AcademyHealth Responds to
National Library of Medicine Request for Information (RFI):
Next-Generation Data Science Challenges in Health and Biomedicine

Introduction

Data science is a multi-dimensional and foundational component of the broad, interdependent objectives of the National Institutes of Health (NIH) Strategic Plan. On behalf of the NIH, the National Library of Medicine (NLM) seeks information from researchers, clinicians, administrators and others in public and private sectors concerning promising data science approaches to the generation, characterization, management, storage, analysis, visualization, integration, and use of large heterogeneous sets that are relevant to health and biomedicine. NLM's long-term commitment to the principles of open science and evidence-informed practice provides a strong foundation for its expanded mandate to become the "intellectual and programmatic epicenter" for data science at NIH.[1]

As the professional home for more than 4,000 health services researchers, policy analysts, and practitioners whose work helps us understand and improve care for individuals, strengthen the performance of the health system, and enable better health outcomes for more people, at greater value, AcademyHealth has demonstrated a strong commitment to "openness" with the goal of increasing transparency and collaboration. In particular, the programs of the AHRQ-funded Electronic Data Methods (EDM) Forum encouraged open access and data sharing so that information can be used more efficiently and accessed more quickly to make a greater impact in research and policy. EDM Forum's open access, peer-reviewed journal, *eGEMs*, has published special issues on data governance to improve scientific and technical advances (2014) and on moving evidence into action by improving user interfaces (2015). We have helped to develop and support a variety of communities of practice to help promote multi-sector data sharing, data integration, and dissemination and have supported the development of new research methods for new data sources.

Based on our experience with open science and open data, AcademyHealth is pleased to have the opportunity to comment on promising directions for:

- New data science research in the context of health and biomedicine;
- New initiatives relating to open science and research reproducibility;
- Workforce development and new partnerships.

[1] https://acd.od.nih.gov/documents/reports/Report-NLM-06112015-ACD.pdf

Promising Directions for New Data Science Research in the Context of Health and Biomedicine.

NLM plays a critical role in working with standards development organizations and other initiatives to help to identify common data elements, to provide guidance on preferred data structures, and to help foresee challenges in data infrastructure and curation. Going forward, opportunities exist to expand NLM's efforts to address new research methods and approaches, including those to address the capabilities and risks of machine learning and human factors design, support team science, address social and environmental risk factors, and set priorities for data science research.

New Research Methods and Approaches

As new traditional and non-traditional data sources continue to emerge, including patient-generated data, there will be an increasing need to develop new research methods, governance structures, strategies for patient engagement and inclusion, and more effective approaches to use patient-reported measures in collaborative care planning and research. To that end, some of the most promising and daunting challenges facing data scientists in the future may be machine learning and user-centered design.

Machine learning and computer-research interface issues are an emerging area of activity where NLM's thought leadership can make a positive impact. Machine-driven analysis should be viewed as an opportunity to augment human cognition and research, not replace it. A number of factors are bringing the capabilities for large scale data analytics to bear on complex health and health care challenges. These include the large growth in digital data from all aspects of biomedical research and health, increased mobility and access to data, enhanced uses of cloud storage and computing, and the creation of a wide array of analytic tools and capabilities.

Although natural language processing and artificial intelligence (AI) are new methods, their applications are now implicated in almost all research domains. Computer science and engineering approaches are broadening the usability by researchers, policymakers, and technology developers to support discovery and translational research. Many new opportunities for extending the reach of machine learning to clinical care interpretations and augmenting health care delivery processes exist. The pace of this technology enhancement and its application may have profound impact on health care and society.

Furthering fundamental research that enables better understanding of the mathematical and engineering principles, and promoting transparency in these applications, will be needed to assure safe and effective use of these applications in research and health care. Support for open source repositories for algorithms, synthetic datasets, and validation protocols will be valued by the community of data scientists. Further basic and applied research on the enhancement of technical areas such as feature recognition (visual and auditory data), reinforced learning paradigms, and the interfaces of machine learning tools with electronic health record data, sensors and medical devices, and non-traditional health data sources will be needed.

In the near term (1 to 3 years), we anticipate that these technological advances will lead to process improvement, and enhance the quality of data and information available for researchers, policymakers, health care providers, and patients. The applicability of machine learning to healthcare in the longer-term (5 to 10 years), for example through clinical decision support or population health analyses, is intriguing but will require government and industry collaboration to assure the safety, efficacy *and* effectiveness of these approaches. Government-sponsored research and development in partnership with industry and academia can facilitate the adaptation of machine learning methodologies toward healthcare in a variety of means by ensuring a continually updated and educated workforce, infrastructure to support common data applications, and ensuring transparency of the performance of these tools.

