Data Science Opportunities for the National Cancer Institute

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

June 10, 2019

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.
The Sheba Foundation Director
Rush University Cancer Center
Rush University Medical Center
Rush University
System Vice President for Cancer Services
Rush System for Health
Chicago, Illinois

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, New York

Brian Alexander, M.D., M.P.H.
Senior Vice President
Clinical Development
Foundation Medicine, Inc.
Associate Professor, Radiation Oncology
Harvard Medical School
Dana-Farber Cancer Institute
Brigham and Women's Cancer Center
Cambridge, Massachusetts

Regina Barzilay, Ph.D.
Delta Electronics Professor
Department of Electrical Engineering and Computer Science
Member, Computer Science and Artificial Intelligence Lab
Member, The Koch Institute for Integrative Cancer Research
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.
Senior Data Scientist, Real World Data Collaboration
Personalized Health Care Data Science
F. Hoffman-La Roche Ltd.
Basel, 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

Sylvia Katina Plevritis, Ph.D.
Professor and Chair
Department of 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 Medical School
Medical Oncologist
Department of Medical Oncology
Dana-Farber Cancer Institute
Associate Member
The Broad Institute
Boston, Massachusetts

<u>Former Member</u>

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

<u>Ad-Hoc Member</u>

Sohrab Shah, Ph.D.
Chief
Computational Oncology
Memorial Sloan Kettering Cancer Center
New York, New York

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.
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
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
Office of the Director
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 (NCAB) rather than wait for a comprehensive report on the entire data science ecosystem. These recommendations were presented as an interim report and accepted by the NCAB in August, 2018. The interim report is presented in its entirety in Appendix A, but briefly, the areas covered included:

- 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

As a complement to the interim recommendations, the Data Science Working Group has expanded the interim recommendation on training and identified three additional areas as important opportunities for the NCI in data science. The topics covered in this final report include:

- Additional areas of support for data science training and workforce development
- Developing machine learning infrastructure for cancer research
- Facilitating the appropriate use of real world data
- Enabling the cultural shift toward data sharing

As in the interim report, 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

The interim recommendations of the Data Science Working Group highlighted the need for NCI to 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. The targeted recommendations in that report focused on key areas for NCI to begin to achieve these goals through (1) investments in high-value data sets, to enhance the value of existing data sets for cancer research; (2) harmonization of terminology between cancer research data and clinical care data, to improve the ability to utilize clinical care data in cancer research; (3) targeted training programs in data science for pre-doctoral and post-doctoral students, to ensure that the next generation of cancer researchers are properly equipped to work in an increasingly data-intensive environment; and (4) opportunities for funding challenges and prizes, to explore alternatives to supporting computational research in cancer.

This final report of the Data Science Working Group further highlights opportunities for NCI to achieve the goals outlined above, as well as two emerging research areas in which NCI should engage. As emphasized in our interim recommendations, a workforce equipped with the knowledge and tools to work with data is fundamental to cancer research; without the people to leverage the data being generated, we cannot make progress in cancer research. We then highlight two emerging scientific areas in which it will be critical for NCI to engage, machine learning and real world data. Machine learning has the potential to transform cancer research, but NCI must pave the path forward in order for machine learning to move beyond low-hanging fruit and tackle difficult scientific questions in cancer research. Real world data is increasingly being leveraged by the research community; again, NCI must lead the way forward to enable researchers to understand and leverage real world data appropriately. Finally, given the inherent importance of data to enabling scientific discovery and in data becoming a public good which researchers funded by the NCI should be sharing for the benefit of broader cancer research, we end our recommendations with a statement on data sharing and recommendations for how NCI can accelerate the cultural shift toward data sharing.

Data Science Training and Workforce Development

Introduction

This recommendation is an expansion of the data science training recommendation made in the interim report of this Working Group to the National Cancer Advisory Board (NCAB) (see Appendix A). Because data science training represents a large need for the cancer research community and incorporates several broad areas of discussion or focus, the Working Group elected to create two recommendations. The first was submitted to and approved by the NCAB in August of 2018 and primarily focused on pre-doctoral, post-doctoral, and extramural training, targeting strategies to provide data science training for cross-disciplinary skills acquisition or to create an alternate career path for PhDs. As discussed in our initial recommendation, there is a dearth of data scientists available to fill the needs in both the commercial and academic sectors. McKinsey predicts that "the demand for data scientists is increasing so quickly that … by 2018 there will be a 50 percent gap in the supply of data scientists versus demand."[1]

