when ranked by priority: Scope of Challenges/Issues, Standards Development, and Data Accessibility.

Collectively, respondents recommended that NIH address data and informatics challenges by not only supporting an infrastructure, but also by supporting output and utilization of data needs such as enhanced organization, personal development, and increased funding for tool development.

Contents

Background

NIH REQUEST FOR INFORMATION

In response to the exponential growth of large biomedical datasets, the NIH ACD formed the Working Group on Data and Informatics. The Data and Informatics Working Group (DIWG) was charged with the task of examining issues related to data spanning basic science through clinical and population research; administrative data related to grant applications, reviews, and management; and management of information technology (IT) at NIH. The ACD will make recommendations on the management, integration, and analysis of large biomedical datasets.[15]

To help inform the development of recommendations, the Working Group announced a request for information (RFI), "Input into the Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics" (NOT-OD-12-032),[16] to gather input from various sources, including extramural and intramural researchers, academic institutions, industry, and the public. For the RFI, the Working Group identified the following issues and sub-issues as important to consider when developing recommendations:

- Scope of the challenges/issues
  - Research information lifecycle
  - Challenges/issues faced by the extramural community
  - Tractability with current technology
  - Unrealized research benefits
  - Feasibility of concrete recommendations for NIH action
- Standards development
  - Data standards, reference sets, and algorithms to reduce the storage of redundant data
  - Data sharing standards according to data type (e.g., phenotypic, molecular profiling, imaging, raw versus derived, etc.)
- Secondary/future use of data
  - Ways to improve efficiency of data access requests (e.g., guidelines for Institutional Review Boards)
  - Legal and ethical considerations
  - Comprehensive patient consent procedures
- Data accessibility
  - Central repository of research data appendices linked to PubMed publications and RePORTER project record
  - Models and technical solutions for distributed querying

[15] http://acd.od.nih.gov/diwg.htm
[16] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-032.html

    - Comprehensive investigator authentication procedures
- Incentives for data sharing
    - Standards and practices for acknowledging the use of data in publications
    - "Academic royalties" for data sharing (e.g., special consideration during grant review)
- Support needs
    - Analytical and computational workforce growth
    - Funding for tool development, maintenance and support, and algorithm development

Respondents were asked to consider the identified issues as they responded to the following three questions:

1. For any of the areas identified above and any other specific areas you believe are worthy of consideration by the Working Group, please identify the critical issues(s) and impact(s) on institutions, scientists, or both.

2. Please identify and explain which of the issues you identified are, in your opinion, the most important for the Working Group to address and why.

3. Please comment on any specific ways you feel these issues would or should affect NIH policies or processes.

The online submission process was open from January 10, 2012 through March 12, 2012. This report is an analysis and summary of the public comments and will serve as a tool for the Working Group to use as part of its process for making concrete recommendations to the NIH Director on ways to improve data management and informatics of large biomedical datasets.

THE ROLE OF RIPPLE EFFECT COMMUNICATIONS, INC.

Ripple Effect Communications, Inc. was engaged by the NIH Office of the Director to perform an analysis of the data received through the RFI. As an independent contractor, Ripple Effect staff is not invested in the ACD committee deliberations and therefore has no bias toward the outcomes of the assessment; however, Ripple Effect is uniquely positioned to bring a continuum of working knowledge and expertise about NIH to the analysis process. Our staff's diverse knowledge about NIH allow an open interpretation of respondents' thoughts and ideas, which not only ensures full expression but also provides context for understanding potentially complicated messages.

Ripple Effect was established in 2006 to provide "Intelligent Project Management"^{TM} to the federal government and is often called upon to provide support in one or more of the following areas: Communications, Program and Policy, Technology, Conference and Events Management, Organization and Process Improvement, Research and Analysis, and Project Management. We assess, plan, manage, and execute projects that aid the government (with the current focus on increasing transparency) in transforming into a “people-centric, results-driven, and forward-thinking” organization.

Methods

We engaged both quantitative and qualitative research methods as part of the analysis process. While focusing on and maintaining the integrity and structure of the issues identified by the Working Group, we remained open to the data. We used grounded theory data analysis methods to capture the ideas that were either pervasive enough to warrant their own codes or went beyond the issues identified by the Working Group.

ABOUT THE DATA

A total of 50 respondents provided feedback to the RFI. Respondents provided a total of 244 comments, which were individually coded. All 50 were received through the online submission process that was open from January 10, 2012 through March 12, 2012. Seventy percent of respondents provided feedback from an individual perspective, while 30% identified an organizational affiliation.

ANALYSIS PROCESS

All submissions were uploaded and organized into a central SharePoint database. The data was parsed into individual comments, coded according to the issues identified by the Working Group, and others that emerged from the data, and then analyzed using both SharePoint and Excel.

Code Development

Code development began using the six issues and seventeen sub-issues identified by NIH as the conceptual foundation of the coding scheme. Team members further developed the coding scheme using an iterative, grounded theory approach, which involved studying the data, suggesting themes for inclusion, reviewing code application by other team members, and resolving disagreements.

Conceptually, the codes that emerged from the data were all at the sub-issue level. In addition to the seventeen sub-issues identified by NIH, three additional "data-driven" codes were developed and applied to the data. The final coding scheme (including code descriptions) included six issues and twenty sub-issues (Appendix A). The table below illustrates the conceptual levels and code names used throughout the report.

Issue	Sub-Issue
Scope of Challenges/Issues	Research Information Lifecycle

Issue	Sub-Issue
	Challenges/Issues Faced Tractability with Current Technology Unrealized Research Benefits Feasibility of Recommendations to NIH
Standards Development	Reduction of Redundant Data Storage Standards According to Data Type Metadata Quality Control^ Collaborative/Community Based Standards^ General Guidelines^
Secondary/Future Use of Data	Improved Data Access Requests Legal and Ethical Considerations Patient Consent Procedures
Data Accessibility	Central Repository of Research Data Models and Technical Solutions Investigator Authentication Procedures
Incentives for Data Sharing	Acknowledging the Use of Data "Academic Royalties" for Data Sharing
Support Needs	Analytical and Computational Workforce Growth Funding and Development for Growth

^Data-driven sub-issues

Priority

To assess the priority of the issues for each respondent, we included only the comments in which one of the following conditions was met:

1) The comment was included in response to Question 2, “Please identify and explain which of the issues you identified are, in your opinion, the *most important* for the Working Group to address and why.”
2) The commenter expressed priority by using words such as “critical,” “important,” or “essential.”

If no priority was indicated or if the commenter explicitly expressed that the item was NOT a priority, the comment was not included in the priority analysis.

Analysis was a straightforward count of the number of people who identified each issue and sub-issue as a priority. Priority is presented as an order based on the frequency with which each person identified a code, not as a mathematical rank. Analysis of this sub-group is presented in Section Two of the Findings.

NIH Responsibility

To assess how the respondents believed issues would or should affect NIH policies or processes, we captured and quantified comments that either explicitly expressed an action for NIH to take in order to improve data and informatics or that suggested the issue coded fell under the purview of NIH. Specifically, we included comments only when one of the following conditions was met:

1) The comment was located in response to Question 3, “Please comment on any specific ways you believe these or other issues would or should affect NIH policies or processes.”
2) The commenter specifically stated that NIH should be responsible.
3) The comment addressed an existing NIH program.

If the respondent explicitly stated that the item should NOT be the responsibility or purview of NIH or the comment was general and did not explicitly state NIH responsibility, it was not included in the NIH responsibility analysis.

Analysis occurred in two steps. First, we compared the frequency distribution of all sub-issues identified as an NIH responsibility with the overall dataset. Second, we reviewed data for overarching themes that informed explicit recommendations for NIH. Analysis of this sub-group is presented in Section Three.

Findings

Findings are divided into three sections that reflect different conceptual levels of analysis and respond to the questions posed in the RFI. The first section includes analysis in response to Question 1: “For any of the areas identified above and any other specific areas you believe are worthy of consideration by the Working Group, please identify the critical issues(s) and impact(s) on institutions, scientists, or both.” This section provides a quantitative overview of the primary categories and issues, as well as a quantitative distribution and qualitative analysis of the twenty sub-issues.

The second section addresses Question 2: “Please identify and explain which of the issues you identified are, in your opinion, the most important for the Working Group to address and why.” We coded and quantified the data for respondents that explicitly identified priority issues.

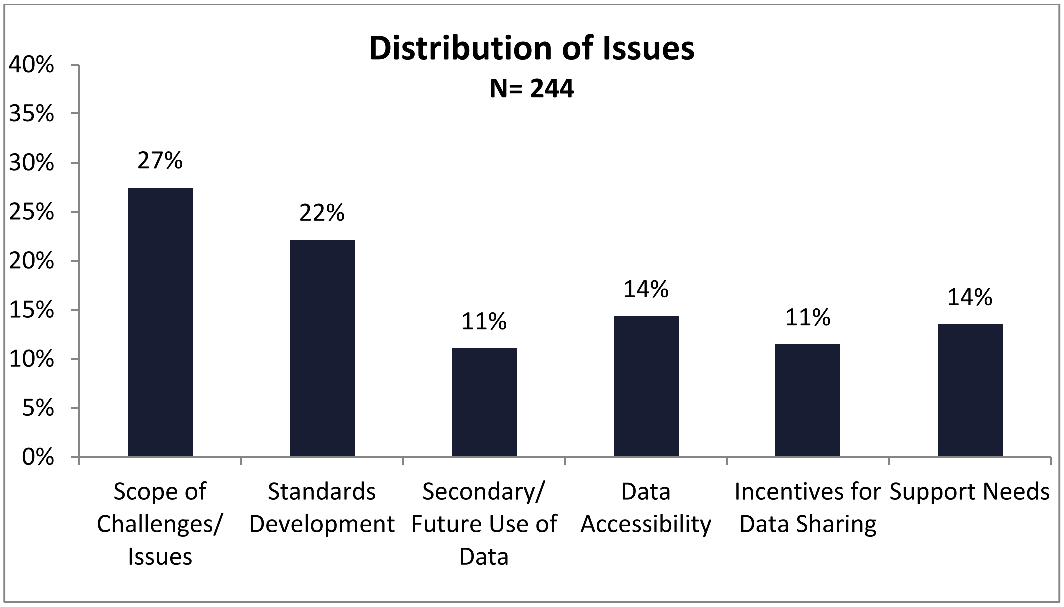
The third section includes a descriptive summary of the ideas commenters presented as relevant to Question 3: “Please comment on any specific ways you believe these or other issues would or should affect NIH policies or processes.” We coded and quantified the comments that referred to specific recommendations for NIH.

SECTION ONE: QUANTITATIVE AND QUALITATIVE ANALYSIS OF CRITICAL ISSUES

A total of 50 (100%) responsive submissions were received and parsed into 244 individual comments. Each comment received one code (corresponding to one sub-issue) and was analyzed for frequency and content.

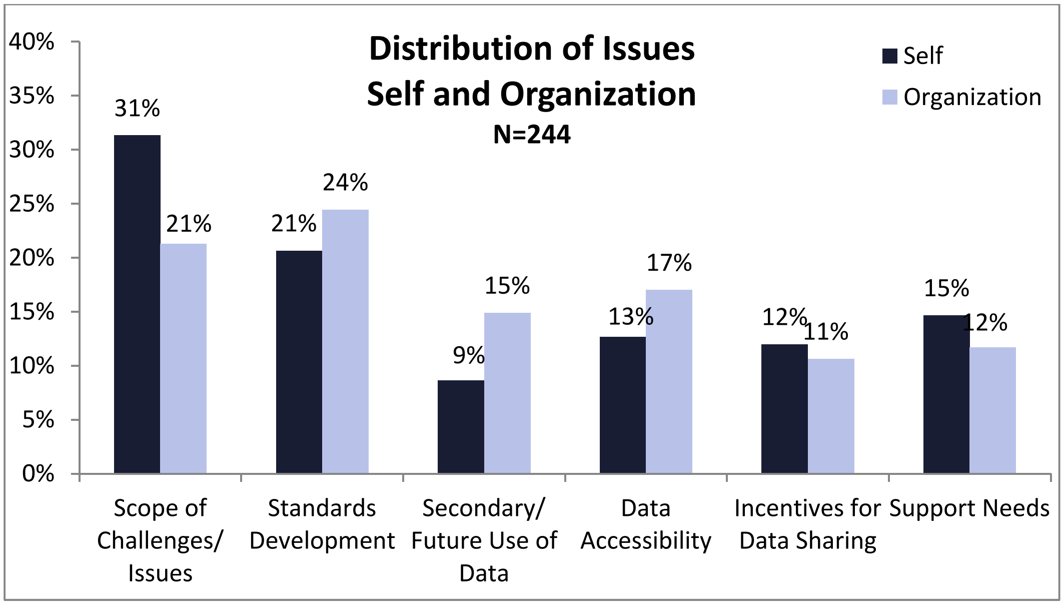
A Quantitative Overview of the Issues

Of the six issues identified by NIH, respondents most frequently commented about the Scope of the Challenges/Issues. The other top issues identified were Standards Development and Data Accessibility. When combined, these top three issues represent approximately two-thirds of all comments.



Issues by Respondent Affiliation

Respondents self-identified with one of two types of affiliation: as an independent individual (self) or on behalf of an organization (organization). Of the total 244 comments received, 150 (61%) were from those identifying as "self" and 94 (39%) were from those identifying as "organization." Those who responded from a personal perspective commented more frequently than organizations about Scope of Challenges/Issues, Incentives for Data Sharing, and Support Needs. Those responding on behalf of an organization commented most frequently on Standards Development, Data Accessibility, and the Secondary/Future Use of Data.