More broadly, data science, mathematics, engineering, and related fields are needed to address such areas as structured data, metadata, data architecture, best practices for validation, etc. NLM should consider creating various avenues and approaches for common discussions and collaborative projects among biomedical researchers, health services researchers, and data science experts to ensure that context and the appropriate use and application of advanced analytic methods. Building the computational capabilities among public health, social science, and clinical research communities should be a high priority for advancing communication and information technology scientific fields. The NLM should consider collaborating with the Agency for Healthcare Research and Quality in these aspects.

In addition, broadband capabilities and telemedicine applications are providing new pathways to rapidly translate modern biomedical science and public health findings to broad populations at scale. By anticipating the rapid advances in computing and digital information capabilities, we can begin to envision new solutions to accelerating the transfer of new knowledge across the care delivery system and to improve coordination and information-sharing along the care continuum.

Information science and its application can also benefit by further expansion of human factors research. Building on the basic foundations of its essential role in improving patient safety, continued advances in human centered design of computer and information tools can lead to environmental influences that encourage better communication and understanding among care providers, care givers and patients. In addition, further fundamental research on neural networks and decision-making can be applied to patient comprehension and adaptability toward biobehavioral changes favoring healthier lifestyles that are augmented by machine learning enabled tools. Overall, we can anticipate that enhancing the machine-human cognitive interfaces will lead to improved efficiency and outcomes of health care practices, and have a major impact on disabilities and health disparities.

Addressing common information sources, common practices for uses of these technologies in large data analysis, descriptive methodologies for adaptation to health services research (HSR), and the many disciplines involved in HSR, would place NLM at the leading edge for these issues. Fostering dialogue and enhancing educational materials for the public will ensure understandings about the intent and useful applications of these new scientific capabilities.

Team Science

The predominant culture in biomedical, health services, and health systems research reflects a hierarchical, competitive ecosystem reinforced by academic incentives that reward researchers' competitive behaviors and practices. Researchers rarely share their data sets, and even in some of the larger practice networks, the practical and technical barriers of sharing and harmonizing data created for different purposes are prohibitive.

NLM could support development and use of new collaboration platforms, and could fund research on what makes communities of practice more or less successful, etc. For instance, NIH/NLM could create a platform for researchers, policy makers, students, and others to request or suggest data sets that could be of high value in their research or other applications. A demand-driven open data environment would be an interactive way in which citizens can connect with government researchers and resources to prioritize the establishment of data resources that are of high value and worthy of the effort to establish and maintain them. NLM could also directly tackle the barriers to team science by showcasing, rewarding and otherwise valuing individuals and institutions who adopt the principles of team science and community learning (e.g. by establishing national awards, including these behaviors in research review criteria, etc…).

Social and Environmental Risk Factors

After many years of focusing on the impact of the health care system in predicting health outcomes, attention is now shifting to the social and environmental risk factors that contribute to disparities and disproportionate impacts on low-income communities of color. For example, there are several efforts underway to develop processes and codes to capture social, economic, and other risk factors in electronic health records (EHRs).

NLM is already participating in some of the important efforts to spur information sharing between and among health and non-health sectors. We urge NLM to continue to work with its colleagues at HHS, including AHRQ, in bridging different communities toward the common goal of integrating information across health, social services, welfare, housing, transportation, criminal justice and other sectors affecting health outcomes. For example, the ONC-funded Community Health Peer Learning (CHP) Program worked toward population health improvements by using a peer learning collaborative. Specifically, program participants demonstrated how communities can link critical information within and outside of health care to address population health challenges ranging from birth outcomes to pediatric asthma to housing insecurity. By revealing key themes and challenges, and offering technical assistance in areas such as data governance, community engagement, systems infrastructure, and sustainability, these community-based efforts have the potential to inform strategy for NLM and NIH overall and align with other delivery system reform efforts driving toward better care, smarter spending, and healthier people.

NLM's expertise in data structure and data standards could play an invaluable role in accelerating the integration of data across all these sectors influencing health outcomes.

Priorities for Data Science Research

First, the biomedical research community could help advance the understanding around data science by working with the stakeholder communities to develop common definitions and a lexicon of terms that can help provide context to the massive interest in data and science. In many ways, the technological capabilities of the information age have far outpaced our recognition of the societal and cultural implications, and common understanding about the impact of the principals and tools.

For a working definition, NLM could begin by adapting or adopting the definition of **Data science**, or **data**-driven **science**, as an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from **data** in various forms, either structured or unstructured, similar to **data** mining. Further, NLM could encourage cross-disciplinary dialogue among STEM professions and biomedical and behavioral training programs along with biomedical sciences to hasten the career development pathways for multidisciplinary research. Further conceptualization of how data science can augment processes and methods in support of translational research, patient-centered outcomes research, and clinical trials should be a high priority.

Secondly, for the purposes of advancing evidence based medicine practices and health systems research, NLM could work with the rest of the NIH Institutes and Centers and AHRQ to advance the principles of open data and establish the scientific management framework needed to assure data sharing is a requirement that is delivered upon by researchers and their institutions. To incentivize this, NIH and AHRQ could use its management and funding authorities to reinforce acknowledgement and standard citations for researchers using datasets developed from publicly funded research.

