Data Science Opportunities for the National Cancer Institute

Interim Report of the National Cancer Advisory Board Working Group on Data Science

August 14, 2018

NATIONAL INSTITUTES OF HEALTH
National Cancer Institute
National Cancer Advisory Board

Ad Hoc Working Group on Data Science

CO-CHAIR

Mia A. Levy, M.D., Ph.D.
Director
Cancer Health Informatics and Strategy
Ingram Associate Professor of Cancer Research
Associate Professor
Biomedical Informatics and Medicine
Vanderbilt-Ingram Cancer Center
Vanderbilt University
Nashville, Tennessee

CO-CHAIR

Charles L. Sawyers, M.D.
Chairman
Human Oncology and Pathogenesis Program
Memorial Sloan Kettering Cancer Center
Investigator, Howard Hughes Medical Institute
Professor of Medicine
Weill Cornell Medical College
New York, NY

Regina Barzilay, Ph.D.
Delta Electronics Professor
Department of Electrical Engineering and Computer Science
Member, Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
Cambridge, Massachusetts

John D. Carpten, Ph.D.
Professor and Chair
Department of Translational Genomics
Director, Institute of Translational Genomics
Keck School of Medicine
University of Southern California
Los Angeles, CA

Amanda Haddock
President
Dragon Master Foundation
Kechi, Kansas

George Hripcsak, M.D.
Vivian Beaumont Allen Professor of Biomedical Informatics
Chair, Department of Biomedical Informatics
Director, Medical Informatics Services
New York-Presbyterian Hospital
Columbia University
New York, New York

Mimi Huizinga, M.D., M.P.H.
Vice President and Head of Strategic Data,
US Oncology
Novartis
Nashville, Tennessee

Rebecca Jacobson, M.D.
Vice President of Analytics
University of Pittsburgh Medical Center
Enterprises
Pittsburgh, Pennsylvania

Warren A. Kibbe, Ph.D.
Professor and Chief,
Translational Biomedical Informatics
Department of Biostatistics and
Bioinformatics
Chief Data Officer
Duke Cancer Institute
Duke University School of Medicine
Durham, North Carolina

Michelle LeBeau, Ph.D.
Arthur and Marian Edelstein
Professor of Medicine
Director
The University of Chicago
Comprehensive Cancer Center
The University of Chicago
Chicago, Illinois

Anne Marie Meyer, Ph.D.
Director, Epidemiology
Real World Evidence, IQVIA
St. Prex, Switzerland
Adjunct Assistant Professor
Department of Epidemiology
Gillings School of Global Public Health
The University of North Carolina
at Chapel Hill
Chapel Hill, North Carolina

Vincent Miller, M.D.
Chief Medical Officer
Foundation Medicine
Cambridge, Massachusetts

Sylvia Katina Plevritis, Ph.D.
Professor
Department of Radiology and Biomedical
Data Science
Co-Chief, Integrative Biomedical
Engineering
Informatics at Stanford (IBIIS)
Stanford University School of Medicine
Stanford, California

Kimberly Sabelko, Ph.D.
Senior Director
Scientific Strategy and Programs
The Susan G. Komen Breast Cancer
Foundation, Inc.
Dallas, Texas

Lincoln Stein, M.D., Ph.D.
Head, Adaptive Oncology
Ontario Institute for Cancer Research
Professor, Cold Spring Harbor Laboratory
Cold Springs New York
Professor, Department of Molecular
Genomics
University of Toronto
Toronto, Ontario
Canada

Nikhil Wagle, M.D.
Assistant Professor
Department of Medicine
Harvard Medcal School
Medical Oncologist
Department of Medical Oncology
Dana-Farber Cancer Institute
Associate Member
The Broad Institute
Boston, Massachusttes

Ex Officio Members

Daniel Gallahan, Ph.D.
Deputy Director, Division of Cancer Biology
National Cancer Institute
National Institutes of Health
Bethesda, Maryland

Anthony Kerlavage, Ph.D.
Acting Director, Center for Biomedical
Informatics and Information Technology
Office of the Director
National Cancer Institute
National Institutes of Health
Bethesda, Maryland

Lynne Penberthy, M.D., M.P.H.
Associate Director
Surveillance Research Program
Division of Cancer Control and Population
Sciences
Office of the Director
National Cancer Institute
National Institutes of Health
Bethesda, Maryland

Louis M. Staudt, M.D., Ph.D.
Director
Cancer for Cancer Genomics
Office of the Director
National Cancer Institute
National Institutes of Health
Bethesda, Maryland

Executive Secretary

Elizabeth Hsu, Ph.D., M.P.H.
Biomedical Informatics Program Manager
Center for Biomedical Informatics and
Information Technology
National Cancer Institute
National Institutes of Health
Bethesda, Maryland

Table of Contents

Executive Summary

In May 2018, Dr. Norman Sharpless charged the Data Science Working Group to provide general guidance to the National Cancer Institute (NCI) on opportunities for NCI in data science, big data, and bioinformatics to further cancer research. Given the quickly evolving pace of data science, the Working Group decided to issue a series of targeted recommendations over time for the consideration of the National Cancer Advisory Board rather than wait for a comprehensive report on the entire data science ecosystem. This strategy allows NCI to move in a more accelerated fashion to address pressing needs, particularly those for which there is rapid consensus. The recommendations presented in this interim report represent the first set of recommendations identified as priority areas where the NCI could move quickly but do not represent a final set of recommendations. The areas covered in this initial set of recommendations include:

- Investments to leapfrog data sharing for high-value datasets
- Harmonization of terminology between cancer research data and clinical care data
- Support of data science training at the graduate and post-graduate level
- Opportunities for funding challenges and prizes

Each recommendation is presented as a stand-alone recommendation to enable targeted action by the NCI, but there are inter-relationships between all the recommendations and together they address important areas of opportunity for NCI in data science.