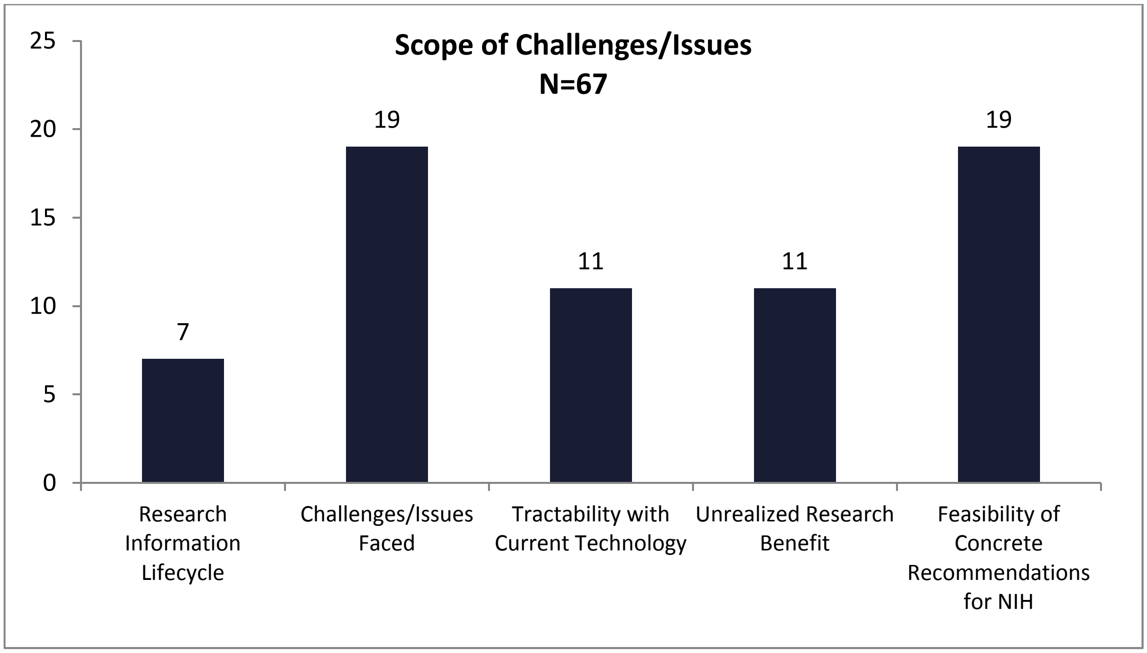


Quantitative and Qualitative Analysis of Issues and Sub-Issues

The six issues and twenty sub-issues, as identified by NIH and derived from the data, are illustrated and discussed here in detail. A graph that summarizes the frequency distribution of comments across all sub-issues is provided in Appendix B. Where relevant, the NIH-identified sub-issues are shown in blue, while data-driven sub-issues are shown in orange.

Issue One: Scope of Challenges/Issues

This issue targeted challenges regarding the management, integration, and analysis of large biomedical datasets. This issue was the most frequently mentioned; approximately one-quarter of all commenters were concerned with the Scope of the Challenges/Issues. Within this category, three leading topics emerged: Feasibility of Concrete Recommendations for NIH, Challenges/Issues Faced, and Tractability with Current Technology. These topics together made up two-thirds of the responses for this issue.



Research Information Lifecycle

For this sub-issue, one respondent outlined a data lifecycle model by describing a scientific community-driven collection of data with a national data infrastructure. In such a community-driven lifecycle, creators of a data set would generate data and input parameters as the first stage. In subsequent stages, other members of the research community would add to the existing data by providing additional context, such as how the data was generated. At the publication and preservation stages, a final detailed description of the data then would be available.

> *An example life cycle is the migration of data from a project collection, to a collection shared with other researchers, to a digital library for formal publication of vetted results, to a reference collection for use by future researchers. (#42)*

When describing the national data infrastructure, one respondent explained that each stage of the community-driven collection would be governed by policies.

Another respondent referred to Charles Humphrey's 2004 overview on research data lifecycles[17] stating that it is applicable to a variety of research disciplines. The respondent noted that, when considering management of analysis of datasets, the roles and responsibilities of the researcher needs to be determined by focusing on documenting the stages of the research lifecycle:

> *Design of a research project*
> *Data collection processes and instruments*
> *Data organization in digital format*
> *Documentation of data analysis process*
> *Publication or sharing of results*
> *Dissemination, sharing, and reuse*
> *Preservation, long-term conservation, and long-term access (#46)*

Other comments revolved around hiring technicians involved in designing methodology, requiring electronic notebooks, maintaining better recordkeeping, and preserving and storing data.

Challenges/Issues Faced

This sub-issue referred to the challenges and issues presented by datasets in the biomedical field. Overall, respondents' comments were divided among data infrastructure, the need for well-trained individuals, and data accessibility, although most comments focused on data infrastructure. One respondent specifically stated that there was a lack of data infrastructure:

> *There are two major barriers to sharing of data: 1) Lack of an infrastructure for data sharing. It's not easy to share. Currently, scientists or universities need to set up their own sharing system (we are doing this using DATAVERSE) but there should be a system put in place by NIH/NLM for widespread sharing of data. Once the systems are in place, scientists will use them. (#1)*

One respondent stated that "we have the information, but we do not know how to use it." Others felt that a data system should be created to integrate data types, capture data, and create "space" for raw data.

Regarding the need for well-trained individuals, one respondent spoke passionately about laying off programmers due to lack of funding. Comments were emphatic about how much harder it is to replace a competent programmer than a lab technician.

[17] Humphrey, C. & Hamilton, E. (2004). Is it working? Assessing the Value of the Canadian Data Liberation Initiative." *Bottom Line*, *17* (4), 137-146.

Regarding data accessibility, most respondents spoke to the difficulty of finding useful data and databases for their particular area of interest, whether it be patient records, health care, or biomedical research. Encountering access issues in our current age of digital technology and electronic records was seen as especially frustrating. One respondent believed that there should be some type of direct access to data records that would facilitate many advances in the biomedical field.

> *What is most puzzling and distressing is that, in spite of our increasingly sophisticated technology and electronic data systems, researchers' direct online access to federal vital records data has become increasingly limited over time, impeding and sometimes precluding potentially valuable etiologic investigations. (#2)*

Tractability with Current Technology

For this sub-issue, there was consensus around a need for tracking current technology for data standards and standardized software. Suggestions to develop standards ranged from performing an analysis of the technology that has been successful or unsuccessful to understanding limitations posed by available computing hardware. Several respondents provided examples of current technology uses and suggestions to accommodate future growth. For example, a suggestion to improve electronic health records (EHRs) was:

> *… to significantly increase the size of the sample (one billion visits per year), the diversity of the population, and the length of follow-up time compared to what is currently feasible. (#4)*

The Nuclear Receptor Signaling Atlas (NURSA) and Beta Cell Biology Consortium (BCBC) were viewed as highly effective efforts that have evolved into successful management of large scale data.

Unrealized Research Benefit

Respondents to this sub-issue consistently agreed that research products involving datasets, data sharing, and administrative data are not being properly utilized. Large amounts of data are not being considered or analyzed. Reasons for such underutilization included poor planning of grant resources, negative results, poor documentation, lack of data sharing compliance, and lack of data retention. Respondents called for open access and offered the Open Government Initiative and the International Household Survey Network as model examples.

> *Great progress has been made in data sharing in many disciplines such as genomics, astronomy, and earth sciences, but not in public health. Developments such as the Open Government Initiative by the US Federal Government and the International Household Survey Network supported by the*

> *World Bank provide a promising start but will require a wider support base for a paradigm shift for data sharing in public health. (#31)*

Respondents believed that providing a more open forum to data sources would improve success rates.

Feasibility of Concrete Recommendations for NIH

This sub-issue captured comments that provided feasible recommendations for NIH to improve data sharing, data storage, data management, etc. Many commenters suggested that NIH maintain an up-to-date data directory, create an organizational structure, obtain adequate memory for computer systems, and develop algorithms.

One respondent contributed step-by-step procedures to manage the influx of large datasets.

> *More pointedly, as NIH moves to larger and larger data sets, and federations of data sets, it will discover that the I/O performance of most systems will be inadequate to handle the volume of data in a timely fashion. Solving this problem requires getting many things right, from organizing the data so that it can be accessed efficiently, to picking representations that allow it to be manipulated efficiently in the available memory of the computer systems, to developing algorithms and data management interfaces that work well with peta- to exabytes of data, and, last but not least, to designing the storage and I/O systems to maximize the transfer rate between disks and memory. (#35)*

Another respondent elaborated on the same concern by providing specific examples in software development and hardware configuration.

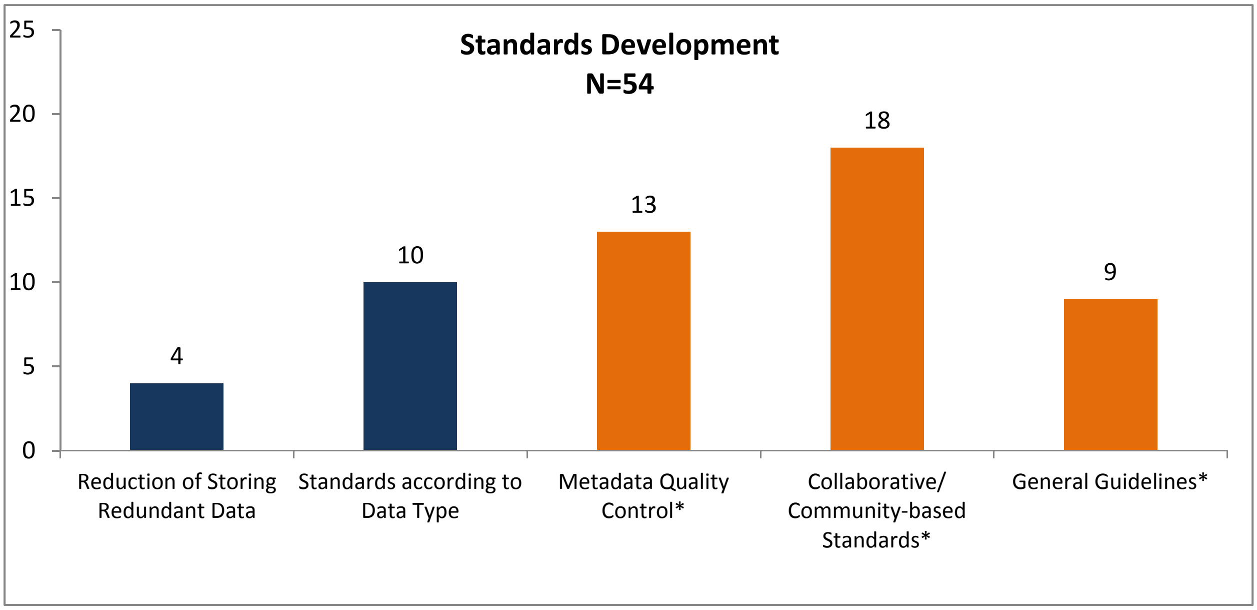
> *What are the non-mainstay innovations that will/could be required? To meet some of the challenges in terms of "population scale" analysis we need a fundamental change in how software is being developed, the methodologies used and the under lying hardware configurations. Such forward thinking seems to be within the remit of the group. Examples of innovations could include: considering how affordable and usable HPC can be made available (e.g. easier to use programmable chips or GPUs, extensions to PIG or other scripting systems for distributed processing/HDFS) or how we can develop scalable/affordable/usable software more easily without introducing constraining requirements on teams (e.g. education, reuse of open-source initiatives (see section 3)). (#14)*

A common suggestion from respondents was the integration of data into a master system. While respondents agreed upon the need for a system, some suggested the goal of this system was data management while others wanted to create a system for publications, patient records, or enforcement of diversity sampling.

Another respondent identified the need for increased training grants that would provide biostatisticians and bioinformatics specialists with strong scientific backgrounds to provide the appropriate level of technical support to assist with large datasets.

Issue Two: Standards Development

Within this issue, respondents felt that it was important to develop organized standards for current data and to also establish standards for future data. The sub-issues originally identified for this issue were joined by three additional sub-issues that emerged from the data (Metadata Quality Control, Collaborative/Community-based Standards and General Guidelines).



Reduction of Redundant Data Storage

Most comments within this sub-issue expressed the opinion that redundancy is an issue primarily because of the increasing amount of data that is being created without oversight or coordination. Respondents suggested strategies for reducing redundant data:

- Establish standards and policies
- Disseminate and preserve data

- Build a proper support network

One respondent commented that data tends to be dispersed; therefore, cross referencing the data is not simple. Possible solutions to remedy the issue were offered.

> *There is a need for better: i) schema integration, ii) schema mappings to navigate from one data source to another, iii) complex join across databases, iv) support for provenance data, v) flexible resource discovery facilitated by a richer metadata registry. [This] item reflects implicit needs for better metadata that will facilitate the selection and the location of distributed data resources. (#43)*

In general, respondents agreed that the identification of data standards, reference sets, and algorithms were strategies to reduce the storage of redundant data.

Standards According to Data Type

Respondents believed that standards should be developed for distinct data types, such as phenotypes, molecular profiling, imaging, raw versus derived, clinical notes, and biological specimens. One universal theme was the need for a consortium to handle the variety of data types, especially because some respondents believed that creating one general standard would be difficult or impossible.

> *While "universal" standards are theoretically appealing, in practice they have proven difficult, if not impossible, to implement. The WGDI must, therefore, avoid a one-size-fits-all approach and should consider a variety of data sharing models and standards to accommodate the diversity of data types. (#18)*

Respondents emphasized the diversity in data types by highlighting features such as the abundance of non-genomic data associated with patients (EEG reports, imaging, biochemical workups, and reactions to therapeutic interventions). To take this concept one step further, one respondent suggested developing a "biomaterials enterprise interlinked for data access and integration."

> *Coordination of acquisition sites for data uploading is a key factor, as is coordination of databases (or synchronization mechanisms if a federated archive is deployed) by data type, e.g., image data vs. genetic data. Biospecimen banking may be optimally conducted elsewhere or separately from the data coordinating center, with the biomaterials enterprise interlinked for data access and integration as needed by project or user. (#27)*

Additionally, respondents agreed that a system should be developed to create consistency in annotating data standards.

*Metadata Quality Control^*

This sub-issue evolved from the data and captured comments related to organizing data and/or improving data quality control with respect to uniform descriptors, index categories, semantics, ontologies, and uniform formats. One respondent specifically noted that this issue was not addressed in the RFI.

> *The current list of areas does not identify data quality as an area of focus for this agenda. There currently exist no established data quality assessment methods, no established data quality standards, and no established data quality descriptors that could be attached to each data set. In the absence of data quality descriptors, a down-stream user of the data has no ability to determine if the data set is acceptable for the intended use. A data set that is acceptable for one use may or may not be acceptable for a different use. (#7)*

Other respondents agreed that data sets lacked well-formed metadata. They believed that the development of standards for metadata is fundamental in ensuring that data will survive and remain accessible in the future.

*Collaborative/Community-Based Standards^*

Some respondents specifically addressed the process of standards development, stating that community-led collaborative efforts were needed to muster broad support for new standards and reduce competing standards. All should have a voice and a stake in this "information ecosystem": researchers, government agencies, universities, students, publishers, industry, associations, educators, librarians, data scientists, patients and study subjects, the public.