However, the issue is not just the lack of a workforce, but also the ability of data science community to understand the specific research and clinical needs of the cancer community and the unique challenges in using cancer data sources. According to Andre Sionek of *Toward Data Science* online magazine, "The biggest issue with most data scientists is that they are not actually scientists" and they therefore don't understand the data with which they are working.[2] This underscores the need for cross-training and the development of cross-domain, "hybrid" researchers who understand both the biology and clinical aspects of the data and who also have the computational skills required to analyze and gain new insights from these data. Computational scientists in undergraduate programs may not have knowledge of opportunities to work in medical research, such that they often look to companies such as Google for opportunities in the commercial sector rather than considering biomedical informatics. Phil Bourne and Michelle Dunn also recognize this need in their publication *Building the Biomedical Data Science Workforce*, where they state that one of the primary goals of the BD2K training initiatives is to "establish biomedical data science as a career path."[3] Additionally, cancer researchers themselves may not realize the gap in their own knowledge, or be aware of data science training programs available to them, despite the need to understand data science in order to do their work and advance their career. Bench scientists and clinicians need to collaborate with data scientists to analyze the data they are generating, but without at least a fundamental understanding of data science they will be unable to do so. Finally, there are two additional audiences/communities for whom data science training would be tremendously useful: patients and patient advocates, and K – 12/undergraduate students. Patients and patient advocates would benefit from high-level data science training, since data sharing has become a ubiquitous call from these groups. Training that provides them with the

[1] https://blog.alexa.com/know-data-science-important/

[2] https://towardsdatascience.com/do-you-really-need-a-data-scientist-fcdfc226f4e4

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5517135/

understanding and language to speak with researchers and data scientists would be very beneficial. The K – 12 and undergraduate audience for training would focus on how to engage young students as early as possible to provide understanding of the career paths available in cancer research data science. Training that fills these gaps for scientists and clinicians, and that addresses undergraduates who are making career choices, is a clear need for the cancer research community.

Recommendation Summary

We recommend several steps to increase the number of cross-trained researchers and data scientists and to attract computational scientists to the field of cancer research. (1) Develop and disseminate an online catalog of existing data science training resources; (2) Convene a workshop to determine: skills needed, existing training, strategies for accessibility of training, potential partnerships across industry, non-profit, academia, and government, and approaches to fill the needs; (3) Engage computational scientists who want to work in cancer research; (4) Add data science training opportunities for biologically-trained investigators; (5) Support "hybrid" researchers, including training grants, strategies to recognize and acknowledge the hybrid individual and her/his contributions, and clear opportunities for advancement; (6) Partner with organizations offering data science training to young people (K – 12 and undergraduate) to expose students to this domain early, and to enhance that training with biological and cancer-specific knowledge.

Background

Many cancer researchers currently have insufficient skill sets in data science to allow them to collaborate with data scientists effectively. Additionally, these same scientists may have no knowledge of training resources or programs available to them to fill this gap. Computational scientists or young people interested in data science, at the undergraduate or even the high school level, may have no exposure to the opportunities available to apply their skills and knowledge to pursue a career in data science for biomedical research. There is a clear need for individuals who can breach the gap between computational and data science and cancer biology and research, and yet training to create these individuals is not readily available or is not adequately publicized. Finally, a "hybrid" workforce that works across domains and acts as the glue between the cancer biologists and the computational scientists, facilitating the collaboration and the analysis and discovery, needs to be recognized for the critical role they play and there must be a formal means of acknowledging their contributions.

Recommendations

1. **Develop and disseminate an online catalog of existing data science training resources:** We must recognize that numerous resources have been developed for data science training; however, we lack a "one-stop shop" to help guide cancer researchers seeking data science training – a clearinghouse that lists and explains all the available options. Many academic institutions now incorporate data science training into their curriculum, with certificate, undergraduate and graduate programs (examples: Stanford, University of Pennsylvania, Yale, Johns Hopkins). Limited listings do exist, such as those provided by the NIH Data Science Workforce Development Center, but are not tailored for a cancer research

audience. A comprehensive online catalog would be a great resource for the community. This resource would also benefit patients and patient advocates, who may be seeking more in-depth knowledge of data science to enhance their advocacy work.

2. **Convene a data science training workshop**: NCI has the ability to gather the community to collaborate on the needs in data science. The cancer research and data science education communities could be brought together to determine: the fundamental informatics skills needed (for cancer researchers, to understand the computational field); the domain-specific skills needed (for bioinformaticists, to understand the science); existing training that meets fundamental needs in data science; strategies for accessibility; potential partners for government programs (public/private partnerships); and to develop strategies to address these needs, including recommendations as to training that needs to be developed to address unique issues encountered in cancer research. Working on these issues as a community would create momentum in the development of programs, identify the gaps, create a mechanism for feedback and collaboration, and publicize both the needs and what is currently available (which could also be a contribution to the online catalog of resources outlined in the first recommendation).