Introduction

Being able to collect and access vast amounts of patient data, along with new technologies which generate vast amounts of research data, has led to unbridled enthusiasm for “big data,” “real-world” data and evidence, and precision medicine. However, it is important to recognize that while unprecedented amounts of data are being generated, much of the data may not be research-ready, and many cancer researchers do not have the appropriate training, skillsets, or experience to appropriately leverage these data. In particular, oncology is a data hotspot with considerable promise, but the immediate application can be oversold and ‘hyped’; considerable attention must also be paid to issues such as interoperability of independently generated datasets, informed consent and protection of patient privacy for the use of data science. The NCI should play a leadership role in funding the creation and sharing of cancer research data, ensuring sustainability of its investment in the creation of such data, creating mechanisms for these data to be made broadly available to the research community, defining responsible data use policies and processes, and supporting the training of the next generation of cancer data scientists. While the promise of data science is real and its application to important problems in cancer care and cancer research tangible, there is significant hard work to be done in data management, data harmonization, computational training, improving data sharing, and in workforce development before the promise can be fully realized. These recommendations and those that the Working Group will make in future are designed to help NCI maximize the benefit of data science in cancer research.

Investments to Leapfrog Data Sharing for High-Value Datasets

Introduction

NCI has the opportunity to enhance the value of existing high-value datasets by providing funding to support collection of additional data to fill gaps in these cancer datasets, as well as support the work required to ensure interoperability of these data through sharing in a public repository such as the Genomic Data Commons (GDC). Support of these efforts will increase the usefulness of these datasets and enable novel insights in cancer biology, treatment, and outcomes by the global cancer research community. High-quality, accessible datasets are foundational for data science approaches to maximize the knowledge extracted from these datasets.

Recommendation summary

Create funding opportunities to support identification, enrichment, curation, harmonization, annotation, and publishing of existing high-value datasets through the NCI Cancer Research Data Commons (CRDC). Eligible datasets would include those fully collected and annotated, but not yet shared in a public repository. These funding opportunities should also support datasets that would be greatly enhanced by additional data generation, data linkages, and/or data collection, such as genomic datasets needing additional clinical information or annotation and vice versa. The goal of this recommendation is to enable broad data sharing of high-value datasets, in particular to enable multiple smaller, well-annotated, well-designed studies to contribute to the publicly available data in the NCI CRDC, and to emphasize the importance of data sharing for solving important problems in cancer research. In addition, an analysis of the current data landscape would help inform NCI's identification and prioritization of existing high-value datasets, beyond those identified through these funding opportunities.

Background

The maturation of the GDC and the ongoing development of the NCI CRDC provides a platform for enabling better data management and data sharing for cancer research. While the existing GDC has many of the features required for a powerful data sharing platform, submitting small to mid-scale studies to the GDC requires more effort and expertise than is typically available to a project team. Further, there is often limited clinical information accompanying the genomic data. Public datasets enable insights into cancer; datasets such as The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and AACR Project Genomics Evidence Neoplasia Information Exchange (GENIE) have transformed our understanding of the cancer somatic tumor mutational landscape. But there is a pressing need for more datasets linking genomic data with clinical outcomes, particularly longitudinal data in the setting of metastatic disease where many genome-based treatment decisions are most commonly made today and in the foreseeable future. Examples of these types of studies include the CoMMpass Study, Count Me In, and Beat AML. Beyond genomics, open repositories such as The Cancer Imaging Archive (TCIA) and the NCI CRDC need complementary datasets to maximize their value to the cancer research community.

The proposed funding opportunities would provide the resources for investigators to prepare and deposit their datasets to the GDC and other components of the NCI CRDC, as well as potentially enrich the data with additional clinical data elements (for instance diagnosis, staging, therapy, therapeutic response, or imaging data) or additional genomic data (such as RNAseq or whole exome sequencing).

Supporting such funding opportunities also provides NCI an opportunity to influence critical conversations on database development and articulate key data requirements for specific types of research. Quality, completeness, or utility of a dataset can vary greatly depending on the research question and are complex constructs to globally define. Some datasets are considered very "wide" and include a large variety of variables on large samples of patients while other datasets are "deep" indicating a very precise and extensive level of detail on fewer key variables (e.g., tumor mutations).