> *Development of such a workforce should be modeled on exemplar efforts such as the NSF DataNets, the Digital Curation Center in the UK, and the Australian National Data Service. This community is needed to help shape and support general policy and infrastructure within and among agencies, and to help spread data expertise into the educational and research communities. At the same time, grass-roots 'communities of practice' must engage disciplinary scientists in order to determine how to implement general agency policies. (#45)*

*General Guidelines^*

Some respondents emphasized the need for guidelines on data management, access and sharing, and some included the necessity for training in guideline usage and compliance. Some respondents specified particular guidelines (e.g., for research funders) for archiving and accessing paper records of public health data for future needs. Some focused on cost issues, others on how to determine who should have access. Some listed concrete suggestions for policy:

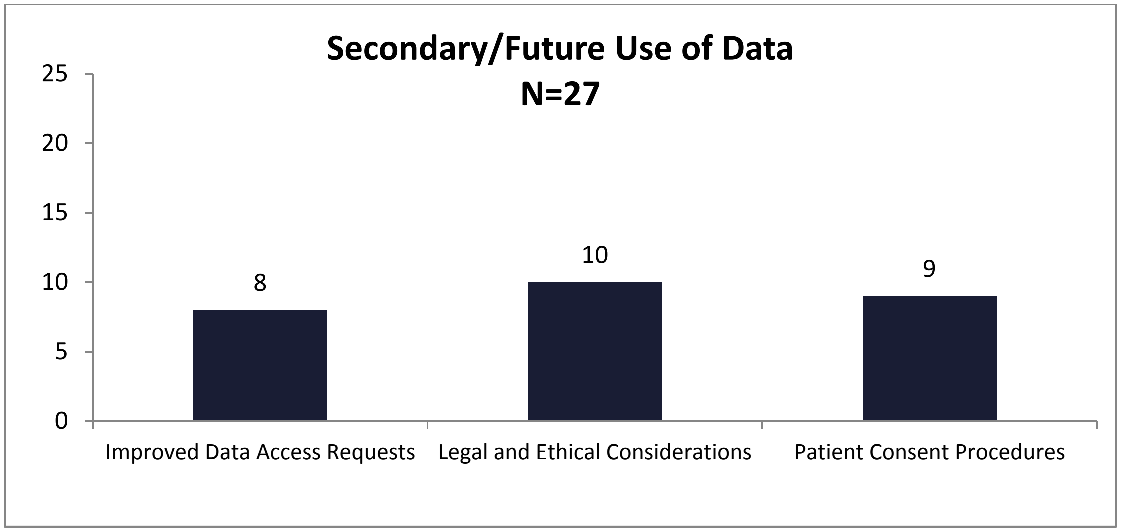
> *Data sharing needs to be built into the research and publication workflow — and not treated as a supplemental activity to be performed after the research project has been largely completed. Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances. Data relevant to public policy should be shared as quickly and widely as possible. (#46)*

All commenters in this category declared that the development of standards and guidelines and policies for data management, access, and sharing, was of critical importance for organizing and utilizing large biomedical datasets.

Issue Three: Secondary/Future Use of Data

Respondents' main suggestion regarding facilitation of the use of data through secondary sources of data was to create commonly-defined data fields with specific structure and standard definitions for methodologies. One respondent spoke to the possible role of the librarian in assisting with building an infrastructure.

> *Again, AAHSL and MLA maintain that librarians have the skills and expertise to assist researchers in understanding the necessity for, and applying the criteria for data definitions so that it can be shared in the future. Librarians can play an important role from the early planning of research proposals to the implementation of data management once a project is funded and should be part of the research team. (#29)*



Others believed that in order to support data for secondary and future use, guidelines and policies would need to be developed to address improvements in data access requests, legal and ethical issues, and patient consent procedures.

Improved Data Access Requests

Several respondents identified the Institutional Review Board (IRB) as a means for improving the efficiency of the request for access to data. In general, respondents felt that IRBs lacked clear guidelines, took a long time to provide approvals back to investigators and project managers, and slowed down the pace of research. The question was posed by a few respondents, "how do we protect privacy without imposing on the pace of many phases in research?" Changes to IRB policies and procedures could improve data access requests.

Legal and Ethical Considerations

Respondents noted that legal and ethical issues complicated data sharing and they relayed concerns that the development of guidelines and regulations for legal and ethical considerations was necessary. In particular, some respondents wanted to ensure that access to secondary data would continue to be free of charge to avoid an unfair barrier for researchers with less funding.

> *To facilitate the discovery process through secondary analyses and data repurposing, database access is optimally free of charge to authorized investigators, regardless of location or primary discipline, with costs of data management and curation underwritten by each e-infrastructure funding source(s) (mostly, NIH), at realistically sufficient levels of funding support. Fee-for-access, even by a sliding scale arrangement, encumbers discovery science by limiting it to the financially privileged. Establishing and maintaining a level playing field in access, scientific community-wide, is thus vital to the data informatics or e-structure enterprise. (#27)*

Developing a framework for determining ownership of data from publically-funded projects was cited as necessary to reduce duplicative claims of ownership by investigators and institutions. Policies of global health agencies and the Bill and Melinda Gates Foundation were cited as exemplars that reflect the key principles that should be included in such a framework.

> *The Bill & Melinda Gates Foundation identified eight principles: promotion of the common good, respect, accountability, stewardship, proportionality, and reciprocity. In a joint statement, global health agencies proposed that data sharing should be equitable, ethical and efficient. Most of these principles call for: 1) a recognition or reward structure for data collection efforts, 2) responsibility in data use that safeguards privacy of individuals and dignity of communities and 3) the use of data to advance to public good. (#31)*

Respondents highlighted the need for data security, especially with respect to information released through secondary sources or presumed for future use. Appropriate privacy protection must be guaranteed and considered as part of the original design of the data sharing and management system. One respondent referenced the Genome-Wide Association Studies (GWAS) results “that restricted info can be obtained by asking the right questions about data.” (#35)

Patient Consent Procedures

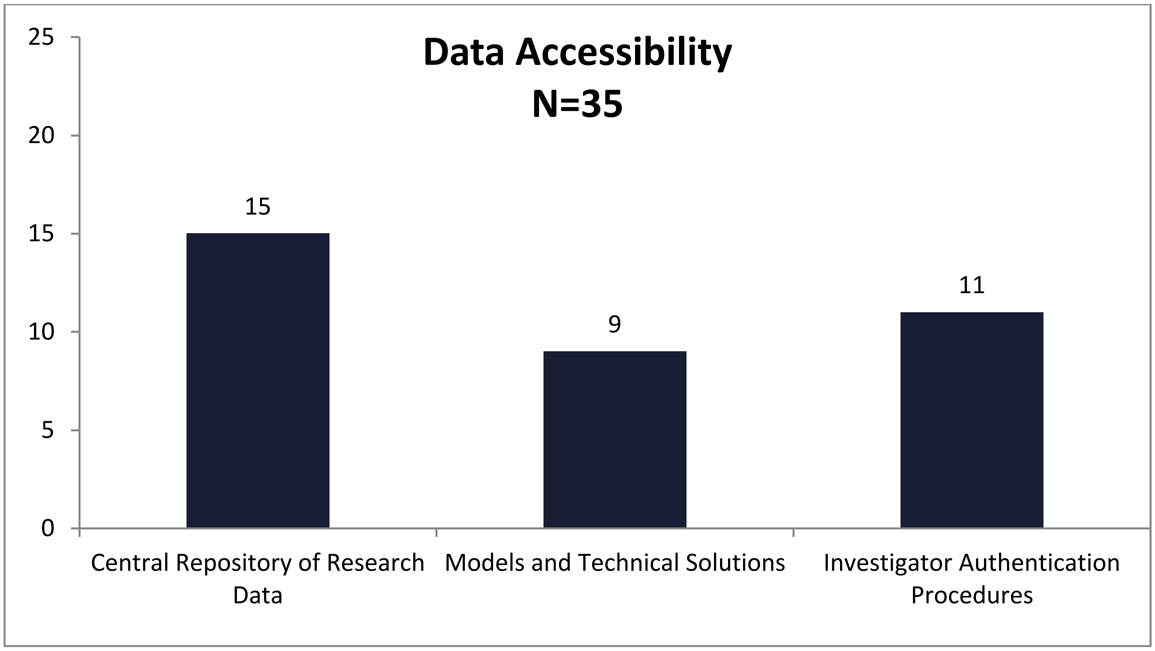
Many respondents believed that the current patient consent procedures are inefficient. One respondent reflected on how the consent process is impeded because there is no clear directive on who owns patient/human subject data.

> *Further, the extent to which data could be shared is constrained by questions of ownership of the data. Funders may feel that taxpayers supported the creation of study-specific data, so that NIH would own the data on behalf of taxpayers. However, in cases where researchers work at health care organizations and build datasets based on the organizations’ data, the parent company may reasonably argue that they own the data and that NIH’s contribution was a modest value-add. Health care organizations will have a need to shelter their data to protect their business from competition and from reputational risk and a duty to safeguard the confidentiality of their patients. Scientific investigators also have a stake in the ownership of the research data; since they invested their knowledge – including knowledge acquired outside of the study-specific work. (#23)*

Comments from other respondents ranged from promotion of an open-ended policy that would allow patients to designate that their data could be used in an unspecified manner to enactment of stricter access policies with governance and oversight (such as a Data Sharing and Publication Committee to control a HIPAA-compliant data system).

Issue Four: Data Accessibility

Most respondents had suggestions about how NIH could provide guidelines and regulations to assist with making data more accessible. One commenter suggested employing the same methods as the journal *Nature,* including requiring the full disclosure of all materials. Another commenter suggested the use of a distributed-computing paradigm or computing “cloud.”



Central Repository of Research Data

Many respondents suggested that a central repository of research data should be developed and handled by NIH. One respondent believed that NIH should work with "university libraries, disciplinary societies, research consortia, and other stakeholders to distribute the many responsibilities associated with establishing and maintaining a trusted repository for digital data" (#15). Others remarked on the financial burden that repositories pose for institutions and emphasized how vital it was for NIH to play a key role to help reduce some of the cost burden.

Respondents acknowledged that there are many existing data repositories and they called for a "directory" of repositories to identify existing datasets. Such a central indexing repository would include links to other repositories, which would help increase access to data. However, respondents recognized that "this is a tremendous undertaking and many datasets that are not federally funded may be excluded from such an approach" (#29). Many suggested that NIH should fund or maintain such repository aggregators.

> *Making public data more visible, navigable, and useful can be accomplished by financing repository aggregators…Financing more projects and tools that promote domain specific databases to push and pull their data to the aggregators and to the Semantic Web will support data sharing. (#49)*

Models and Technical Solutions

One respondent indicated that computational models should be designed to answer specific questions and not for a general purpose. Another respondent called for NIH support so that tools to share data across sites could be streamlined. One comment mentioned the need to develop specialized tools that will provide assistance with "the use and understanding of common data elements and promote open architecture to enable software development for data mining" (#27). These tools will help in data exploration by alleviating limited usability of a database. A commenter reported that building an infrastructure to query several repositories would add value because new discoveries rely on putting together different pieces of information.

Investigator Authentication Procedures

The comments on this sub-issue identified comprehensive procedures that authenticated the data provided was the investigator's own work. One respondent suggested that NIH create a digital author identifier which would provide a digital signature broadly recognized by datasets.

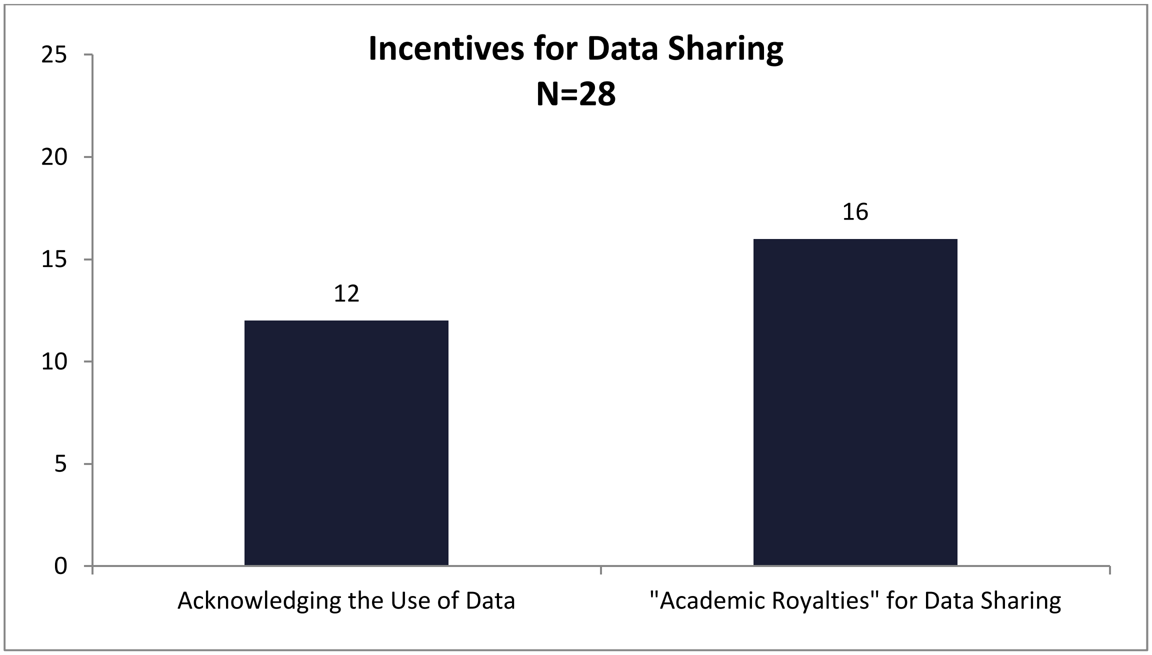
> *Digital Object Identifiers (DOIs) seem to be the best scheme today. Provenance requires that disambiguated authors be assigned to these datasets and as of today no widely accepted scheme exists to provide this identification. (#17)*

Other suggested procedures included the use of social networking tools for investigators to create a catalog and the protection of rights to the use of intellectual property by investigators.

Issue Five: Incentives for Data Sharing

Respondents either agreed that NIH **promote** policies and incentives to encourage data sharing or that NIH **require** data sharing.

> *The NIH should promote data sharing policies and incentives that will encourage data sharing. Without such incentives, researchers may see data sharing as an overhead activity, requiring time and effort with little reward. (#28)*
>
> *The NIH must become less passive with regard to enforcing data sharing by its grantees. If grantees are spending federal research dollars, it is incumbent upon them to preserve the research that these dollars purchase. (#38)*



Acknowledging the Use of Data

Developing standards, policies, and practices for acknowledging the use of data was deemed important by respondents, especially since many commented that researchers do the "bare minimum" to satisfy journal and publication requirements. One respondent stated,

> *There should be incentives for researchers to provide consistent and detailed meta-data annotation to the experimental data they are submitting. Special credit should be given during funding decisions to scientists who not only publish good papers, but also whose data are used by many other people.* (#13)

One respondent suggested that cultural differences play a role in the unwillingness to share data because of the fear of being "scooped." Creating clear incentives for data sharing could combat this fear. Specifically, developing a widely-accepted way to identify the creator of a dataset (such as the use of unique identifiers) would enable tracking of the impact and usefulness of data, as well as provide an easy way to reference data as part of an author's publication record.

