HealthManagement.org
Promoting Management and Leadership

NIH's Role in Biomedical Big Data

Across many areas of science, technological and conceptual advances are resulting in the increasingly rapid generation of large amounts of data (called 'big data') in many formats and at all levels. 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. 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.

A report of the National Institutes of Health (NIH) Advisory Committee of the Director from its Data and Informatics Working Group (DIWG) in June 2012 recommended the establishment of a broad and inclusive trans-NIH programme for addressing the opportunities and challenges presented by biomedical big data. As a result, 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 Big Data to Knowledge (BD2K) initiative consisting 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 centres of excellence in data science.

Approach

Taking into account the substantial current investment of individual NIH institutes and centres 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 (ie, production, handling and utilisation of data). Outreach to the diverse community of stakeholders mentioned above has been at the forefront of BD2K efforts through targeted Requests for Information (RFI), workshops, consultation with leaders in various fields, and discussions across agencies and within the NIH about the relationship(s) between potential BD2K components and other ongoing activities.

The resulting input strongly ratified the DIWG's recommendations that supporting the development of software and applications for analysing biomedical big data was only part of the solution. Further support came from the

Holdren/OSTP memo issued in February 2013 and directing all agencies of the government to ensure the results of federally funded scientific research are made broadly available.

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.

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. 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 DDI will enable advanced approaches to search, integrate, and facilitate visualisation of data. 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.

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.

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 behaviour. 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 catalyse long-term solutions for the biomedical research community.

Reference:
R Margolis, L Derr, M Dunn, M Huerta, J Larkin, J Sheehan, M Guyer, E D Green; J Am Med Inform Assoc
doi:10.1136/amiajnl-2014-002974
(http://jamia.bmj.com/content/early/2014/07/09/amiajnl-2014-002974.full?g=w_jamia_open_tab)
Image credit: Wikimedia Commons

Published on : Mon, 11 Aug 2014