Recommendation

- Provide funding opportunities that support resources to increase the accessibility and utility of high-value datasets best poised for data sharing, recognizing the effort required for meaningful data sharing.
- To support investigators funded under these opportunities, the NCI should also provide the following:
    - A simplified timeline estimating when the NCI CRDC and its various components will be ready to work with different types of datasets.
    - An NCI facilitation team to work with the funded researchers and with the teams building the NCI CRDC on harmonization of datasets.
    - An education and outreach plan to support and encourage cancer researchers to share cancer research data; this should include support for interpreting NIH and NCI data sharing policies and requirements and for understanding NCI-supported mechanisms for sharing data, such as the NCI CRDC.

Benefit for the cancer research community and NCI

In addition to enabling the broader sharing of high-value datasets, such funding opportunities would also achieve the following:

- Advertise the capabilities and capacity of the NCI CRDC.
- Define the training and experience necessary to contribute to and utilize the NCI CRDC (for example, data harmonization and publishing to NCI CRDC) for the next generation of data driven science and discovery.
- Provide a vehicle for coordination and engagement of the funded investigators and other groups across the cancer research community to collaborate to define best practices for data collection, data generation, data harmonization, terminology services and making data FAIR (Findable, Accessible, Interoperable, Reusable). For example,

groups funded through these funding opportunities would likely work with the teams defining data commons and the semantic infrastructure for the NCI CRDC. This would help facilitate agreement on how to represent data types, including outcomes, across cancer types, identify improvements to data models, and improve/simplify the import/export of tools available through the NCI CRDC.

- A better understanding of the data landscape for oncology (both private and publicly funded).
- A clearer definition and prioritization of the research questions and "use cases" which require additional investments in data repositories for the NCI.
- A more systematic way to identify "lowest hanging" fruit, e.g. valuable databases that require minimal investment to turn them into high(er)-value datasets.

Appropriately funding the resources needed to support data sharing will encourage all members of the cancer research community to be avid data sharers. Sharing of datasets, as demonstrated by TCGA, can have a transformative impact on the velocity and trajectory of cancer research and maximize the impact of cancer research on reducing the burden of cancer. In the short-term, the proposed funding opportunity will increase the number of high-value, well-annotated cancer datasets shared in the NCI CRDC as well as the number of longitudinal, real-world data sets with genomic data linked with therapeutic information and clinical outcomes.

Terminology Harmonization

Statement of the Problem

Electronic health record (EHR) data may contribute significantly to cancer care research, but such use requires agreement on what the data mean. There is a plethora of data standards involved in cancer research; the recent adoption of the Clinical Data Interchange Standards Consortium (CDISC) standards, Clinical Data Acquisition Standards Harmonization (CDASH) and Study Data Tabulation Model (SDTM), by NCI's Cancer Therapy Evaluation Program (CTEP) for National Clinical Trials Network (NCTN) reporting was an important step for clinical trials research, particularly because the Food and Drug Administration (FDA) is now requiring electronic submissions. However, CDISC standards cannot necessarily be adopted easily in the collection of data for clinical care. In parallel, organizations such as the Office of the National Coordinator of Health Information Technology (ONC), the National Library of Medicine (NLM), and SNOMED are standardizing clinical representations in the EHR. Unfortunately, clinical care data standards and cancer research data standards are not well harmonized, limiting the benefit of using EHR data for cancer research. The large size of the US healthcare system adds to the challenge, as adoption and modification of EHR data standards is a slow and expensive process.

Recommendation summary

To improve the ability to utilize clinical care data in cancer research, the working group recommends NCI work with standards development organizations to augment EHR data standards such as RxNorm, LOINC, and SNOMED CT. In addition, the working group recommends that NCI fund research related to achieving near clinical trial grade data within traditional clinical care settings. Such work may be carried out in collaboration with NLM, ONC, FDA, other branches of HHS, and standards organizations such as SNOMED. To support these recommendations, NCI should identify and prioritize existing standards development organizations and activities with whom to collaborate, particularly in closing the gap between cancer research data and clinical care data. Once these partners have been identified, NCI should work with them and the broad informatics community to harmonize standards for interoperability between cancer research and clinical care.

Background/History

In all fields, there is a typical data-standards lifecycle where competition among groups generates multiple standards early on until one becomes predominant. In cancer research, however, two dichotomies have significantly increased the data standards multiplication.

First, because of both the nature of cancer and its treatment, as well as the independent and strong funding for cancer, multiple cancer-specific data and terminology standards have been developed. For example, whereas most non-cancer diagnoses are encoded in vocabularies like ICD9-CM (pre-2015 billing), ICD10-CM (post-2015 billing), and SNOMED CT (clinical), cancer uses vocabularies tailored to cancer (e.g., ICD-O). Most other disease areas have insufficient resources to develop specialized vocabularies, and must therefore work with general vocabulary developers to incorporate as many terms as possible. This is not always the best approach since each disease has its own specific needs. The cancer vocabularies, on the other hand, can incorporate stage and grade as axes in its diagnoses, whereas an approach that did not involve development of disease-specific vocabularies would require incorporating those features into a general vocabulary like SNOMED CT. With over 100,000 terms, the NCI Thesaurus (NCIt) includes wide coverage of cancer terms as well as mapping with external terminologies.