"Academic Royalties" for Data Sharing

Most examples of incentives for "academic royalties" were provisions for special considerations in funding decisions. One respondent suggested a sixth scored review criterion for research awards entitled "data sharing track record" to include:

> *1) the number of publications that re-used the data from your lab and you serve as a coauthor of the papers; 2) the number of publications that re-used the data from your lab and you are not a coauthor of the papers. (#10)*

Another respondent believed that “the incentive to share data for the public good for individual investigators and their institutions will be outweighed by the incentive for personal (and institutional) gain.” While this public good versus personal gain theory was seen as a barrier, the respondent thought that an international system may help.

> *An international registration system of collected data in health sciences or publication of datasets after peer review would provide opportunities for considerations of data collection and sharing practices during manuscript or grant reviews and could form an additional basis for promotion and tenure. (#31)*

Respondents shared concerns about unintended consequences of increased data sharing.

> *More significantly perhaps, it is not in the interest of the community if publicly-funded shared data favors researchers with loose ethical standards by granting them exclusive access to a valuable resource. NIH should establish and enforce guidelines to ensure that incentives for data sharing do not compromise existing standards in the scientific community, such as for example standards of academic authorship... (#37)*

> *Policies that support new indicators (e.g., bibliometric measures other than first or senior authored publications) of individual contributions to collective work need to be developed. Further, the federal funding data deposition policy, although requiring data deposition as part of publication, does not yet have a method to track the use of the dataset, nor a dedicated resource for sustaining access to the dataset after deposition. A system for dataset tracking and acknowledgement along with inclusion of metadata and provenance is needed. Such a system would give researchers a rich resource to evaluate for extant datasets BEFORE starting experiments of their own, therefore avoiding duplication of efforts and wasted research resources (money and time). (#43)*

Issue Six: Support Needs

This issue targeted the role of NIH in providing resources to support the needs of the extramural community.   Respondents stated that NIH would provide this support through workforce growth or funding and development opportunities.

Analytical and Computational Workforce Growth

Respondents addressed ways in which guidelines, training, and education could meet recent growth in the analytical and computational workforce. Suggestions spanned four topics:

- Need for trained specialists

  Many respondents commented on the lack of biostatisticians and bioinformaticians. Suggestions to increase the workforce included training individuals in data collection, formatting, algorithms, design, programming, and integration, as well as to make the career more attractive.

- Undervaluing of current professionals

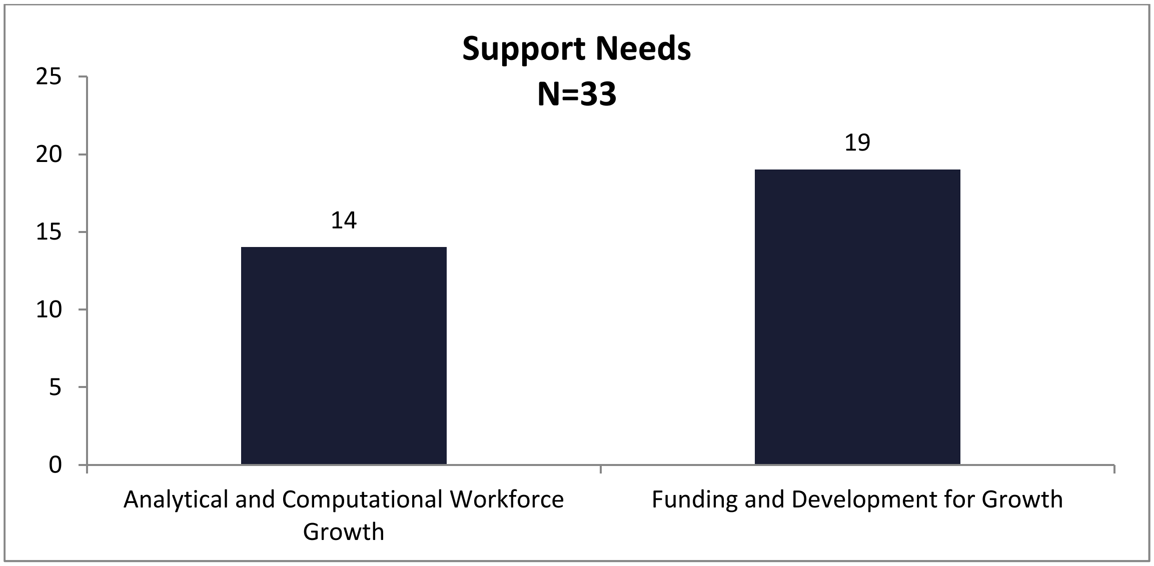
  Another point made by respondents was the undervaluing of workers: "professionals supporting data and the infrastructure to make that data available need to be recognized and suitably supported." (#17)

- Development of training programs

  To support an increase of trained individuals in the data information systems workforce, curriculum development will play a major role and should include approaches to data annotation and storage.

- Establishment of data management tools

  Respondents shared their need for help in managing duties and resources; they believed that new management tools would be beneficial in this regard.



Funding and Development for Growth

Comments included the desire both for new programs that support technological developments and additional grants for methodologies and tools to maintain evolving software systems. One respondent wanted tools developed to quickly search and access relevant data. Another felt that tools were available but their values were unknown; therefore, standards to measure the value of tools needed to be developed. In regard to developing new methodology and tools for software efforts, respondents argued for increased funding from NIH. One commenter articulated this response more fully, concentrating on the fact that currently no one has taken on the responsibility of incurring the cost.

> *You have identified issues related to these questions, but the reality is that, at present, no funding agency has the responsibility and resources to do the very real, detailed work needed to create an agreed common physical and software infrastructure for practical long-term management and archiving of the data flows we are now seeing, much less the data flows that are coming soon. (#25)*

Other concerns that arose were the development of lab notebook software, filling of missing repository gaps, and international cooperation.

SECTION TWO: PRIORITY ISSUES

Respondents generally recognized the challenges inherent with managing large datasets. While it was rare for respondents to rank the order of the issues and sub-issues they identified as priorities, some provided a short paragraph or two identifying the issues they felt were most important.

To give a perspective on how many people identified which issues and sub-issues were a priority, we have presented the priority data from the individual perspective (as opposed to code application frequencies, which represent the total number of comments that received a particular code). Of the 50 respondents who provided feedback to this RFI, 36 (72%) identified at least one priority sub-issue.

Priority of Issues

The distribution of the top three issues based on priority criteria matches the distribution of the top three issues found in the overall comment analysis: Scope of Challenges, Standards Development, and Data Accessibility. However, in the priority analysis, the final three issues were Incentives for Data Sharing, Support Needs, and Secondary / Future Use of Data.

Order of Priority by Issue	Number of Respondents (N=36)

Order of Priority by Issue	Number of Respondents (N=36)
Scope of Challenges	24
Standards Development	24
Data Accessibility	16
Incentives for Data Sharing	14
Support Needs	12
Secondary / Future Use of Data	7

Priority of Sub-Issues

A summary of the top ten sub-issues is provided below for overall respondents and self-reported affiliates; a complete list of prioritized sub-issues is provided in Appendix C. Priority order was established based on the total number of respondents that expressed priority for each sub-issue.

Priority of Sub-Issues: Overall

Of the sub-issues, the greatest single priority was placed on Collaborative / Community-Led Standards, followed equally by Central Repository of Research Data, and Academic Royalties for Data Sharing. The sub-issues rounding out the top ten are shown in the table below.

Issue	Sub-Issue	N*	Priority
Standards Development	Collaborative/Community-based Standards	10	1
Data Accessibility	Central Repository of Research Data	9	2
Incentives for Data Sharing	Academic Royalties for Data Sharing	9	3

Issue	Sub-Issue	N*	Priority
Scope of Challenges/Issues	Feasibility of Concrete Recommendations for NIH	8	4
Standards Development	Metadata Quality Control	8	5
Support Needs	Analytical and Computational Workforce Growth	6	6
Support Needs	Funding and Development for Growth	6	7
Scope of Challenges/Issues	Challenges/Issues Faced	5	8
Scope of Challenges/Issues	Unrealized Research Benefit	5	9
Incentives for Data Sharing	Acknowledging the Use of Data	5	10

*N=Number of Respondents

Priority of Sub-Issues: Self

Those who reported from their own individual perspectives expressed greatest priority for Collaborative/Community-based Standards and "Academic Royalties" for Data Sharing. Metadata Quality Control, Central Repositories for Research Data, and Feasibility of Concrete Recommendation for NIH complete the top five priorities for individuals.

Issue	Sub-Issue	N*	Priority
Standards Development	Collaborative/Community-based Standards	7	1
Incentives for Data Sharing	"Academic Royalties" for Data Sharing	7	2
Standards Development	Metadata Quality Control	6	3
Data Accessibility	Central Repository of Research Data	6	4

Issue	Sub-Issue	N*	Priority
Scope of Challenges/Issues	Feasibility of Concrete Recommendations for NIH	5	5
Scope of Challenges/Issues	Challenges/Issues Faced	4	6
Incentives for Data Sharing	Acknowledging the Use of Data	4	7
Support Needs	Funding and Development for Growth	4	8
Support Needs	Analytical and Computational Workforce Growth	3	9
Data Accessibility	Investigator Authentication Procedures	3	10

*N=Number of Respondents

Individuals who provided feedback from an organizational perspective offered limited comments with regard to prioritizing the issues and, therefore, the analyzed priorities are not presented.

SECTION THREE: RESPONDENT RECOMMENDATIONS FOR NIH

Our analysis for this section involved two approaches. The first approach was to compare code frequency distributions across the entire dataset with the subset of data created to represent specific ideas for NIH. The second approach involved qualitative analysis of the subset of data to identify common themes across respondent suggestions.

Code Frequency Comparison

Comparing the distribution of issues between the total dataset and the subset of NIH Responsibility revealed many differences. The order of frequency distribution for most of the issues differed except for the least identified issue (Secondary/Future Use of Data). The table below illustrates the overall order of frequencies for both subsets.

NIH Responsibility Subset	Total Dataset

NIH Responsibility Subset	Total Dataset
Support Needs	Scope of Challenges/Issues
Incentives for Data Sharing	Standards Development
Scope of Challenges/Issues	Data Accessibility
Data Accessibility	Support Needs
Standards Development	Incentives for Data Sharing
Secondary/Future Use of Data	Secondary/Future Use of Data

Qualitative Themes

A number of specific suggestions were presented throughout Section One; in this section, we analyze the subset of NIH Responsibility data to present a more holistic view of respondent recommendations. The recommendations were at the issue and sub-issue level. The table below shows the number of codes marked NIH responsibility according to issues and sub-issues.

Issues and Sub-Issues	N*
Scope of Challenges/Issues	20
Research Information Lifecycle	1
Challenges/Issues Faced	2
Tractability with Current Technology	0
Unrealized Research Benefit	2
Feasibility of Concrete Recommendations for NIH	15
Standards Development	18
Reduction of Storing Redundant Data	1
Standards according to Data Type	4
Metadata Quality Control^	4
Collaborative/Community-based Standards^	7
Develop Guidelines^	2
Secondary/Future Use of Data	8

Issues and Sub-Issues	N*
Improved Data Access Requests	3
Legal and Ethical Considerations	2
Patient Consent Procedures	3
Data Accessibility	18
Central Repository of Research Data	11
Models and Technical Solutions	2
Investigator Authentication Procedures	5
Incentives for Data Sharing	23
Acknowledging the Use of Data	7
"Academic Royalties" for Data Sharing	16
Support Needs	36
Analytical and Computational Workforce Growth	14
Funding and Development for Growth	19

*N=Number of codes marked NIH responsibility

Support Needs

To adequately address data and informatics challenges, respondents made several suggestions that NIH support not only an infrastructure, but also output and utilization of data needs, such as enhanced organization, personnel development, and increased funding for tool maintenance.

Increase Funding to Develop and Maintain Data Applications

Comments included suggestions for investing in the development and maintenance of tools. For example, there was interest in new projects that created data repositories. One respondent claimed that NIH supported certain sub-types of data more than others (i.e., genomic/transcription over biological/biochemical). Similarly, others requested less emphasis on translational goals and more on basic science. The creation of an up-to-date directory describing databases and tool development projects was also recommended.

Specific comments were to increase funding for tool development in the areas of technology transfer, data capture, standards compliance, and data integration. Software for lab notebooks that would be freely accessible and available from NIH was suggested (often-cited tasks that the software must accomplish included assisting with documenting lab work, allowing links to figures, storing raw data in several file formats, and providing storage locations). Referring to the issue of exponential data growth, one commenter requested that NIH not only invest in hardware, but also invest in algorithms and techniques. With the emergence of these new tools, respondents asked for curriculum materials to develop improved understanding of data annotations and storage.

Increase Professionals in the Field/Workforce Growth

Respondents urged NIH to fund projects and programs that placed more bioinformaticians and statisticians in the workforce. Some respondents requested resources for hiring and retaining technically-trained personnel. One comment suggested fellowships that would develop the skills of the workforce that already existed in most institutions, such as librarians.

> *In partnership with computational bio-informaticists and statisticians, librarians undertaking additional training opportunities can address data stewardship principles and practices including: data archival methods; metadata creation and usage; and awareness of storage, statistical analysis, archives and other available resources as part of a data stewardship training curriculum. (#29)*

While respondents called for an increase in funding to ensure growth in the workforce, the comments emphasized the need to “fund people and not projects.”

Respondents also suggested support for data curation as a profession, stating that NIH should improve recognition programs for data curation and create alternative career paths. Respondents recommended that NIH stipulate guidelines for data curator positions to be filled by highly-qualified individuals with advanced degrees; these individuals would annotate datasets for high levels of accuracy and ensure data integrity.

Some respondents suggested NIH develop new training programs for data management and sharing. These programs would emphasize coherent strategies for the analysis of large datasets. One respondent suggested the need for new training programs in health agencies to prepare the next generation of investigators and public health staff with the mindset for data sharing.

Data Sharing

The second most cited area in which respondents made recommendations to NIH was in Data Sharing. Many comments suggested the need to make biomedical data more readily available and to address issues regarding the need for incentives to support data infrastructure.

Make Biomedical Data Available

Respondents suggested that NIH develop guidelines and standards. Specifically, they asked for guidelines around comprehensive patient consent procedures that would make data available. Respondents felt that the challenge lies in answering the question of who owns the data: researcher/scientist, institution, or government.

> *Funders may feel that taxpayers supported the creation of study-specific data, so that NIH would own the data on behalf of taxpayers. However, in cases where researchers work at health care organizations and build datasets based on the organizations' data, the parent company may reasonably argue that they own the data, and that NIH's contribution was a modest value-add. Health care organizations will have a need to shelter their data to protect their business from competition and from reputational risk and a duty to safeguard the confidentiality of their patients. Scientific investigators also have a stake in the ownership of the research data; since they invested their knowledge – including knowledge acquired outside of the study-specific work. (#23)*

One suggestion was to provide a place in the grant application to list shared data; another suggestion was that each researcher's data sharing record be evaluated in peer review. As described in Section One, one respondent suggested a sixth scored review criterion on the data sharing track record.