3. **Support and engage computational scientists who want to work in cancer research:** Computational scientists may be at a disadvantage if they move to a faculty position immediately after their graduate work because of lack of exposure to biomedical research. NIH grants are very problem-driven; method is important, but secondary to understanding the nuances of the data and domain, which can make it challenging for computational scientists to be successful at receiving funding from NCI/NIH. NCI can help provide opportunities for data scientists and informaticists to gain the training required in cancer biology, clinical research, and clinical care that would allow them to work in the field of cancer research. These could include workshops or targeted courses that issue a certificate which would indicate the range of skills acquired, along with the opportunity to do wet lab work that involves generating data. Additionally, we recognize that because of the high demand for data scientists in the private sector, it may be difficult to recruit computational scientists to work in cancer research. Prestigious fellowships such as the United States Digital Service, Presidential Innovation Fellows, and the Department of Health and Human Services Entrepreneurs-in-Residence Program serve as mechanisms to entice people working in industry, often computational or data scientists, to work in the public sector for short periods of times. NCI could consider establishing similar fellowships, perhaps in partnership with academic institutions, to entice computational or data scientists to work in cancer research.

4. **Add data science training opportunities for biologically-trained investigators:** Just as it is important to engage computational and data scientists in cancer research, it is likewise important to ensure that cancer researchers, both clinical and basic scientists, have knowledge of the skills and tools they need to understand data science. There are three ways NCI can support this need. First, the courses and assets described in the catalog outlined in the first recommendation should include the ability to search based on specific field of interest and the current skill set of the researcher. A scientist may have minimal

understanding of data science and informatics and wish to gain basic competency, or a moderate level of knowledge and a desire to go deeper. Courses fulfilling these needs should be easy-to-find in the catalog. Second, NCI can support post-doctoral training programs that specifically provide data science and informatics training to biologically-trained PhDs. This kind of training ensures that young scientists enter the workforce with at least a basic knowledge of how data science and informatics work with and support the needs of cancer researchers. Finally, NCI can support the creation of short-term courses such as those offered by Cold Spring Harbor Labs, where researchers attend for an intensive, off-site training that lasts two or three weeks, which supports basic, core skills for more experienced investigators. Researchers in this category may not have the time or appetite for a long-term course or to pursue an additional degree, but recognize the need to understand data science and informatics as critical to collaborating with computational colleagues and to developing new knowledge from the data they generate.

5. **Create training grants to support cross-domain, "hybrid" researchers**: NCI should provide specialized grants to support the cross-domain/cross-skill training. The recommendations above are designed to allow for transition from computational science to cancer research, to provide training for computational scientists to work in the field of cancer, or to train cancer researchers to use the tools needed to work with data scientists. The "hybrid" grants and programs would focus on the individuals who wish to straddle both fields as a career path, acting as the "glue" between the two domains and supporting the necessary collaboration between the two. These could include grants to institutions who are offering dual track programs, or grants to allow individuals already trained in one domain (biomedical or data science) to pursue additional training in the other. Stanford's Cancer Systems Biology Scholars Program could be a model for this work. NCI can support broad communication and education to academic institutions about supporting the transition of traditional biomedical researchers into data science, which can be intimidating for many. NCI can support the community in developing a clear path for advancement as well as clear means to acknowledge the contributions of hybrid-scientists. These could include inclusion in publications, specialized grants related to sharing and analysis of data.

6. **Partner with organizations providing training for K – 12 and undergraduates:** Organizations such as NSF offer general data science training, which could be leveraged to introduce younger audiences to the domain. Some institutions are also offering internships for high school and undergraduate students to work within a computational lab doing research that is useful to the investigators while providing background and skills that will inform students' choices of majors and careers. NCI could partner with NSF and other institutions to provide the cancer-specific and biology training that would enhance the informatics and data science expertise young people can develop in the more general training programs.

Conclusion

The need to build the workforce equipped to handle the diversity and wealth of data supporting cancer research is well recognized and cannot be understated. We believe the recommendations outlined both here and in the interim report lay the needed groundwork for NCI to support needs of the cancer research workforce in data science training and workforce development.