Second, there is a dichotomy between data collected for research and data collected in EHRs for the provision of care, be it in cancer or not. In clinical research, there tends to be a manageable number of variables, and the data tend to be manually collected, curated, and validated. There is a complex management structure (from funders through research management to research coordinators who collect the data) that permits some degree of top-down imposition of standards. The CDISC data model is one well-known and widely implemented standard which is particularly well-suited for clinical research intended to generate evidence for FDA submissions, but adoption of CDISC may not translate well into data collected for the provision of care.

Data collected from the provision of care, on the other hand, are driven by the two trillion-dollar health care industry, which in turn is largely driven by its financing, resulting in standards focused more on billing than on clinical workflow. Although adopted widely, clinically-driven standards are therefore more difficult to impose. Typical data models include Observational Medical Outcomes Partnership (OMOP), Patient-Centered Outcomes Research Network (PCORnet), Accrual for Clinical Trials (ACT), and Sentinel (the latter for claims data only). These models use general rather than cancer-specific data standards. They typically incorporate hundreds of thousands of terms across medicine, which are used by millions of health care workers to document health care on hundreds of millions of patients. It is not a nimble system, yet it contains critical information of great value to cancer research. Many data elements collected in routine clinical care, which are critical for oncology, are not collected as structured data elements nor with the same definition rigor as those in clinical trials. This limits everything from clinical trial patient-matching to real-world data generation. Prime examples of this limitation are the lack of data elements for tracking of disease progression and recurrence of disease, and even data fundamental to cancer, such as stage and tumor grade. The NCI is uniquely positioned to contribute to this conversation, and to work productively across NIH as well as with other HHS agencies, including the ONC, FDA, and the Centers for Medicare and Medicaid Services (CMS), to incorporate structured elements into reporting requirements.

One question in translating between clinical terminologies is whether it is feasible to convert data from general clinical care to cancer care research. This has been an ongoing discussion across NIH and standards communities in exploring secondary uses of EHR data. While not specific to cancer, informatics research has looked at the translation among clinical terminologies. An important translation for research is from billing terminologies like ICD9-CM and ICD10-CM to clinical terminologies like SNOMED CT. Reich et al. (Journal of Biomedical Informatics 2012) found that translations among ICD9-CM, SNOMED CT, and MedDRA caused differences in research cohorts, but that studies performed on those cohorts produced the same results. In a study of nine clinical phenotypes, Hripcsak et al. (submitted for publication) showed that cohort errors caused by translating from billing to clinical vocabularies can be limited to 0.26%, a rate far lower than other sources of errors in observational research. Ruijun Chen et al. (work in-progress, funded by NCI) reported that the accuracy of cancer care can be inferred from EHR data that have been converted to a research database using a set of standard vocabularies like SNOMED CT. They used manual chart review and a cancer registry as the gold standards and found a positive predictive value of 96% in detecting any cancer and a sensitivity of 99%, with a calculated specificity of 99.9%. In a related study of cancer treatment, exposure to any chemotherapeutic agent was correctly identified 100% of the time (in 50 cases), but only 82% were actually used for cancer therapy as opposed to non-cancer indications. Similarly, hormone therapy was correctly identified in 98%, with 84% receiving it for cancer treatment. Immune therapy was 100% and 94%, respectively. Radiation therapy was correctly identified in 86%, although most of those errors represented lack of information to prove it had been given; only 4% were clearly miscoded. Given these good-to-excellent results on a first attempt, along with recognizing a non-zero error rate in the traditional approach to generating cancer registries, it appears that EHRs may be able to support cancer care research.

Cancer registries may serve as a bridge between cancer research and cancer care. Standards can be imposed in registries in a manner similar to cancer clinical trials (i.e., cancer-specific and top-down). While there are efforts underway at the NCI to improve data abstraction for registries in an automated, near real-time fashion, as well as to extend the depth of information contained in registries, registries have not traditionally recorded the nuances of care that are in health records, including comorbidities, past history, and detailed responses to treatment. In addition, the resource burden of registry participation and maintenance may limit the scalability unless additional automation is employed.

Recommendations

We believe that there is a strong need to augment existing terminology standards and to harmonize terminology standards between cancer research and clinical care. We have identified three initial areas that can lead to improved cancer research, along with a longer-term approach on how to achieve improved data sharing.

- Identify resources to augment the work of the NLM Lister Hill Center, which is tasked with maintaining standards for medications and laboratory tests, on RxNorm to ensure that new cancer therapies are covered in a timely manner; to assess the extent to which investigation drugs should be included; and to identify resources to augment work on LOINC’s coverage of cancer-related laboratory tests.

- Identify resources to extend SNOMED CT to better cover the types of information required in cancer care (e.g., stage, grade). This could take the form of funding for the Lister Hill Center, which holds the US license for SNOMED CT and manages the Unified Medical Language System (UMLS), or SNOMED International to carry out work, funding of the SNOMED community, or funding of outside groups. This work will need to be closely coordinated with NCI Enterprise Vocabulary Services (EVS) to help ensure alignment of SNOMED CT terms with required CDISC NCIt terms. For example, the Observational Health Data Sciences and Informatics (OHDSI) Oncology Workgroup for the OMOP Common Data Model is working on mapping ICD-O to SNOMED CT.