Respondents indicated the importance of engaging in shared policies and guidelines to determine best practices and systems for data citation. To address this need, respondents recommended accelerating the production of guidelines for researchers to ensure best practices. In line with this suggestion, one respondent was concerned with the ethical compromises inherent when guidelines are not readily available or accessible and suggested that NIH endorse or provide a set uniform data use agreement (DUA).

Incentives to Support Infrastructure

Many respondents called for improved incentives that would help facilitate data sharing by establishing data generators or intramural infrastructure. One respondent thought NIH should promote data sharing; otherwise, investigators may see it as a thankless overhead activity.

> *Without such incentives, researchers may see data sharing as an overhead activity, requiring time and effort with little reward. This perception will not encourage development of high-quality metadata...Better incentives for sharing data, standards for describing data, and clarity of policies for secondary/future use of data are all vitally important to making contribution and reuse of high-quality data a more achievable goal. (#28)*

NIH was encouraged to promote investigator compliance by rewarding and recognizing datasets as research output, thereby ensuring that data sources are included in applications for funding.

Furthermore, some respondents believed that NIH had become less rigid in enforcing data sharing. One respondent recommended that NIH "require" data sharing, not just ask or suggest that it occur.

> *The current policies in NIH RFAs and PAs only ask that applications describe plans to share data and software products created by their publicly funded research. They do not (at least in the cases I have seen) actually require funded projects to share their data. It would not seem unreasonable for NIH to require that projects share data in standard (or at least commonly-accepted formats), especially if those formats were developed thanks to NIH funding in the first place. (#36)*

Standards Development and Data Accessibility

Respondents thought that standardization and data housing would be most efficiently handled by a central source (which was often suggested as NIH). One respondent recommended the development of consortia for each subject, allowing researchers to make decisions specific to the discipline. The Beta Cell Biology Consortium was used as an example where selection of cell type, treatment, and antibody is well discussed.

Standards regarding regulatory policies and procedures were recommended in an effort to advance the strong need for archiving data by developing technical solutions and standard templates for data sharing. Others suggested the need for appropriate digital signatures, such as digital object identifiers (DOI).

The issue of consistency was frequently mentioned. Some respondents proposed that repository requirements should establish minimal service criteria to be met by repositories as a method of unifying and preserving the data repository. Those who identified this issue as important suggested support to navigate and maintain the repositories since there are many repositories available for different types of data.

> *The researcher must know about all the different repositories in order to search for what they need, and the number of such repositories is only growing…making public data more visible, navigable, and useful can be accomplished by financing repository aggregators. Financing more projects and tools that promote domain specific databases to push and pull their data to the aggregators and to the Semantic Web will support data sharing. (#49)*

While most respondents believed that there were several repositories that met their needs, a few believed that some new repositories should be identified. One respondent suggested that a new database for RNAi data should be widely accepted. Another respondent recommended an alternative repository using the Supplementary Information (SI) section of research articles, thereby allowing publishers to commit to storing the data themselves.

Feasibility of Recommendations to NIH (Scope of Challenges/Issues)

One respondent felt that NIH had fallen behind in data informatics, recommending that NIH move to the cutting edge of the field to catch up with current developments. For example, The Cancer Genome Atlas was not able to release data online until the demise of CaBIG in 2011.

Respondents highlighted the dual problems of large amounts of data produced in many sites and the inadequacy of most systems to handle large volumes. Suggestions for solving these problems were organizing data, picking appropriate representatives, developing algorithms and managing interfaces, and designing systems that maximize transfer between disc and memory.

Others articulated the need for a site to host linked data, stating that their current systems compose a series of “patchworks of exception.”

Collaborative/Community-based Standards (Standards Development)

On the mind of several respondents was the need for NIH to facilitate collaborations for data and informatics topics. Increased collaboration and coordination were consistently identified as important for improving data sharing and data management issues. Respondents called for collaboration on a variety of levels and emphasized the involvement of everyone, including agencies, institutions, and the U.S. and international scientific communities, in a discussion about the development of data standards.

Collaboration with NIH, Federal Agencies, and Institutions

In addition to NIH, respondents suggested partnerships with sister agencies and grantee institutions to develop approaches for supporting mid-level IT infrastructure as a way to meet agency needs and, in return, avoid inflicting operating inefficiencies on grantee institutions. One respondent highlighted ongoing collaborations to improve the grant making process by the Research Business Models Working Group of the National Science and Technology Council and

Federal Demonstration Partnership. This suggestion was for NIH to work in conjunction with working groups in order to facilitate progress towards more developed and maintained IT infrastructure.

Respondents urged NIH to develop data standards to assist investigators who are less familiar in one research area in understanding and using datasets from another research area, thereby leveraging previously funded resources. To facilitate such standards, the National Library of Medicine was suggested to serve as a model for the extramural community in its reliance on the experience and expertise of its librarians.

> *Librarians can be essential team players, not only in helping to develop standards and ontologies, but also in making their research communities aware of the resources available through NIH and other research groups and agencies. (#29)*

The important question, "Who owns the dataset?," emerged from a few commenters. The respondents recommended that NIH, in consultation with researchers, clinicians, and patients, address this issue, giving sufficient weight to the common good.

Community Collaborations

Respondents believed NIH could promote effective coordination of standards by helping to identify problems that standards will solve. Creating initiatives on sharing through use of community reporting standards would encourage good data stewardship. Repeatedly, respondents suggested that NIH support community-initiated efforts for standardized data representation. One respondent used the example of The Gene Ontology to support the notion of collaboration.

> *The Gene Ontology was developed by multiple model organism database developers who saw the benefits of collaborating on a common standard. Its wide adoption demonstrates the success of data standards developed collaboratively by researchers trying to solve practical problems. (#26)*

One respondent recommended that NIH require some minimum amount of diversity analysis and reporting on data collected under diversity sampling requirements.

> *It is nonsensical that NIH requires, and goes to great pains to enforce, diversity in sampling; yet has no coincident requirement to conduct and report on differential validities due to race, gender, age, etc. Consequently, very little of this sort of research is ever conducted despite having sufficient data. (#3)*

Global Collaborations

Respondents believed NIH could use its considerable influence to promote and improve collaboration around the world. Respondents suggested that NIH coordinate support between the U.S., Europe, and Asia where uniform standards are often needed. One suggestion was for NIH to work with other funders, such as The Welcome Trust or Biotechnology and Biological Sciences Research Council (BBSRC), to establish consistent data policies where regional laws permit. The ultimate goal would be to make data interoperable, regardless of geographic origin or funding source.

Appendix

A. FULL CODING SCHEME: DESCRIPTION OF ISSUES AND SUB-ISSUES

Issue 1: Scope of the Challenges/Issues

Issue: Understanding the challenges and issues regarding the management, integration, and analysis of large biomedical datasets

Sub-Issue	Description
Research Information Lifecycle	Strategies for managing research information/data from the time it is created until it is terminated
Challenges/Issues Faced	Challenges and issues presented by use of datasets in the biomedical field
Tractability with Current Technology	Ability to manage and control current technology
Unrealized Research Benefit	Acknowledgement that datasets, data sharing, and administrative data have many research benefits that are not being explored
Feasibility of Concrete Recommendations for NIH	Recommendations for NIH action regarding biomedical data

Issue 2: Standards Development

Issue: The development of data standards

Sub-Issue	Description
Reduction of Redundant Data Storage	Identification of data standards, reference sets, and algorithms in order to reduce the storage of redundant data
Standards according to Data Type	Identification and differentiation of data sharing standards according to data type (e.g., phenotype, molecular profiling, imaging, raw versus derived, etc.)

Sub-Issue	Description
*Metadata Quality Control^*	Development of standardized metadata (uniform descriptions, indexing categories, semantics, ontologies, formats, etc.) to organize data from different sources and improve data quality control
*Collaborative/Community-based Standards^*	Development of processes that involves community-led collaborative efforts
*General Guidelines^*	Development of guidelines for data management, access and sharing, and training for compliance

^Data-driven issues

Issue 3: Secondary/Future Use of Data

Issue: The facilitation of the use of data through secondary sources or data presumed for future use

Sub-Issue	Description
Improved Data Access Requests	Development of procedures and policies that will improve the efficiency of the request for access to data (e.g., guidelines for IRB)
Legal and Ethical Considerations	Development of evolving guidelines and regulations for legal and ethical considerations
Patient Consent Procedures	Development of comprehensive procedures and policies regarding patient consent to share their information

Issue 4: Data Accessibility

Issue: The ability to access data

Sub-Issue	Description
Central Repository of Research Data	Development of a central repository of research data appendices (e.g., developing links to PubMed publications and RePorter project record)
Models and Technical Solutions	Development models and technical solutions from multiple heterogeneous data sources

Sub-Issue	Description
Investigator Authentication Procedures	Development of comprehensive procedures that authenticate the data provided are the investigator's own work

Issue 5: Incentives for Sharing Data

Issue: The need to have incentives in order to encourage/influence others to participate in data sharing

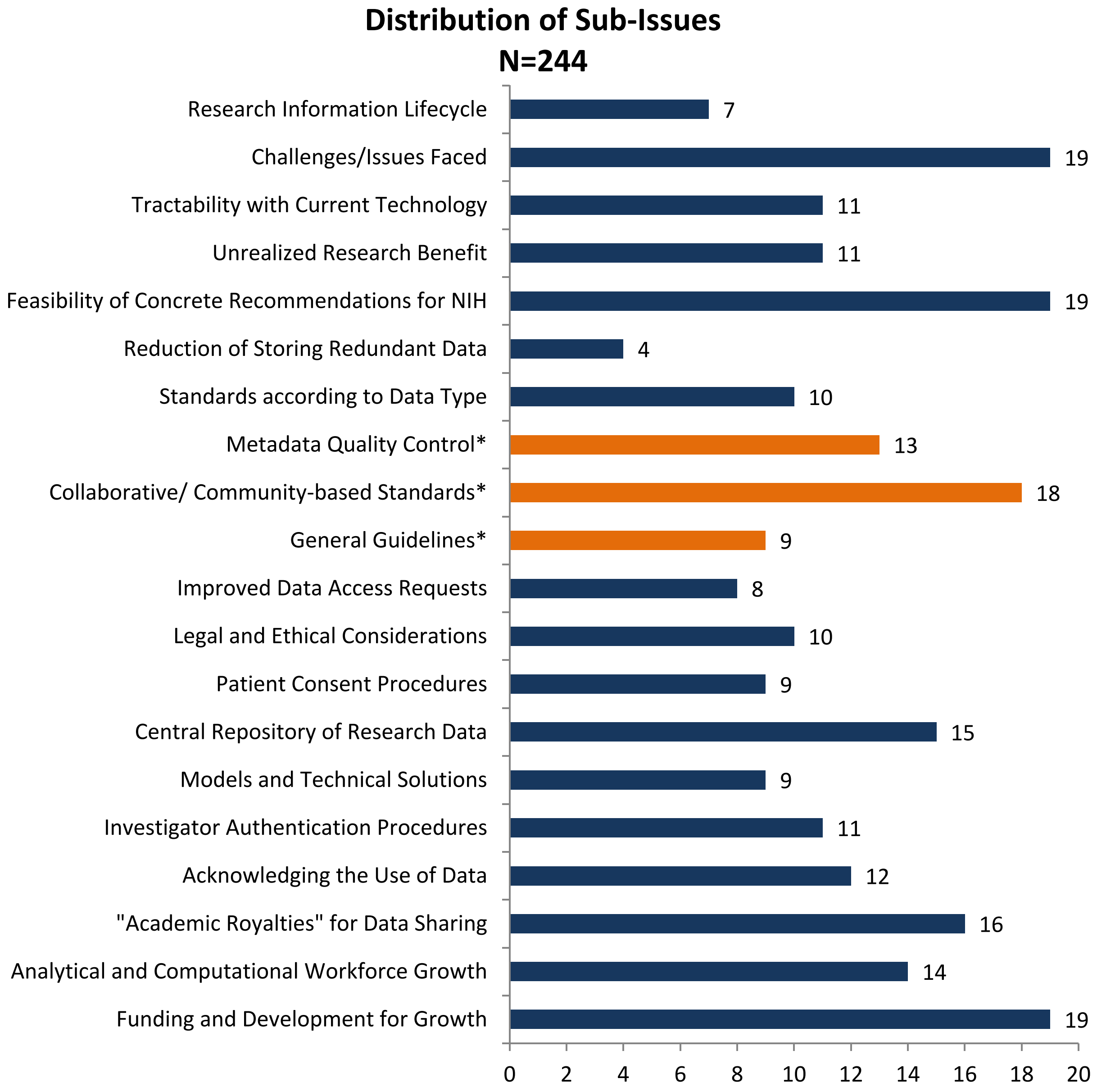
Sub-Issue	Description
Acknowledging the Use of Data	Development of standards/policies for acknowledging the use of data in publications
"Academic Royalties" for Data Sharing	Creation of policies for providing "academic royalties" for data sharing (e.g., special consideration during grand review)

Issue 6: Support Needs

Issue: The role of NIH to provide supportive needs to the extramural community

Sub-Issue	Description
Analytical and Computational Workforce Growth	Provision of guidelines, training, and education to facilitate growth in the analytical and computation workforce
Funding and Development for Growth	Provision of funding and development for tools, maintenance and support, and algorithms

B. SUMMARY OF FREQUENCY DISTRIBUTION ACROSS ALL SUB-ISSUES



C. ORDER OF PRIORITY: ALL SUB-ISSUES

Order of Priority: Overall (N=36)

Issue	Sub-Issue	N*	Priority
Standards Development	Collaborative/Community-based Standards	10	1
Data Accessibility	Central Repository of Research Data	9	2
Incentives for Data Sharing	"Academic Royalties" for Data Sharing	9	3
Scope of Challenges/Issues	Feasibility of Concrete Recommendations for NIH	8	4
Standards Development	Metadata Quality Control	8	5
Support Needs	Analytical and Computational Workforce Growth	6	6
Support Needs	Funding and Development for Growth	6	7
Scope of Challenges/Issues	Challenges/Issues Faced	5	8
Scope of Challenges/Issues	Unrealized Research Benefit	5	9
Incentives for Data Sharing	Acknowledging the Use of Data	5	10
Standards Development	General Guidelines	4	11
Secondary/Future Use of Data	Improved Data Access Requests	4	12

Issue	Sub-Issue	N*	Priority
Data Accessibility	Investigator Authentication Procedures	4	13
Scope of Challenges/Issues	Research Information Lifecycle	3	14
Scope of Challenges/Issues	Tractability with Current Technology	3	15
Secondary/Future Use of Data	Legal and Ethical Considerations	3	16
Data Accessibility	Models and Technical Solutions	3	17
Standards Development	Reduction of Storing Redundant Data	1	18
Standards Development	Standards according to Data Type	1	19
Secondary/Future Use of Data	Patient Consent Procedures	0	20

*N=Number of Respondents

Order of Priority: Self (N=26)

Issue	Sub-Issue	N*	Priorit y
Standards Development	Collaborative/Community-based Standards	7	1
Incentives for Data Sharing	"Academic Royalties" for Data Sharing	7	2
Standards Development	Metadata Quality Control	6	3
Data Accessibility	Central Repository of Research Data	6	4
Scope of Challenges/Issues	Feasibility of Concrete Recommendations for NIH	5	5
Scope of Challenges/Issues	Challenges/Issues Faced	4	6
Incentives for Data Sharing	Acknowledging the Use of Data	4	7
Support Needs	Funding and Development for Growth	4	8
Support Needs	Analytical and Computational Workforce Growth	3	9
Data Accessibility	Investigator Authentication Procedures	3	10
Scope of Challenges/Issues	Tractability with Current Technology	2	11
Scope of Challenges/Issues	Unrealized Research Benefit	2	12
Standards Development	General Guidelines	2	13

Issue	Sub-Issue	N*	Priority
Secondary/Future Data Uses	Improved Data Access Requests	2	14
Data Accessibility	Models and Technical Solutions	2	15
Scope of Challenges/Issues	Research Information Lifecycle	1	16
Standards Development	Standards according to Data Type	1	17
Secondary/Future Data Uses	Legal and Ethical Considerations	1	18
Standards Development	Reduction of Storing Redundant Data	0	19
Secondary/Future Data Uses	Patient Consent Procedures	0	20

*N=Number of Respondents

6.2 National Centers for Biomedical Computing Mid-Course Program Review Report

National Centers for Biomedical Computing
Mid-Course Program Review Report
July 13, 2007

Introduction:

In response to a request from the Roadmap Implementation Coordinating Committee (RICC), an external panel was convened and charged to assess the status and progress of the National Centers for Biomedical Computing Initiative and to provide guidance for the future course of the program. The panel was asked to address 7 questions in their review and to make recommendations for future investments by NIH as part of the ongoing NIH Roadmap Initiative.

