The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data

Ronald Margolis,[1] Leslie Derr,[2] Michelle Dunn,[3] Michael Huerta,[4] Jennie Larkin,[5] Jerry Sheehan,[4] Mark Guyer,[6] Eric D Green[6]

[1]National Institute of Diabetes and Digestive and Kidney Diseases, NIH, Bethesda, Maryland, USA
[2]Office of the Director, NIH, Bethesda, Maryland, USA
[3]National Cancer Institute, NIH, Bethesda, Maryland, USA
[4]National Library of Medicine, NIH, Bethesda, Maryland, USA
[5]National Heart, Lung and Blood Institute, NIH, Bethesda, Maryland, USA
[6]National Human Genome Research Institute, NIH, Bethesda, Maryland, USA

Correspondence to
Dr Ronald Margolis, National Institute of Diabetes and Digestive and Kidney Diseases, NIH, 6707 Democracy Blvd, Room 693, Bethesda, MD 20892-5460, USA; margolisr@mail.nih.gov

Received 8 May 2014
Accepted 17 June 2014
Published Online First 9 July 2014

To cite: Margolis R, Derr L, Dunn M, *et al*. *J Am Med Inform Assoc* 2014;**21**:957–958.

ABSTRACT

Biomedical research has and will continue to generate large amounts of data (termed 'big data') in many formats and at all levels. Consequently, there is an increasing need to better understand and mine the data to further knowledge and foster new discovery. The National Institutes of Health (NIH) has initiated a Big Data to Knowledge (BD2K) initiative to maximize the use of biomedical big data. BD2K seeks to better define how to extract value from the data, both for the individual investigator and the overall research community, create the analytic tools needed to enhance utility of the data, provide the next generation of trained personnel, and develop data science concepts and tools that can be made available to all stakeholders.

INTRODUCTION TO THE PROBLEM

Across many areas of science, technological and conceptual advances are resulting in the increasingly rapid generation of large amounts of data. While research has always involved the collection and organization of data, the volume, variety, and velocity of current 'big data' production presents new opportunities and challenges in both scale and complexity. At the same time, there is a broader cultural shift underway from approaches that kept data mostly private with sharing of resultant knowledge in the form of publications to an information-based culture that dynamically engages the scientific community through the active sharing of both data and publications. Big data are not only a new reality for the biomedical scientist, but an imperative that must be understood and used effectively in the quest for new knowledge. Needed are new approaches for data management and analysis that allow scientists to better access and extract value from data so as to advance research and discovery.[1 2] Key stakeholders in the coming biomedical big data ecosystem include data providers and users (eg, biomedical researchers, clinicians, and citizens), data scientists, funders, publishers, and libraries. Implementation of such an envisioned biomedical big data ecosystem will depend upon cultural changes and require updated policies related to funding, data sharing, and data citation. Metrics on the efficacy of data sharing and re-use may help inform important decisions at funding agencies and research institutions about the support for and career advancement of biomedical scientists, and could be invaluable in incentivizing necessary cultural changes across the biomedical research enterprise. The aim of this perspective is to describe how the National Institutes of Health (NIH) has started catalyzing such changes through its Big Data to Knowledge (BD2K) initiative.

BACKGROUND

A report to the NIH Advisory Committee to the Director from its Data and Informatics Working Group (DIWG) in June 2012 (http://acd.od.nih.gov/diwg.htm) recommended the establishment of a broad and inclusive trans-NIH program for addressing the opportunities and challenges presented by biomedical big data. Indeed, the spirit of the report suggested that failure of the NIH to act at this crucial juncture would represent 'institutional malpractice' (as stated by one DIWG member). To emphasize the need for a coherent and forward-looking response to these opportunities, the position of Associate Director for Data Science (ADDS) at NIH was created with the goal of placing big data and data science at the highest level of decision-making at the NIH. The first ADDS, Dr Philip Bourne, has begun his tenure at the NIH, and among his responsibilities is oversight of the BD2K initiative. BD2K has been formulated as a programmatic response to the DIWG recommendations and consists of four focused areas: (1) improving the ability to locate, access, share, and use biomedical big data; (2) developing and disseminating data analysis methods and software; (3) enhancing training in biomedical big data and data science; and (4) establishing centers of excellence in data science.

APPROACH

Taking into account the substantial current investment of individual NIH Institutes and Centers in bioinformatics and computational biology within their own spheres of interest, BD2K has first focused on identifying the pressing general but unaddressed needs related to biomedical big data, assessing community expectations, and identifying current policies and practices related to the production, handling, and utilization of biomedical big data. Outreach to the diverse community of stakeholders mentioned above has been at the forefront of these efforts through targeted Requests for Information (RFI), workshops, consultation with leaders in various fields, as well as extensive discussions across agencies and within the NIH about the relationship(s) between potential BD2K components and other ongoing activities (http://bd2k.nih.gov/workshops.html).

The resulting input strongly ratified the DIWG's recommendations that supporting the development of

software and applications for analyzing biomedical big data was only part of the solution. Further support came from the Holdren/OSTP memo issued in February 2013 (http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research) and directing all agencies of the government to ensure the results of federally funded scientific research are made broadly available.