- Improve understanding of the potential barriers to achieving near clinical trial grade data within traditional clinical care settings through research. In order for patients to fully benefit from the innovation in oncology care, we must better understand the barriers within the care delivery system. Electronic health care data that is incomplete, incorrect, implausible, non-concordant, or out-of-date can negatively impact patient care (Weiskopf and Weng, JAMIA 2013). The NCI could provide funding opportunities for research focused on understanding and exploring the gaps and barriers between clinical trial data and clinical care data. This research could contribute strategies and approaches to translation between standards and work to incorporate cancer-specific data elements into more general terminology standards such as SNOMED CT.

Longer term approach

- Identification of standards efforts that are critical to data sharing in cancer research.
- Engagement with content and domain experts to determine the appropriate process and recommend or confirm priority standards and activities.
- Involvement of key terminology and data consumers and implementers to ensure the end product is fit for purpose.
- Involvement of both cancer and non-cancer researchers and clinicians to ensure interoperability, reduce the gaps and barriers, and improve understanding between the groups.

Factors for success

- *Collaborative efforts, not parallel efforts*: When creating data standards and designing tools to use those standards, we offer caution in proceeding too strongly in parallel to and separate from the general informatics community. The recommendations of the May 2018 NCI Workshop on Semantics to support the NCI Cancer Research Data Commons are excellent, well thought out, and right on-target, but may encourage a parallel and separate path for cancer. The recommendations must be carried out within the general community using general standards.
- *Work in the context of the broader community*: Similarly, as data standards expand in scope, such as the addition of metadata standards to support new technologies, this should be done in the context of the broader community. Demonstrations of harmonization of standards can be carried out across broader clinical community, including the NCI-funded OHDSI cancer data standards project, the PCORnet network, and the ACT i2b2 network for Clinical and Translational Science Awards (CTSAs).
- *Leverage existing standards development organizations and activities to incorporate cancer-specific terminology, especially the general terminology standards such as SNOMED*: By leveraging non-cancer specific vocabularies, there is an opportunity to speed the availability and use of cancer terminologies more broadly. The field of oncology should focus development on the areas that are most unique to oncology (e.g.,

staging, progression, repeated progression, molecular characteristics), and in addition strive to develop terminologies within existing vocabularies where possible. The NCI should evaluate existing cancer terminologies and determine which data elements could be developed within existing vocabularies. If no existing vocabulary meets the needs for oncology, then the NCI should evaluate adapting existing vocabularies versus creating an oncology specific vocabulary.

Definition of success

We see success as a set of collaborative efforts between the broad cancer research community, the NCI, and the clinical care community to harmonize data and terminology standards, and the uptake of these standards in both research and clinical care. This will enable better patient care, more efficient research strategies and better real-world evidence generation. Oncology is one of the most rapidly innovating fields of medicine and can lead the way in bridging data from clinical trials to improving patient care.

Data Science Training

Introduction

There is no disputing that the era of biomedical Big Data has brought with it a host of challenges. Making data accessible for sharing, storing, and analyzing is a strategic imperative in healthcare. This is particularly true for cancer, where integration between genomic, proteomic, clinical, imaging data, and many other data types, and the ability to manipulate and analyze these data, will be required for the breakthroughs the research community is seeking. Because of this, there is a significant need for data scientists who can facilitate these breakthroughs. But the availability of these critical resources has not kept up with the demand for them. For example, in January, 2017, hiring platforms Glassdoor and Indeed listed over 100,000 available data science jobs.[1] This is only expected to grow exponentially. According to IBM, by the year 2020, the number of data science and analytics jobs across all industries will reach 700,000 annually.[2] In a survey conducted by the MIT Sloan Management Review, 40% of companies indicated they were "struggling to find and retain" data scientists.[3] Similarly, biomedical scientists and clinicians need exposure to data science to understand the challenges, the necessity of collaboration with informaticists, and to expand their own expertise and abilities. Additionally, the ability of PhD students to land tenure-track positions has diminished greatly in recent years, and data science may represent another path to solid employment.[4] The NIH recognized this critical need in the Strategic Plan for Data Science, stating that "more needs to be done to facilitate familiarity and expertise with data-science approaches and effective and secure use of various types of biomedical research data" and that there is a "need to grow and diversify the pipeline of researchers developing new tools and analytic methods for broad use by the biomedical research community."[5] Clearly, training and education to increase the availability of skilled resources needs to be a priority for the cancer research community and for NCI.

Recommendation Summary

We recommend that the NCI increase the number of graduate training programs and trainees in Cancer Data Science using four approaches: (1) dedicating a specific T32 training program in cancer data science, (2) collaborating with the National Library of Medicine (NLM) T15 training programs, (3) contributing to the National Institute of General Medical Sciences (NIGMS)

[1] Dunn MC, Bourne PE (2017) Building the biomedical data science workforce. PLoS Biol 15 (7): e2003082. https://doi.org/10.1371/journal. pbio.2003082

[2] Columbus, Louis (2017) IBM Predicts Demand for Data Scientists Will Soar 28% by 2020. https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020

[3] Business.com (2017) Big Data, Big Problem: Coping with Shortage of Talent in Data Analysis. https://www.business.com/articles/big-data-big-problem-coping-with-shortage-of-talent-in-data-analysis

[4] Offord, Catherine (2017) Addressing Biomedical Science's PhD Problem. https://www.the-scientist.com/careers/addressing-biomedical-sciences-phd-problem-32258

[5] NIH Strategic Plan for Data Science. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf

Medical Scientist Training Program, and (4) developing a short term training program to increase opportunities for additional training in cancer data science for clinicians and biological scientists.