For many years, scientists supported by NIH have advanced the frontiers of computing and its methodological infrastructure. This work has provided valuable biomedical computing support for a variety of biomedical research areas and applications to medicine, as well as the informatics infrastructure important to both. The 1999 BISTI report (Botstein, et al. 1999) recognized the critical impact that computational science and infrastructure could make on the advancement of discovery in biomedical science. The four overarching recommendations of that report were: 1) to establish five to 20 National Programs of Excellence in Biomedical Computing, 2) to develop principles and best practices for the storage, curation, analysis and retrieval of information, 3) to support the development and adoption of software tools for biomedical computing and 4) to foster a scalable national computer infrastructure. The investment by NIH in the establishment of 7 National Centers for Biomedical Computing directly addresses the first and third recommendations made in the BISTI report.

The planning process for a Roadmap for Medical Research in the 21st Century (http://nihroadmap.nih.gov) also recognized the importance of developing sustainable infrastructure that spans multiple NIH Institutes and Centers for advancing biomedical computing. The National Centers for Biomedical Computing are poised to address several of the Roadmap themes: “New Pathways for Discovery” as part of its focus on new tools and methods, “Research Teams of the Future”, developing sites where training of cross disciplinary researchers takes place and “Re-engineering the Clinical Research Enterprise”, where NCBC advances in informatics and biomedical computing provide critical support to that research arena as well as computational tools that facilitate the delivery of its findings to medical environments.

This focus on support for biomedical computing is not new at NIH. For over four decades, the NIH has supported research and development (R&D) on mathematical and computational methods and systems crucial to the advancement of biomedical research. The panel was concerned to learn that there have been previous extramural programs at NIH to support biomedical computing centers that were subsequently abandoned. Thus, in the past, the NIH has failed to develop stable administrative structures at NIH to support this critical research area. It is of paramount importance that the NIH recognizes the massive scale of computational needs anticipated in the future of biomedical research and that these NCBCs, though necessary, are not sufficient. The panel sees sustained investment in these seven NCBCs as only the beginning of the investment required for the creation of a stable computational platform to sustain biomedical research.
The breadth of the Project Team for the NCBC program is very encouraging, as are the large number of

Institutes and Centers represented in the Bioinformatics and Computational Biology Implementation Group. This must continue as the NCBC program and future programs related to it evolve, – not only for the sake of stability, but also to ensure that the biomedical computing centers have a comprehensive view of challenging R&D opportunities available to them and that they and other computational scientists are welcomed at frontiers being pursued by many of the ICs.

Current funding provided to the NCBCs does not appear to permit substantial investments of effort outside of their basic R&D missions, although a number of very worthwhile cross-center activities are in place. The panel recommends the consideration of increasing the budget for the entire program, to facilitate more interactions between the ICs and NCBCs, and also to increase the impact of education and outreach programs.

The panel enthusiastically endorses the investment NIH has made in these critical areas through the NCBC program. The panel believes that these 7 centers, although young, are meeting the challenge of developing a national network of research centers in biomedical computing. They are effectively engaging in multidisciplinary team-based research, developing an extensive collection of useful software tools, providing training opportunities and promoting biomedical computing as a discipline through education and outreach to the community.

Charge to the panel:
"In response to a request from the Roadmap Implementation Coordinating Committee (RICC), the NCBC initiative will undergo a mid-course program review on June 11, 2007. An external panel will be convened to assess the status and progress of the NCBC initiative and to provide guidance for the future course of the program. The members of the review panel have been selected for their expertise in the diverse scientific areas impacted by the NCBCs and for their ability to provide objective input and advice. The chair of the review panel will responsible for writing a report summarizing the views and recommendations of the panel. The report will be sent to the Office of Portfolio Analysis and Strategic Initiatives (OPASI) and forwarded to the RICC. The RICC is scheduled to review the NCBC initiative on August 14, 2007."
"*The review panel will be asked to consider the following questions:*
1) To what extent does the vision and direction of the NCBC initiative promote biomedical computing?
2) In what ways has the NCBC initiative advanced biomedical computing?
3) Are the NCBCs interfacing appropriately?
4) What new collaborations have been formed through the NCBC initiative?
5) What new training opportunities have the centers provided?
6) What changes could make the program more effective in the future?
7) What lessons have been learned from the NCBC initiative that can guide future NIH efforts in biomedical computing?"

Executive Summary:
The panel concurred that a long-term investment in biomedical computing by the NIH is critically important to addressing the health care needs of the country. The panel recommends the following actions to ensure the success of this important effort.
1) Continue the support of biomedical computing as a key part of the NIH research portfolio over the long term. Computational biology, theoretical research and the development of robust software tools are critical to the understanding of biological processes and disease.
2) Begin developing a process to sustain and expand this effort now to anticipate support beyond the 10 year Roadmap funding horizon. The panel is concerned that the viability of this program and of biomedical

computing in general depends on the relatively unstructured cooperative interactions of different NIH Institutes and Centers.
3) Focus research within and across the NCBC Centers on ambitious problems that other programs are unlikely to support, such as the development of cheaper, safer drugs, or new methods for multi-scale modeling of biological processes. Consider partnership with industry, to address difficult problems of national importance, by taking advantage of longer more stable funding periods not possible within the biotech industry. Coordinate tool development with industry, since this is where the tools may have their biggest impact.
4) Continue to support the model of multidisciplinary, team based, collaborative research within the NCBCs. Extend the reach of the individual centers to collaborators outside the centers to increase the impact of the Centers on the community. Do not require interaction between NCBC centers where there is no obvious programmatic advantage. Continue the All Hands Meeting as an effective means for sharing best practices among the Centers and for fostering high impact Center-wide activities.
5) Develop an additional approach beyond the R01 and R21 collaborative grant program for developing and supporting collaborations with the Centers. The current peer-review system imposes delays in getting projects started and creates an additional administrative burden on the Centers to provide support to potential collaborators engaged in proposal submission. Support the NCBC Project Team in developing and implementing alternative approaches, such as streamlining the review process or providing funds for exploratory research, data collection or software design.
6) Develop a process to assess the impact of the software tools developed by the centers. Develop a simple assessment instrument to gauge how the software tools are advancing research and achieving widespread use within the community.
7) A focus on educating the next generation of computational scientists is critical to the success of biomedical computing as a discipline integrated within biomedical research. Continue to support the NCBCs and other programs in training multi-disciplinary researchers through collaborative research and outreach to the community. Leverage the efforts of the NCBCs and expand the educational programs designed to foster the education and training of computational biologists.

Answers to Questions:
The review panel was asked to address the following set of questions in their report. The panel's responses are based on their review of materials provided by program staff, additional material provided by the NCBC PIs, information provided by program staff and discussions within the panel.

1) To what extent does the vision and direction of the NCBC initiative promote biomedical computing?
The establishment of the seven NCBCs is an excellent start to what the panel hopes will be a long-term investment in biomedical computing. NIH has been less consistent than other federal agencies in recognizing the power of computing and promoting its multiple roles for advancing the biomedical sciences. This very visible program effectively reinforces the importance of biomedical computing to the research community. Moreover, by providing a longer planning horizon than is available in individual grants, or in the biotech industry, the NCBCs can propose projects that could not be accomplished otherwise. This program has also encouraged significant matching funds, and its visibility has helped the NCBCs to recruit and develop talent.

2) In what ways has the NCBC initiative advanced biomedical computing?
Despite the short time that these centers have been in place, many success stories are already evident. The NCBCs have developed widely available new software and web-based systems, created visibility for the discipline, and developed much-needed training programs. They have effectively paired experimentalists with computational biologists, and involved computer scientists and engineers in problems of biomedical interest.

They have also generated a large number of relevant R01 and R21 collaborations. Besides their individual achievements, it is noteworthy that in such a short time, the NCBCs have collaborated as a whole on the categorization of biomedical ontologies by their degrees of acceptance in research communities, developing software yellow pages and have begun to coordinate collaborative activities that center on Driving Biological Projects (DBPs).

3) Are the NCBCs interfacing appropriately?

The panel concurred that there is ample evidence of a considerable amount of appropriate trans-NCBC activity. For example, the July 2006 NIH Roadmap NCBC All Hands Meeting was particularly useful in this regard. These meetings should continue to be held and continue to include successful components such as the Building Bridges Compendium and Dissemination Events.

It is apparent that many productive interactions grew out of the All Hands Meeting, including the Software and Data Integration Working Group (SDIWG). The charter of the SDIWG is to promote software interoperability and data exchange, and to bring the collective knowledge and practices across the centers to wide publication. The SDIWG appears to have tapped a wellspring of endogenous enthusiasm in the Centers and has moved forward with leadership from within the centers to conduct regular electronic (and in some cases face-to-face) group conferencing to coordinate and direct the activities.

Three domains of SDIWG activity include: the Software Yellow Pages, Categorization of Scientific Ontologies, and Driving Biological Projects and Impact Working Group.

- The Yellow Pages project. Led by Ivo Dinov (CCB) and Daniel Rubin (NCBO) this project includes a NCBC iTools Prototype that supports a visualization interface for browsing and query of available NCBC tools.
- The Driving Biological Project Interactions, led by Andrea Califano (MAGNet) and Brian Athey (NCIBI), focuses on determining the research community needs for tools, data and methodologies for the analysis of cellular networks, with a focus on their use in complex trait and biological process analysis in current and future Driving Biological Projects. Currently, this activity is represented in a searchable graphical Interactome of potential DBP interactions among NCBCs.
- The Scientific Ontologies Group led by: Zak Kohane (i2b2), Suzi Lewis and Mark Musen (NCBO) aims to create a succinct categorization of available ontologies and terminologies. As a result, the particularly useful contribution of this effort has been the evaluation and categorization of existing biological ontologies into three groups, as (1) Fully Endorsed, (2) Promising and used with some reservations, or (3) Not quite ready for use, underdevelopment and for use under protest. An interactive table of these results is available at http://www.berkeleybop.org/sowg/table.cgi.

These activities show an appropriate focus on the tool development mission of the NCBCs. It is also encouraging that the ongoing interactions make programmatic sense in creating technical and biological synergies. *The panel recommends encouraging the NCBCs to continue to apply this metric and not develop forced interactions and collaborations that do not make programmatic sense. At the same time, the panel recommends encouraging the NCBC Project Team to broaden their search for additional potentially synergistic interactions outside the scope of the NCBCs where it makes programmatic sense, for example with P41s and with other agencies.*

4) What new collaborations have been formed through the NCBC initiative?

All seven of the NCBC centers have developed new collaborations as a result of the NCBC initiative. Those collaborations include interactions within the individual centers, among the different centers, and include a wide array of collaborations with new entities outside of the centers. (Specific examples are well detailed in the annual progress reports and summary documents provided by the Center Directors.) Within individual centers, collaborations have been forged across the individual core components, seminar series have been established and nascent links between biological, bio-informational, and computational components have been expanded and solidified, even beyond the activities proposed in the original applications.

• All NCBCs have cooperatively engaged in scientific and technical discussions of common interests under the auspices of the NIH Software and Data Integration Working Group (SDIWG) as described above. Other examples include the supplementary postdoctoral opportunity that helps bridge across centers, commonly attended conferences such as the DREAM (Dialogue on Reverse Engineering Assessment Methods Workshop: http://www.iscb.org/events/event_data.php?454), and examples of associations that have been developed with other NIH sponsored networks (e.g., caBIG, BIRN, and CTSA).

• Probably the most diverse set of new collaborations are those with entities from outside the NCBC initiative. Such collaborations have been spawned by the inclusion of DBPs (and the new round of DBPs that are being considered), the development of the R01 and R21 "Collaborating with NCBC" initiatives, and by the increasing visibility/capability of the individual centers. Examples include collaborations with industry, vendors, academia, hospitals, and foreign institutions, international, and healthcare organizations. The lists provided by the centers are truly impressive.

• The workshops, websites, and other dissemination efforts developed by the centers are serving to bring together diverse groups of people who might not otherwise interact. These efforts are catalyzing interactions and are likely to lead to new collaborations.

• The review committee believes that all centers have developed significant collaborative interactions. One useful tool used to build these interactions and to get tools out to the community is the Driving Biological Projects. However, the panel rose the question of what the optimal size/number of DBPs should be since they also place a burden on the other center components; affording additional flexibility may be warranted. The R01 and R21 initiatives help build collaborative activities with investigators "outside" the centers; but the peer review process (first at the NCBC and then the NIH) may unnecessarily delay the start of meritorious projects. *The panel recommends some form of flexible short term funding to jump start new collaborations, and/or funds to hire postdocs who bridge centers. The panel also recommends facilitating interactions between NCBC centers, (where it makes programmatic sense) with other P41s, caBIG, BIRN, Virtual Cell, etc - or other agencies.*

5) What new training opportunities have the centers provided?

Training is central to creating the next generation of multi-disciplinary scientists and to broadening the skills of existing scientists to pursue cross-disciplinary collaborations that are advancing the frontiers of biomedical sciences and their applications today. For recipients already committed to careers in this area, the training is of immediate value. For other students, exposure to major nationally supported centers where such multidisciplinary

research is thriving may become an important element in their choice of careers and in turn, a commitment to the very substantial preparation required to become leaders in this area.