Thirdly, to advance the scientific agenda for data science, NLM could work with the rest of the NIH Institutes and Centers and AHRQ to encourage greater access to new sources of data along with clinical and research data sources through linked data and machine learning practices. NLM could also hasten the development of metadata standards and repositories, and encourage the foundation of an open source community for modular algorithms to be used in analytic methods.

Promising Directions for New Initiatives Relating to Open Science and Research Reproducibility

Reproducibility, or the replication of research and capacity for data to be duplicated and reused, is indispensable for maintaining public trust and advancing discovery. However, the data, codes, and detailed methodologies for many studies are not available, accessible, or transparent.

We view the role of NLM as being critical in promoting data sharing through language included in grant announcements, ensuring consistent application of open data and open science by NIH program officers, and supporting and participating in multi-sector and international initiatives that increase the dissemination of findings through multiple channels.

To advance the goal of reproducibility, NLM can support or conduct research:

- Demonstrating the value of data sharing and transparency in experimental methodology, observation, and collection and curation of data;
- Evaluating the difference in update of research findings in policy and practice by making scientific data publicly available and reusable;
- Communicating about research in accessible and transparent ways, including a greater emphasis and reliance on data visualization science;
- Strengthening the data infrastructure by creating and supporting web-based tools that facilitate collaboration among communities of practice;
- Supporting researchers' use of the most appropriate and robust dataset for their studies by developing and maintaining a central clearinghouse of available datasets across public and private sectors; and
- Supporting professional development and fellowship opportunities that foster future leadership in data science, open science, information science, and data visualization science.

Promoting Adoption of the FAIR Principles

NLM should work with the rest of the NIH Institutes and Centers and AHRQ to play a leadership role in promoting adoption of the FAIR principles in data sharing to help guide open science and data science. Recognizing that publicly funded research can have sustainable impact and enhanced societal benefit by the sharing and reuse of data, policies and practices among NIH researchers should be encouraged to ensure that data objects are available for use by machines and people in formats that are *findable, accessible, interoperable* and *reusable* by others (i.e., FAIR principles). Further, NLM should work with the rest of the NIH Institutes and Centers and AHRQ and undertake management policies to encourage applicants, reviewers, and stewards of publicly funded research to recognize the use of, and ensure the production of, data resources that follow the FAIR principles. As a result, discovery and translation of research results will be achieved at a faster pace and attain greater society good as a result. The extended benefit of this at a global level could be achieved through advances in scientific understanding and application in low and middle-income countries that do not have the resources to support fundamental basic research in health and medicine.

Promising Directions for Workforce Development and New Partnerships.

Workforce Development

From the perspective of formal academic training, the number of data science programs is growing extremely rapidly to meet market demand. While there is not yet a consensus on the core skill set for degree programs, most programs provide training in programming, including data cleaning, wrangling, manipulation, and presentation; statistics, including modeling and simulation; communication, including data visualization; and some understanding of the healthcare enterprise, including research methods and data sources.

NLM has a commendable track record of supporting dozens of informatics fellows who have emerged as thought leaders for the field. NLM should continue to support efforts to develop and update core competencies for emerging areas of specialization and also should continue to support training and professional development opportunities for informaticians. These training opportunities should expand to include more data scientists, health professionals, human factors experts, software engineers, and others whose diversity of perspective will help to move information science forward faster. In our view, data science competencies go beyond just technical expertise and also include an understanding of team science, change management, collaborative methods of evidence generation, and community-building in order to realign the analytics ecosystems to support collaborative methods. A number of these competencies also align well with the recently published Learning Health Systems Research Core Competencies. We encourage NLM to partner with AHRQ as they further strengthen their workforce development programs and investments.

Communities of Practice and Learning Health Systems

To ensure that future generations continue to build on the transformational work led by NLM informatics fellows and other NIH fellows who have been trained in a way that promotes continual learning and evidence-based improvements, NLM should engage collaborative networks of researchers as mentors for fellows in projects that promote transparency of methods, data-sharing, and joint publications.

NLM could gain significant insights by increasing its participation and support for communities of practice and learning networks that expand peer-to-peer learning. We have found that these more innovative methods for translation and information-sharing make it much easier for junior researchers and faculty in data and information science to become involve in methods development, an activity that is often reserved for very senior individuals.

Summary and Closing

The sustainability of the national research enterprise depends on increasing understanding of the value of data science methods and open science paradigms, an understanding that needs to be actively modeled and encouraged by NIH at all levels.

NLM leadership in these areas will also be reliant upon a sustained and reliable commitment to supporting new data science research in the context of health and biomedicine; new initiatives relating to open science and research reproducibility; and workforce development and new partnerships.