Background

The current workforce in cancer data science is insufficient to support the demand for individuals with this unique combination of technical, analytic, and scientific skills. One reason for the undersupply is the small number of targeted training programs that prepare students to succeed in such varied positions. This recommendation addresses (1) pre-doctoral (MS, PhD) and postdoctoral training, and (2) extramural training, but is broad enough to include, (3) programs leading to an enhanced workforce across academics, industry, and government. Current limitations of training in data science are characterized as follows:

- There are currently no data science focused **T32 training programs** at NCI, and few across NIH as a whole. Historically, NCI has preferred to keep the T32 program announcement as broad as possible. One consequence of this approach is that it reduces the impact that NCI can have on development of programs in this important and rapidly evolving field.
- The National Library of Medicine (NLM) **T15 training program** offers a strong foundation for building additional capacity in cancer data science. Although this program has included additional slots from disease-focused Institues and Centers (National Institue of Environmental Health Sciences, National Heart Lung and Blood Institute, National Institue of Dental and Craniofacial Research), there have never been NCI directed slots.
- **Medical Scientist Training Programs** in some institutions provide a strong potential recruiting pool for data scientists but they can be quite disconnected from computational and quantitative graduate programs that may be in other health sciences schools or outside of the health sciences.
- There are few short-term and no long-term training programs targeted specifically towards data science as an alternative career pathway for recent PhDs and postdoctoral scholars who are not able to obtain faculty positions. Such programs could arguably increase the data science workforce while simultaneously addressing the oversupply of postdocs and PhDs.

Recommendations

1. **Initiate a specific T32 program focused on cancer data science training**. It is anticipated that programs best positioned to train PhD students will include: (a) cross-disciplinary faculty including those from genetics, epidemiology, computer science and statistics departments, (b) curricula which build on existing foundational programs (genetics, computational biology, systems biology, computer science, statistics, biomedical informatics, data science, etc.), (c) curricula that reinforce interdisciplinary learning and teaching, (d) integration with other cancer-related training programs within the institutional setting

2. **Support additional cancer-specific biomedical informatics training within the existing NLM T15 program**. This recommendation includes both: (a) providing supplemental funds to existing awardees with strong cancer research programs to support pre-doctoral and post-doctoral trainees whose research includes a cancer-focus, and (b) creating a long-term partnership with NLM to support additional slots in the next funding opportunity announcement (FOA). In the latter case, specific NCI training and curricula requirements can be included for programs to address in their applications.
3. **Enhance the existing Medical Scientist Training Program with funding for MD/PhD research training in cancer data science.** This could be accomplished by providing a supplement to the existing program and/or by co-signing a future FOA. In either case, it is recommended that this be accompanied by a requirement for data science research in a laboratory with clear relevance to NCI's mission.
4. **Support a new set of short term cross-training program (3-12 months),** potentially including online programs, designed to: (a) provide additional training experiences in data science to PhDs in biological and chemical sciences, population sciences and other related disciplines, (b) provide additional training experiences in data science to clinically trained professionals, and/or (c) provide biological or clinical training to encourage entry into cancer-specific fields by computer scientists, applied mathematicians, statisticians, and data scientists. Outreach to multidisciplinary computational sciences programs is encouraged for the development of short-term training content in cancer-specific data sciences.

Conclusion

Working to fill the gaps and current need for bioinformatics resources through a variety of training programs will provide tremendous benefit to the cancer research community. The development of new tools and techniques in data analysis and sharing is imperative to achieve the necessary breakthroughs in cancer treatments. Cancer researchers need the in-depth understanding of data science if they are to collaborate with informaticists and make use of the vast and varied data becoming available. Additionally, provision of cross-training programs creates growth opportunities for informaticists and scientists, who may be looking for ways to advance their careers within or outside the context of academia. Encouraging these kinds of programs through the recommendations herein is an excellent step in moving towards making data science a potential pathway for pre and post-doctoral scholars. We believe additional training programs area also necessary for undergraduate students and other individuals and will address these requirements in a future recommendation.

Challenges/Prizes

Recommendation Summary

The working group recommends that NCI sponsor a series of scientific challenge competitions in the area of data science to address critical issues in cancer research. The challenges should be open to all members of the academic and commercial cancer research communities and overseen by a neutral party who will be responsible for conducting the challenge, vetting the data, comparing the submissions, and adjudicating the winners. The prizes should be a mixture of academic recognition, such as a publication, and financial awards, such as a contract to further develop and disseminate the winner's algorithm/software.