The NCBC centers have provided a variety of training and educational opportunities to members of the centers, affiliated groups, and the broader biomedical scientific community. Most of these are at the postdoctoral level and involve special tutorials, workshops or meetings centered on topics of interest and research strength in a particular center. Training activities are not coordinated across the centers presently, and there is some debate as to whether coordination would be beneficial.

Some examples of training activities are listed briefly below:

• *Pre-college – graduate school training*: The CCB has hosted numerous visits by groups of pre-college students. It has provided both graduate and undergraduate courses for students at UCLA, offering research experience to both. I2b2 hosts a Summer Scholars program for undergraduates across the country, which includes both education and research projects. NCIBI participates in the training of 10 graduate students in the University of Michigan’s Bioinformatics Graduate Program.

• *Postdoctoral training*: Most NCBCs are actively involved in post-doctoral training. Five of the seven NCBCs have close associations with NLM training programs at their universities. Simbios has created the “Simbios Distinguished Post-Doctoral Program.” I2b2's post-doctoral program provides access to its leading investigators and DBP data resources. NCIBI would like to increase their post-doctoral fellowships to 3 years, to aid recruiting. NCBO, which supports two post-doctoral trainees and participates in training some from other programs, urges increasing budgets for training. CCB reports training over a dozen post-doctoral fellows and young investigators, and has a program for visiting scholars.

• *Cross training of established investigators*: Cross-training occurs as scientists collaborate at the Centers, in a variety of meetings, and via other resources. NCIBI produces a year-around weekly seminar series, “Tools and Technologies”, in which NCBO, MAGNet, and others also participate. These also are broadcast live over the Internet via streaming video/audio, and are archived for later use. It also produces web-based interactive training and educational programs. I2b2's “Grand Rounds” seminars, to educate scientists about biomedical computing ingredients for discovery research in academic health-care centers, also are available via streaming video available from i2b2's site. CCB has organized three well-attended (> 100 each) international workshops. Simbios is providing a summer short-course and workshop on the use of some of their software. MAGNet’s second retreat, this April, had about 150 attendees, and it has been recruited by a related network in Europe, ENFIN, to produce a joint conference next year.

6) What changes could make the program more effective in the future?

The panel was impressed with what the NCBCs have accomplished so far on a relatively small amount of funding. Limiting discussion to the constraint we were given of continuing the Centers for an additional five years at the current funding level, several suggestions emerged from the panel’s discussions. These suggestions mirror the recommendations in the executive summary and provide additional detail and rationale for the recommendations.

• Centers should be given sufficient programmatic flexibility to jump-start new projects. For example, a joint post-doctoral program emerged from the All Hands Meeting brainstorm session. While this is a new positive idea *across* centers, there is little doubt that if each center's funds were less constrained and thereby they were given the opportunity to think broadly, many such ideas would emerge and be executed *within* each center.

• More could be done to encourage appropriate collaborations of other non-Center investigators with the Centers, while at the same time avoiding unnecessary collaborations. For example, simplifying the collaborative grant programs for the centers and streamlining the review process may be appropriate. The R01 and R21 programs, while getting strong responsiveness, require a conventional peer-review process that limits the risk-taking needed to quickly jump-start new ideas. It also appears to be particularly burdensome for the centers to "help" to write multiple R01 and R21 submissions most of which will never be funded. Perhaps a mechanism can be established whereby some of the IC funds that would otherwise go toward those R01 and R21 programs could be managed by the NIH NCBC Project Team for the ICs and be assigned to Centers for use in collaboration with others through a streamlined review process. Fundamentally, the development of the NIH NCBC Project Team is a terrific development that can be leveraged to assist the Centers in promoting the use of Center tools and collaborations. Although the Project Team has done a great job of stimulating new projects, they could be encouraged to take a more active role in supporting those projects. The shared ontology database is a good example where the NCBC Project Team stimulated new positive inter-Center work, but where additional support seems appropriate to fully execute and maintain it rather than pressuring the Center Directors to do it without new support.

- Encourage the NCBC Project Team to engage in developing an assessment program for software tools. Just as the Categorization of Ontologies performed by the NCBCs, stimulated by the Project Team, is seen generally as a very positive outcome, an objective method of assessment of the usefulness (effectiveness, ease of use, desirability, efficiency) of new biomedical software tools would be a great contribution. The panel believes the Project Team could develop and lead and/or subcontract a program to develop such an assessment instrument that would account for all the spectra that make a software tool a success (user base, computational efficiency, human-computer interaction, scientific papers/achievements using the software, specific niche it fills and why, etc.).
- Encourage the Project Team to work with the NCBCs to enhance the dissemination of their successful software development efforts. A variety of mechanisms could be used to help investigators adopt and use the software the NCBCs are creating. Websites that are deep enough to help investigators learn to use the software, rather than just download it, would be worth the effort. The NCBCs can lead this effort, since they are the experts in their software, however additional resources should be devoted to this effort would allow the NCBCs to focus on their strengths in building new tools.
- Consider how to leverage the NCBCs to enable training programs in computational biology. The NLM Medical Informatics Training Programs have certainly contributed to this domain and to the ability of the Centers to pursue their research agendas. Perhaps NIGMS, in collaboration with the NLM can establish additional training programs by leveraging the NLM model. These programs could be located outside the NCBCs, ideally linked with them expanding the new joint post-doctoral program.
- Adopt methods to improve the cross-disciplinary collaboration experience, among NCBC PIs, the Project Team, and others, perhaps through social networking approaches. This topic could be discussed at the next All-Hands Meeting.

7) What lessons have been learned from the NCBC initiative that can guide future NIH efforts in biomedical computing?

The panel concurred that the development of a successful NCBC is a very complex and challenging task. Each of the Centers has encountered different obstacles that can slow, or sometime prevent the achievement of the goals of the program. The panel identified several consistent themes that illuminate these challenges.

- Carrying out true interdisciplinary work is hard enough when only two disciplines are involved; it is much more so when trying to achieve this with four or more disciplines (i.e. computer science, mathematics, engineering and biomedicine). NCBC programs will take time to fully develop.
- The concept of “Driving Biological Problems” is outstanding and critical for ensuring focus and utility. The best work is done when experimental biologists and computational researchers are equal partners and there is a tight feedback loop between experimental design and computational analysis.
- Co-location of various disciplines is a big advantage, especially during the initial period of program development. Geographical distribution within a single NCBC can inhibit cross-disciplinary exchange and increases administration complexity, but has the advantage of increasing the range of expertise available to apply to a specific problem. The importance of social networking has been underappreciated and could help ameliorate these issues.
- Interactions among NCBCs, and with other NIH biomedical computing initiatives, may be very important for success but unnatural or artificial interfaces should not be forced just for the sake of interacting. A commitment to interoperability will enhance the success of interactions.
- While it may be sensible in some cases to attempt a broad and diversified portfolio of projects and applications, this approach may dilute efforts and should not be done at the expense of focus.

Increased focus will increase impact.
• Because the NCBCs have such great potential, it is easy for unrealistic expectations and “scope creep” to develop, leading to a serious mismatch between objectives and the resources available to achieve them.
• Partnerships with industry (Pharma, other biosciences, translational medicine and IT) should be encouraged and supported. Open source software combined with the ability to license applications for commercial use may help leverage the work of the Centers and increase their impact.
• Sustainability plans should be formulated now and the NCBC initiative should evolve from a Roadmap activity to one that achieves ongoing and cross-ICD support (with competitive renewal). Partnerships with other Federal agencies (e.g. NSF) should be explored.

Vision and Grand Challenges for the Future of Biomedical Computing
The panel was encouraged to think broadly about a future vision for biomedical computing. The following sections summarize the panel’s thoughts on this important issue.

Historically, some important advances in mathematics have been developed to meet challenges on the frontiers of research in the physical sciences. Today the rapidly advancing frontiers of biology and medicine invite similar opportunities in mathematics, statistics, and computational methodologies. There is a need to evolve productive environments and approaches for inspiring and facilitating cross-disciplinary collaborative research. There is also a need to increase our investments in training scientists who have substantial graduate-level training in both mathematical/computational disciplines and in a biological/medical discipline. Such scientists exist, but we need many more. The NCBCs can contribute in a variety of ways. They can explore and evaluate new approaches to improving cross-disciplinary techniques, research environments, and cultures. They can provide cross-disciplinary research opportunities for students from upper-division through post-doctoral levels.

The existing set of NCBCs is an extremely valuable start toward a maintained focus on biomedical computing at the NIH. While the centers cover a broad range of topics, at present they focus mostly on problems that can be handled with today’s computers running algorithms that are not too far removed from current standard practice. Given the nature of the original NCBC proposal guidelines and the review process, this is not terribly surprising and is an appropriate beginning. Reviewers are naturally cautious at the outset of a new program, and pre-existing NIH funding was required for the candidate Driving Biological Projects. This set of constraints has produced centers that to some extent are plucking the low hanging fruit; that is, they are working mostly in areas that are already well established as part of computational biomedicine, and that have clearly discernible connections to the clinic in the here and now (imaging, genomics, high-throughput data handling, etc.).

From the standpoint of broad-ranging future needs, however, the program must expand to embrace more innovative cutting edge software and hardware development. Biology and medicine still lag far behind other scientific disciplines in the use of large-scale high performance computing (e.g., physics, chemistry, meteorology, climatology). In some circles the perception remains that biology and medicine just “aren’t ready” for large-scale computing. This attitude is probably due to history, ignorance, and the admitted complexities of biomedical research. Even in those areas where large-scale computation is already necessary and established in biology (e.g., quantum mechanical and molecular mechanical simulations) the connection to classical chemistry (as opposed to biology) is strong and the overall scope remains narrow. But “waiting” for

additional areas of biomedicine to be “ready” for large-scale computing will leave the US in a perpetual “catchNCBC up” position compared to other nations, and will dramatically delay breakthrough advances in predictive
personalized health care.

To accelerate innovation, a new mechanism for collaborative interactions should be developed without an obligatory need for traditional, conservative, and lengthy R01/R21 review (as outlined in preceding sections). To make reasonable judgments about what projects are appropriate and productive, recommendations can be obtained from each Center’s External Advisory Committee. Software development components of all interactions should strongly encourage innovation and maintain a long-term outlook on what needs to be achieved for the future, rather than remaining steadfastly centered on short-term applications. Current applications are important, of course, but the strategic outlook of the Centers should include increasing attention to innovative long-term goals beyond the obvious short-term gains.

Grand Challenges in Biomedical Computing

The following are examples of unsolved problems in the biomedical sciences that should be pursued as part of a sustained focus on biomedical computing. None of the current NCBCs are addressing these problems at present and while this does not represent an exhaustive list, these examples illustrate the range of problems that require the type of research infrastructure being developed in the NCBCs.

Quantitative multiscale modeling is one example of an outstanding grand challenge in computational biomedicine. While many existing centers and investigators are already working on different aspects of the problem, no large-scale concerted effort has been created to date. Part of the problem is the need for new mathematical treatments of complex biological systems, and another is the need for new algorithms and approaches to “mesoscale” (cellular to tissue) problems in physiology and medicine. The eventual solution will require sophisticated and long-term software development that couples stochastic and continuous methods applied to problems of very large scale, and therefore will also require careful attention to hardware design and implementation. Despite a petascale computing initiative presently underway at the NSF, no federal agency has yet truly come to grips with the real software development requirements in any discipline. The NIH, in fact, remains conspicuously absent from such initiatives, and even within the NSF funding for petascale applications in biology, is conspicuously absent. One can view the present situation as a problem or an opportunity. To make it the latter the NIH will have to assume a leading role and push the frontier in collaboration with other federal funding agencies, and the NCBCs of the future are an obvious possible venue.

Predictive personalized medicine of the future will encompass routine knowledge of each individual’s genome and proteome, high resolution real-time non-invasive imaging, and individualized modeling and simulation that will encompass molecular to cellular and tissue scales. For example, when a patient presents to the emergency room with an acute onset of heart attack or stroke, treatment decisions in the future will depend on massive amounts of individualized quantitative data coupled to quantitative predictive modeling of the evolving event. Thus the future of personalized interventional health care is similar to what is now taken for granted (but is still under active development) with storm forecasting, which is based on massive real time data collection and continually improving supercomputer models that likely will grow into first-case petascale applications. To accelerate this vision for the future of health care, it is incumbent on the NIH to break through traditional boundaries. New trans-agency funding mechanisms and long-term support must be developed to foster innovative breakthrough thinking in the NCBCs of the future, as well as new additional biomedical computing initiatives that have yet to be envisioned.

For patients with diseases, such as HIV and some kinds of cancer, which require long-term therapy during which there are changes in the patient's basic status (e.g. viability of immune, hematopoietic systems) and the treatment target (e.g. mutations conferring drug resistance), ***individualized optimal therapy*** will require continual modifications in the types and doses of therapies administered, and in their timing. The benefits to patients and the economy could be substantial. To achieve this, sophisticated techniques must be developed for the weighted and possibly model-based integration of the patient's information (e.g. functional, biochemical, imaging). The resulting computational support should be physician-friendly and responsive to suggestions for improvement. NCBCs or other programs with the ability to provide stable support for the long-term crossdisciplinary
R&D that is required here are essential. Recommendations made at the end of the preceding paragraph apply here as well.

Methods for answering complex queries over a continuously updatable semantic web is at the center of a new computational approach to integrating literature searches with methods of experimental science, and is stimulating a new paradigm of "active" computational scientific inquiry. The NCBCs have the potential to accelerate these developments by tying them to modeling, visualization and ontologically-based argumentation, helping researchers pose problems in entirely new ways, checking their conjectures "on the fly" against a synthesis of existing knowledge.
A critical missing component and challenge is how to combine the visual with the logical, simulation (modelbased what-if), and statistical argumentation typical of such scientific reasoning – while also taking into account the ambiguity, risk and uncertainty inherent in many of the arguments. Developing effective tools which suggest new questions automatically from the results of such "computational models of inquiry" is a major challenge for the centers, which may increase the likelihood of "seeding' new ideas and scientific ventures in other bioscience laboratories. The development of such a "biosciences semantic web" will be essential to overcome the current "traditional discipline silos" which are the basis of most knowledge organization, both in the literature and even in the most flexible of current computational search systems.
An even more ambitious challenge is to see if specific experimental designs can be suggested automatically on the basis of a systematic search over current experimental conditions and results that have yielded conflicting interpretations under various model assumptions. This "grand challenge" involves cognitive science and learning in addition to the modeling, simulation, analysis and interpretation typical of current exploratory bioscience.