A first step in fostering the emergence of data science as a discipline relevant to biomedicine is the development of solutions to specific high-need challenges facing the research community. Thus, the first BD2K Funding Opportunity was for investigator-initiated Centers of Excellence in Data Science (RFA-HG-13-009); this called for proposals to test and validate novel ideas in data science that not only focused on particular challenges but also had the potential for broad impact.

The second launched BD2K area aimed to enhance the training of methodologists and practitioners in data science. Skills in demand under the data science 'umbrella' include computer science, mathematics and statistics, biomedical informatics, biology and medicine, and others, all incorporated as 'data science.' At the same time, the generation of large amounts of data together with the complex questions being posed, requires interdisciplinary teams to design the studies and perform the subsequent data analyses. The BD2K training initiatives seek to seed the development of investigators in all parts of the research enterprise who are well-trained in data science. Hand-in-hand with training is the need for cultural changes to assure that the contributions of scientists well-trained in data science are appreciated and rewarded, including the provision of appropriate career paths with commensurate incentives and rewards.

A fundamental question for BD2K is how to enable the identification, access, and citation of (ie, credit for) biomedical data. The DIWG proposal for federated data catalogs, as distinct from data repositories, requires descriptions of and pointers to the data. Inherent in data discovery is the need for a sustainable and scalable plan to create and maintain a discovery system that allows researchers to readily find and cite biomedical data. Indeed, sustainability and scalability are two intertwined issues that must be addressed in order for the advances made possible by BD2K to have a lasting effect. A necessary first step has been recognition of the need to assemble and validate ideas drawn from the broader scientific community in developing a Data Discovery Index (DDI). The goal of the initial efforts in this area is to define an effective and efficient mechanism(s) for indexing that will enable the discovery of relevant, existing datasets through the use of metadata and index terms. The DDI will enable advanced approaches to search, integrate, and facilitate visualization of data. Stakeholders will be encouraged to learn from related efforts in other fields, and conduct short-term pilot studies to explore different ways in which a DDI might be developed and used. Central to development of the DDI will be the ability to link data to associated publications to enhance discovery and facilitate better understanding and interpretation of data and associated analyses. Again, cultural and policy changes will likely be needed to both enhance data-sharing policies and incentivize data sharing.

Among the difficult challenges is how best to create greater value from the expanding availability and use of electronic health records (EHRs). Clinical data from EHRs, together with individual health data captured by various personal devices, offer considerable opportunities for advancing clinical and biomedical research. Data from clinical sources could very well provide a cost-effective means to study different health interventions, to conduct large-scale surveillance of disease incidence and progression in real time, to identify patient cohorts for recruitment into clinical trials, and more. However, unlike most other forms of biomedical research data, clinical data are typically captured outside of traditional research settings and must then be re-purposed for research use. Doing so raises important issues of consent and protection of patient privacy. Changes in policies and practices are needed to govern research access to clinical data sources and facilitate their use for evidence-based learning in healthcare. Improved approaches to patient consent and risk-based assessments of clinical data usage for research are also high priorities. Improved quality and quantity of clinical data available for research (eg, by creating shared libraries of phenotype elements and algorithms and by providing access to key sources of data), as well as new methodologies for analyzing clinical data, are all needed for ethical and informed use of these data.

While a DDI will undoubtedly help to make data findable and citable, large datasets can also give value beyond their original purpose when the data are appropriately annotated so that they can be used in a meaningful way. Critical to all of these activities will be the need for formulating, conducting, and maintaining community-based data and metadata standards efforts for reproducibility that will: (1) be driven by the clearly identified needs of the community; (2) include a core group of developers with appropriate expertise; (3) engage stakeholders across the community from the outset and in an ongoing way; and (4) evolve based on broad input to become increasingly dynamic and responsive to community needs.

CONCLUSION

Addressing the challenges associated with biomedical big data must of necessity engage all parts of the big data ecosystem. While these challenges are complex, they are also addressable. BD2K is deploying an integrated plan of action that will tackle numerous aspects of the big data challenge, including multiple elements of data science, training, policy, and community behavior. By fostering open communication with the stakeholders, BD2K will contribute to the broader vision for the future of biomedical big data. While the NIH is only one part of the biomedical big data ecosystem, it can provide leadership for convening stakeholders, providing seed funding, developing metrics for success, implementing a working process, establishing and modifying policies, and supporting basic infrastructure that can catalyze long-term solutions for the biomedical research community. Through a working partnership between and among the relevant stakeholders, useful solutions will emerge and lead to vibrant and sustainable models for capitalizing on biomedical big data and translating data into new knowledge.

Acknowledgements Helpful edits were provided by Susan Gregorick (NIGMS), Belinda Seto (NIBIB), and Lynda Hardy (NINR).

Contributors RM conceived of the idea for the article, and performed the literature search. RM and JL jointly wrote the article, with contributions from all other authors. RM and JL share joint responsibility as guarantors.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The NIH has a full data sharing policy. All publications must be openly available through the NLM.

REFERENCES

1 Science. Dealing with data. *Science* 2011;331:639–806. http://www.sciencemag.org/content/331/6018.toc#SpecialIssue
2 Nature. Science in the petabyte era. *Nature* 2008;455:1–136. http://www.nature.com/nature/journal/v455/n7209/index.html