Background

We are living through a golden age of biological data science in which sophisticated statistical models, machine learning techniques, and other algorithms are being brought to bear on large molecular and clinical data generated by increasingly sophisticated instruments as 'digital first' data sets. Each year thousands of new and updated algorithms, software packages, and computational tools are published, and many of these tools have been widely adopted by the community. However, there remain persistent problems with biological software, including issues of portability ("it won't install on my system!"), reliability ("it runs but crashes!"), and reproducibility ("it doesn't give the answer that appears in the published paper!"). In addition, the academic publishing model does not provide a way for a researcher to easily choose the software best suited for their tasks. When a new algorithm or software package is published, the authors may benchmark their software against a handful of existing algorithms that perform the same task, but there is a strong bias in these exercises. While there are some communities that have implemented more structure and standardization for testing analytical models, such as the machine learning (ML) research community, issues persist in benchmarking algorithms for many areas of cancer research, such as genomics. For example, algorithm developers may have an incentive to tweak their algorithm to give the best performance and accuracy for their own particular software and a potentially idiosyncratic instrument and sample protocol. Not having the same level of expertise with the competing packages or access to 'gold standard' training and validation data sets, they are unable – or not inclined – to perform the same tuning on others. So new software often appears to be an improvement over existing software and it is very hard to determine how generalizable the improvement may be. The overall impact of these issues is to reduce the research community's access to the best performing tools.

Scientific competitions ("challenges") can be an effective solution to these problems. In a typical challenge, the computational task and a dataset with a known solution ("ground truth") are selected. Challenge participants are given access to a training data set with a known ground truth; this training data set is used to refine their algorithms. In many cases, the groups are also provided with a development training data set to test their models. The participants upload their algorithms, which the challenge administrators run against a test data set to which the participants have been blinded. Challenges can feature multiple rounds of increasing complexity and can incorporate a "leaderboard" feature that allows contestants to see how their submission ranks against their competitors. In several recent challenges, contestants have been required to upload their software into a uniform cloud-based compute environment in which the software is run on the test data and scored in a fully automated fashion, a design that has multiple advantages both in terms of simplifying challenge logistics and promoting software reusability.

Several recent DREAM challenges sponsored by Sage Bionetworks and the NCI (Nat Biotechnol. 2014 Dec;32(12):1213-22; Lancet Oncol. 2017 Jan;18(1):132-142), have shown the effectiveness of the scientific challenge approach in biological data science. Challenges have included prediction of drug response from genomically-characterized cell lines, discovery of prognostic molecular biomarkers in cancer, prediction of response to therapy in rheumatoid arthritis, and a series of challenges in cancer variant identification and interpretation. Each of these challenges garnered a robust set of contestants and were successful in identifying the best performing algorithms and raising the accuracy of the algorithms overall.

Recommendation

NCI should sponsor an ongoing series of data science challenges related to pressing problems in cancer biology and care. A modest number of challenges in the range of 4-8 per year would be an appropriate target.

Several priority areas were considered based on (1) the impact of the problem; (2) the availability of data sets that either have experimentally-derived ground truth, or for which it can be generated synthetically; (3) whether the challenge is logistically manageable. The following series of research areas that are both impactful and have abundant data that can be applied to a challenge were identified:

- Drug response prediction (in cell lines, animal models, human trials)
    - Examples: response to immunotherapy, adverse events
- Discovery of multi-'omic prognostic biomarkers
- De-convolution of heterogeneous tumors
- Cancer diagnosis, grading and staging (histology, imaging, genomics, proteomics)
    - Example: determination of the primary from characteristics of a metastasis
- Facilitation of data access and integration from the ethical, legal, and social implications (ELSI) standpoint
    - Examples: machine-readable participant consents, automated approval of data access requests, protocols for establishing trust/delegation responsibilities amongst data access committees

The priority areas described above address pressing problems in cancer research and care. The ability to perform deep 'omics analysis (genomics, transcriptomics, proteomics) on clinical tissues economically and at-scale makes it possible to collect rich data sets on tumor and host tissues. However, interpreting this data is a challenge, and the community's ability to transform 'omics data into actionable recommendations for treatment has been hindered by the lack of a common testing framework on which to evaluate different algorithms against each other. The highest priorities, therefore, are ones focused on discovery of predictive and prognostic biomarkers, improved diagnosis based on advanced molecular and imaging techniques, and the use of ‘omics and imaging technologies to dissect the cellular composition of complex tumors.

What would be needed for success?

A successful challenge requires:

- *A clearly defined and quantifiable task*, for example, to estimate the LD_{50} of a particular drug against a particular cell line based on genome and transcriptome of the cell line.
- *Appropriate data sets:*
    - *The ground truth is available,* for example, a drug x cell line screening set in which the LD_{50} has been empirically determined.
    - *The ground truth is not already published,* to prevent overtraining
    - *The legal hurdles for accessing the data are not burdensome,* in the case of data sets that require data access committee approval, the approval process must be reasonably easy to affect.
- *An infrastructure to run the challenges on,* including shared storage for exchanging results, a system for executing contestants’ submissions, a system for assessing the accuracy of the submissions, and a mechanism for posting scores and rankings. In most cases, challenges will benefit from having a mechanism to return the information needed for contestants to understand the sources of errors made by their algorithms.
- *An infrastructure for disseminating results to the community,* including publication vehicles, publicized awards ceremonies, and code repositories for making submitted software easily accessible to the community.
- *Incentives,* including guaranteed publications for winners and runners-up, cash prizes, and/or grants and contracts focused on improving the performance and usability of the winning algorithms.