Working towards developing cheaper, safer drugs. The process of drug discovery, validation, development, testing and submission is long and complex and expensive. However, there are a number of points along the pipeline where better computation, the introduction of modeling, and more effective information retrieval and analysis could have a large impact in decreasing the time and the cost of preclinical and clinical development. Learning from a large body of prior experimental data about how to analyze interpretations of experimental results is a challenge that goes far beyond what current (largely static) ontologies and parameterized sets of models can handle. Overcoming this barrier, for even a restricted class of design problems, could generate automatic suggestions for designs within a "standard protocol" of experimental design and discovery, as is the current practice in drug design. To achieve this goal, the integrated analysis of meso or multi-scale (moleculecell-tissue-organ, organism, population-environment system), and heterogeneous models will be required. However, these methods, once developed, could lead eventually to automatic search and discovery heuristics that could yield "cheap, effective drugs and procedures with a minimal number of side-effects."

Creating comprehensive, integrated computational modeling/statistical/information systems and related databases for major biomedical sectors. For example, the software platforms GENESIS, NEURON, and other computational modeling systems serving research in the neurosciences have evolved over the years into critical tools for this community. They provide information related to neuronal model elements and the software that facilitates linking these elements into a model. For one of these programs, an active users' group has made many contributions over the years, e.g. models and parameters for a variety of channels, compartmental models of various neurons.

Are there other important biomedical sectors that might benefit from, and do not as yet have, programs of this general type? – e.g. oncology, virology/HIV, genetics? When we examine the programs and consider overall computational biology tools that might be applicable to research in those areas, can we identify methodological elements that perhaps should be added and integrated – e.g. the ability to embed stochastic models within a deterministic system?

Establishing a resource for advancing and maintaining software valuable for biomedical research and applications. Some non-commercial software may be of substantial use to various biomedical researchers long after the grant that funded its creation, or its creator, have ceased to exist. As time goes on, the operating systems and other elements for using the software change. If the community agrees that the software is well worth maintaining, and improved in specific ways, can NIH provide a resource for doing this?

Training the next generation of computational scientists: The panel recognized that the greatest grand challenge of all is recruiting and educating future computational scientists. The type of complex inter- and multi-disciplinary research, which is the hallmark of the NCBCs, will require long-term educational preparation by future members of such centers, including strong High School and Undergraduate studies in both the formal and analytical aspects of research methodology, as well as experimental science and/or engineering, and biological knowledge. This is a tall order for most people, but successful researchers in the future (and even now) often require the equivalent of two undergraduate degrees to prepare them for successful graduate studies and post-doctoral work in any of the computational biosciences. With this in mind, *the panel recommends that the NIH suggest to colleges and universities that they encourage such joint programs of study so future researchers can be as well prepared for deploying analytical and computational approaches to bioscience as they are for wet-lab experiments.*

The preparation needed for the discipline of computational biology is very different than that for bioinformatics. The former requires much more depth in mathematical modeling, algorithms, and simulation, and theoretical biology, whereas the latter tends to focus on large-scale data mining analysis and evaluation of experimental biological data. Both are needed in most NCBC projects, in addition to imaging and visualization methods, which tend to come from yet other disciplines: Computer Science, Cognitive Science, or Engineering. The appreciation of the complementary roles these different approaches play in experimental design and its implementation within the context of systems biology and the emerging semantic web of information is a central educational need for all the NCBCs. In addition, as we move increasingly into large-scale epidemiological studies, researchers will also have to appreciate the nuances of population modeling and study design, which presents yet another set of methodological challenges. In some sense the NCBCs could help develop a new “systems biology ontology” which has not yet emerged in most other sciences, which still tend to adhere to a physics-based paradigm of science.

Postdoctoral trainees within the NCBCs may be uniquely well positioned to make valuable contributions to research in the computational biosciences which require an exceptionally broad set of hybrid models and theories combining mathematical, logical (including temporal), and linguistic/semantic components to explain their heterogeneous data and knowledge over a wide range of scales, and levels of abstraction. Training experiences within the NCBCs may prepare these individuals to use ontologies in the representation of biological concepts and through “meta-experimentation” of an entirely new computational kind, develop methods to use ontologies in problem solving and biological experiment design, analysis, and interpretation. The present “portfolio” of NCBC centers is likely to change over the years as new problems, methods, and technologies are developed. The only way for the NCBCs to yield productive long-term researchers who can themselves train a new generation of interdisciplinary investigators is to ensure that they disseminate and explain their results and produce educational materials about their novel computational and biological contributions. Connecting the NCBC centers to existing NIH training programs is one approach, though alternatively, once they are sufficiently mature, the NCBCs could develop their own stand-alone educational program(s).

On the practical, technology infrastructure side, the NCBCs are already developing systems and software for advanced modeling and analysis, but these frequently require highly sophisticated prior knowledge of the model assumptions, empirical data constraints and their application within a complex set of software tools. During the past 40 years there have been many misapplications of sophisticated statistical, graphics and image interpretation, mathematical modeling, simulation, data mining programs, and languages. In the future, the more heterogeneous and complex combinations of methods and software will make it yet more difficult for investigators to apply the results of NCBC software and evaluate their results. An advanced educational objective of the centers could involve educating others in the foundational assumptions behind the software, and the “rules of practical application” to specific examples with successful outcomes as well those problems that prove too difficult to solve with the current tools.

The NCBCs could develop an outstanding educational tool by providing a critical perspective on the application of their computational methods to specific biomedical problems. This perspective would be invaluable as a means to educate the next generation of investigators in how questions can be posed differently, problems reformulated, and different models chosen or sought as result of a careful documentation of NCBC “successes and failures”. In contrast, training investigators to solve currently defined problems with current methods and technologies is merely an incremental component of center activities. A unique opportunity exists within the current and future NCBCs to create an educational environment for the next generation of computational scientists that can truly address the challenges of solving the most difficult problems in biomedical science now and in the future.

References:

NIH Biomedical Information Science and Technology Initiative, by the Biomedical Computing Advisory Committee to the Director: D. Botstein, L. Smarr, D. Agard, M. Levitt, D. Lippman, D. Herrington, C. R. Johnson, G Rose, G. Rubin, A. Levison, M. Spence, and H. Smith, C. Peskin, G. Jacobs. (www.nih.gov/about/director/060399.htm), June 1999.
NIH Roadmap for Medical Research [Online] http://nihroadmap.nih.gov

Materials provided to the Panel:

BISTI Report – This report was written in June 1999 by the ad-hoc NIH Working Group on Biomedical Computing to the NIH Director.
RFA-RM-04-022 – The request for applications which established the NCBCs.
NCBC Descriptions – Brief descriptions with links to each center’s website.
PAR-07-249 – The program announcement for collaborative R01 grants with the NCBCs.
PAR-07-250 – The program announcement for collaborative R21 grants with the NCBCs.
NCBC AHM Report – A report of the July 2006 All Hands Meeting
NCBC Management Plan – This is the organizational chart for the NIH management of the NCBC program.

The Software and Data Integration Working Group (SDIWG)
http://namic.org/Wiki/index.php/SDIWG:Software_and_Data_Integration_Working_Group
Biomedical Computation Review: http://biomedicalcomputationreview.org/

Review Panel:

Gwen Jacobs, Ph.D., Chair
Professor of Neuroscience
Asst. CIO & Director of Academic Computing
Cell Biology and Neuroscience
Montana State University

Carol Newton, M.D., Ph.D.
Professor
Department of Biomathematics
School of Medicine
University of California, Los Angeles

Mark Boguski, M.D., Ph.D.
Vice President and Global Head
Division of Genome and Proteome Sciences
Novartis Institute for Biomedical Research, Inc.

Ralph Roskies, Ph.D.
Co-Scientific Director
Pittsburgh Supercomputing Center
Professor of Physics
University of Pittsburgh

Craig Jordan, Ph.D.
Director
Division of Extramural Activities
National Institute on Deafness and Other Communication Disorders

James Schwaber, Ph.D.
Director, Daniel Baugh Institute for Functional Genomics/Computational Biology
Department of Pathology, Anatomy and Cell Biology
Thomas Jefferson University

Casimir Kulikowski, Ph.D.
Board of Governors Professor of Computer Science
Rutgers, The State University of New Jersey

Jonathan Silverstein, M.D., M.S., F.A.C.S.
Assistant Professor of Surgery and Radiology
Director, The University of Chicago Hospitals'
Center for Clinical Information
Scientific Director, Chicago Biomedical
Consortium
Associate Director, Computation Institute
The University of Chicago

Joel Stiles, M.D., Ph.D.
Director
Center for Quantitative Biological Simulation
Pittsburgh Supercomputing Center
Associate Professor
Mellon College of Science
Carnegie Mellon University

NIH Roadmap Bioinformatics and Computational Biology Co-Chairs
Jeremy Berg, Ph.D.
Director
National Institute of General Medical Sciences

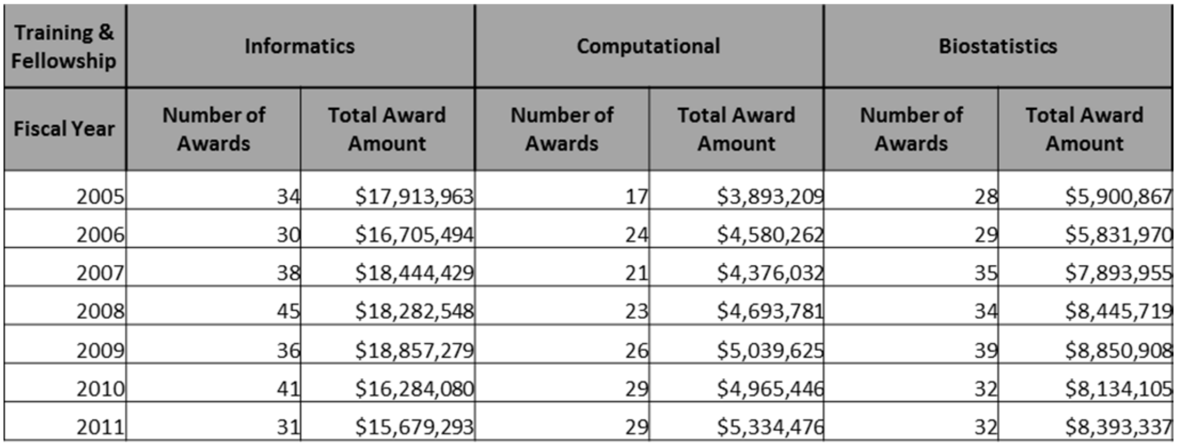
Donald A.B. Lindberg, M.D.
Director
National Library of Medicine

National Centers for Biomedical Computing Project Team
John Whitmarsh, *Leader*
NIGMS
Michael Ackerman
NLM
John Haller
NIBIB
Carol Bean
NHLBI
Michael Huerta
NIMH
Zohara Cohen
NIBIB
Don Jenkins
NLM
Milton Corn
NLM
Jennie Larkin
NHLBI

Valentina Di Francesco
NIAID
Peter Lyster
NIGMS
Greg Farber
NCRR
Grace Peng
NIBIB
Valerie Florance
NLM
Salvatore Sechi
NIDDK
Dan Gallahan
NCI
Karen Skinner
NIDA
Peter Good
NHGRI
Jen Villani
NIGMS

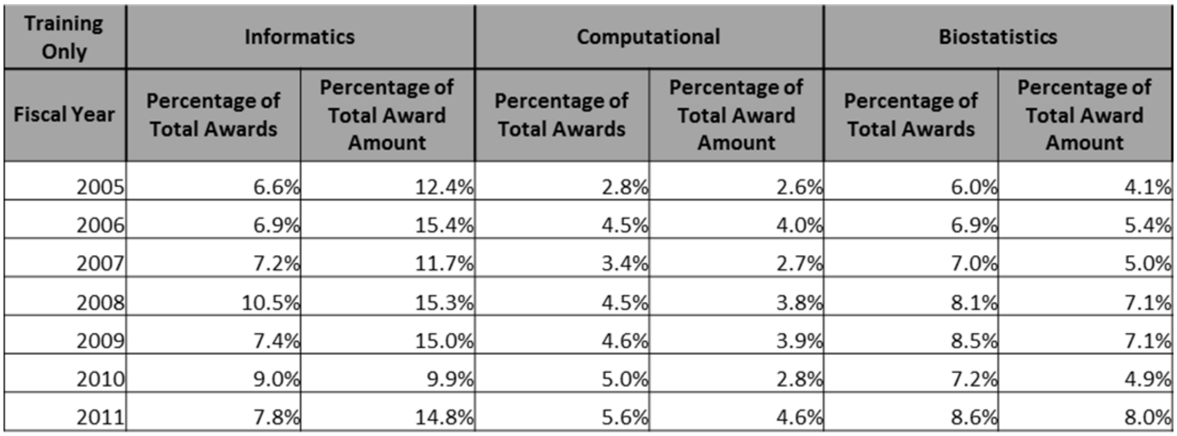
Office of Portfolio Analysis and Strategic Initiatives Staff
Rebecca Lipsitz
OD
Krista Pietrangelo
OD
Anne Menkens
OD

6.3 Estimates of NIH Training and Fellowship Awards in the Quantitative Disciplines



Training & Fellowship	Informatics		Computational		Biostatistics	
Fiscal Year	Number of Awards	Total Award Amount	Number of Awards	Total Award Amount	Number of Awards	Total Award Amount
2005	34	$17,913,963	17	$3,893,209	28	$5,900,867
2006	30	$16,705,494	24	$4,580,262	29	$5,831,970
2007	38	$18,444,429	21	$4,376,032	35	$7,893,955
2008	45	$18,282,548	23	$4,693,781	34	$8,445,719
2009	36	$18,857,279	26	$5,039,625	39	$8,850,908
2010	41	$16,284,080	29	$4,965,446	32	$8,134,105
2011	31	$15,679,293	29	$5,334,476	32	$8,393,337

Total Estimated Number and Funding Amount of Training (T) & Fellowship (F) Awards: These figures are estimates, based on a keyword search of training and fellowship titles in NIH's IMPAC II database.



Training Only	Informatics		Computational		Biostatistics	
Fiscal Year	Percentage of Total Awards	Percentage of Total Award Amount	Percentage of Total Awards	Percentage of Total Award Amount	Percentage of Total Awards	Percentage of Total Award Amount
2005	6.6%	12.4%	2.8%	2.6%	6.0%	4.1%
2006	6.9%	15.4%	4.5%	4.0%	6.9%	5.4%
2007	7.2%	11.7%	3.4%	2.7%	7.0%	5.0%
2008	10.5%	15.3%	4.5%	3.8%	8.1%	7.1%
2009	7.4%	15.0%	4.6%	3.9%	8.5%	7.1%
2010	9.0%	9.9%	5.0%	2.8%	7.2%	4.9%
2011	7.8%	14.8%	5.6%	4.6%	8.6%	8.0%

Total Estimated Percentage of the Number and Amount of Quantitative Training (T) Awards Relative to all Training Awards: These figures are estimates, based on a keyword search of training and fellowship titles in NIH's IMPAC II database.