Another type of challenge that could be utilized in certain situations is an idea or problem-based challenge. In this case, the community is being asked not to solve a specific problem, but instead to help generate new ideas or specific questions. These can be much broader and generally less applied than the more solution-based challenges, which as defined here, are more computational in nature and have a specific tangible deliverable. For example, idea challenges may solicit ideas for new concepts in cancer etiology or new cancer detection tools. Scoring would be somewhat more subjective and generally involve a review panel. Since there is no “gold standard,” results may vary and not be worthy of further pursuit. In order to ensure broad participation, incentives and outreach should be the same as for the problem-solving challenges. The process might then take the path of a more structured challenge or an idea used internally by the NCI.

The Working Group noted several key strategic steps needed to design and deploy a successful challenge. These are described in Appendix A. In addition, the Working Group noted potential synergies among the challenge concept and other Data Science Working Group recommendations. In particular, the creation of curated and standardized data sets promoted by the “Leapfrog” Data Sharing Subgroup, and the harmonization of terminology promoted by the Terminology Harmonization Subgroup could be facilitated by challenges targeted at data standards, harmonization systems, and tests of interoperability. Challenges can also promote the goals of the Training Subgroup by providing cash prizes and other incentives earmarked for trainee participants.

Definition of success

We see success as creating a robust, recurrent system of computational challenges and prizes, which are held several times per year and span an evolving set of topics in cancer biology and care. The challenges would spur research in computational cancer biology, measured as increasing publication rates for journal articles and software packages. Winning tools and algorithms would be further developed and refined, increasing the availability of advanced analytic software to the broader research community, measured as citations of the software packages that participated in the challenges. Most importantly we would expect to see the developed tools and solutions have a direct impact on the understanding and management of cancer.

Appendix A: Steps to Create a Challenge

The Working Group noted several key steps needed to design and deploy a successful challenge.

1. *Select the challenge topic.* A successful challenge requires clearly defined and ideally quantifiable task(s). For some topic areas, a suitable task may not be immediately obvious. Under these circumstances, we recommend beginning with an "idea challenge" in which participants are invited to submit proposals for challenges/prizes in the area of interest. Proposals would be judged by an impartial panel based on significance and feasibility, the latter including such criteria as availability of suitable test data sets. The winning idea(s) become the basis for challenges executed in subsequent phases.
2. *Identify the evaluation criteria.* The next step is to identify the metrics and evaluation framework that will be used to rank challenge submissions. The Working Group recommends piloting the evaluation framework and metrics in advance of beginning the challenge.
3. *Identify the test data set and the "ground truth."* This may be the most difficult data set to identify, because the best data set is often one that has been subjected to experimental validation but is not yet published. The Working Group suggests forming strategic alliances with groups that are privy to unpublished data, such as journal editors, lead investigators on intramural and extramural NCI-sponsored clinical trials, leads of pharmaceutical-sponsored clinical trials, and NCI-sponsored research consortia and networks, e.g., NCI Molecular Analysis for Therapy Choice (MATCH) and Surveillance, Epidemiology, and End Results (SEER) programs, in order to receive advance notice of unpublished data sets that may be suitable for use.

   In some cases, it may be appropriate for the challenge organizers to perform or commission the creation of an experimental data set designed to test a specific type of algorithm. A concrete example might be the commissioning of a CRISPR screen for gene knockouts that modulate the immune response in a T-cell model, in order to test gene network-based predictive models of immunity.
4. *Establish the infrastructure for executing the challenge.* This includes the computational infrastructure for executing submitted algorithms against the test data, ranking the submissions against the pre-established metrics, and feeding back results to the contestants. There is a strong argument to be made for infrastructures in which contestants submit their algorithms to a cloud-based system for automatic execution, because these environments discourage cheating and preserve copies of the submitted software for later replication. However, such environments should not interfere with contestants' ability to debug their algorithms, and we believe it should always be possible for contestants to download the raw test data set for deeper inspection of their algorithm's performance. Ideally the infrastructure is general enough so that it can be reused for multiple challenges.
5. *Perform a dry run.* Perform end-to-end testing of the challenge, confirming that the test data set is accessible, that the submission/ranking system is working, and that the infrastructure has the capacity to handle the expected load.

6. *Establish the incentives for the challenge.* Ideas discussed include cash prizes for challenge winners, a publication, recognition at a scientific meeting, or a contract/grant to further develop the winning software.
7. *Publicize the challenge* in appropriate venues, such as specialty journals, widely-read computational biology/bioinformatics blogs, web sites of professional organizations and NCI-sponsored resources, e.g., NCI-GDC, Twitter feeds, and the funding opportunities section of the NCI web site, as well as publicizing through key stakeholders, such as NCI-designated Cancer Center and NCATS awardees, among others. Particular attention will be required to ensure that notifications of the challenges reach the broadest audience representing diverse expertise. This should include the core machine learning community and targeting researchers who are not traditionally involved in NCI activities.
8. *State the rules, evaluation criteria and deadlines clearly* in the materials distributed in advance of the challenge.
9. *Execute the challenge.* The challenge is run by the administrators according to the rules established.