Machine Learning

Introduction

Cancer care has been transformed by targeted therapies, immuno-oncology and more highly personalized care. This transformation has been enabled by greater access to data and analytics and has driven interest in more and new data. In addition to more traditional clinical, registry and trial data, today's cancer research is informed by genomics and specific mutations, changes in gene expression at the RNA and protein levels, and metabolomics. This enables population risk predictions as well as specific treatment recommendations for individual patients. Integrating these different levels of data from millions of data sets into coherent models of cell-to-population states is beyond current bioinformatics tools. For cellular models, data science could enable and speed the next generation of cancer care through the development of algorithms that integrate results from genomic and chromatin, nascent RNA, mRNA and non-coding RNA, and protein that give statistical models of the relationship between coordinated changes in cells upon perturbation. At the population level, data science could help encode *in silico* populations to improve risk predictions and understand complex questions that are not well suited to addressing with the current clinical trial framework, such as treatment sequencing and combinations. Machine learning and artificial intelligence (AI) can improve all areas of cancer research, from modelling cellular function, to predicting response in a given population, to identifying new areas of potential development.

Recommendation Summary

Core clinical questions concerning diagnosis, treatment personalization, and understanding disease progression are all prediction problems. Therefore, powerful machine learning methods that can leverage large volumes of patient data have the potential to be accurate beyond the abilities of human physicians, to reduce the experiential bias that is part of human learning and training, and to reduce the outcome uncertainty inherent in most of today's medical interventions. To achieve these goals, this working group recommends that NCI supports development and integration of machine learning methods that are contextualized to the unique needs of cancer research and care. This requires curating large, diverse datasets that will enable this research, creating new funding opportunities dedicated to the field of machine learning, and developing outreach to the machine learning and data science community who currently lack significant opportunities to contribute to cancer research. Lack of engagement with the machine learning community is a missed opportunity for cancer research, which could ultimately impact patient care. The need for engaging machine learning experts has been further emphasized by the data science training recommendation in the call to support and engage computational scientist.

Background

In the last decade, machine learning has transformed multiple areas of science and engineering, significantly reshaping the industrial landscape. This progress has been fueled by the advancement of deep learning algorithms, access to large amounts of data, and fast computing capacity. Despite its wide penetration in many areas of our lives, this powerful technology has had limited impact on healthcare, and specifically, on cancer research and clinical care.

According to the Healthcare Information and Management Systems Society, only 5% of hospital providers in the United States currently utilize some form of AI technology. This statistic is representative of the current state of the art worldwide. Despite large amounts of data collected in the healthcare system, today this information is severely underutilized and often unavailable in a way that facilitates the application of AI. The American Society of Clinical Oncology acknowledges that all the clinical decisions today are based on the 3% of the population that participates in clinical trials. While cancer registries serve as a primary source for retrospective population-level studies, they are still manually curated and only partially capture patient information needed for large-scale machine learning.

Nevertheless, despite the slow start, machine learning methods have already demonstrated significant promise in both improving clinical applications and deepening understanding of cancer biology. Today, there are clear successes in the application of computer vision to reading diagnostic images, analyzing pathology slides, and understanding cell-level interactions. On the natural language processing side, most approaches focus on extracting patient information from electronic health records (EHRs), mining scientific literature, and analyzing patient-related social media. Similar progress has been achieved using genetic data to stratify patients in terms of their risk and response. Still, most of these efforts are limited to academic research. Only a small portion of this research is sufficiently mature to be translated into the clinic.

There are several barriers that inhibit the penetration of machine learning technology into cancer research:

1. **Access to data:** The performance of machine learning algorithms is a function of the training data, where both the quantity and the quality of the data drive the utility and generalizability of resulting algorithms. The lack of large, publicly available datasets in the clinical area has been a significant barrier for algorithmic development. For the vast majority of machine learning researchers, the data is not available as it currently resides in specific medical institutions. Moreover, the lack of standard datasets prevents benchmarking and reproducibility, a cornerstone of empirical research.
2. **Algorithmic gaps:** While many problems in cancer research and care can be solved using existing machine learning methods, there are several areas where the technology lags behind the needs. Examples include training from small data, sustaining robust performance in the presence of the distributional shift (e.g., population diversity), and developing interpretable deep learning models.
3. **Challenges in clinical integration:** Most research on cancer/AI focuses on method development rather than on studying mechanisms for their implementation within existing clinical pipelines. The latter question is challenging in its own right. On the one hand, while AI algorithms provide useful solutions, their predictions are not perfect; they can be brittle and hard to interpret. On the other hand, today's medical workforce is not familiar with this technology and may be reluctant to embrace it. The current gap between AI developers and their clinical users is a barrier for large-scale adoption.
4. **Burden of manual data curation:** Today, much of the data curation for retrospective studies, population health, and clinical trials supported by NCI is done manually. This reliance on manual labor significantly increases the cost of the studies, limits their

scope, and slows down research progress. Despite the broad usage of natural language processing (NLP) tools for similar purposes in other fields, NCI investigators are still not utilizing them routinely.

5. **Funding strategy:** Today, NCI lacks a funding program dedicated to AI. Much of the machine learning relevant work supported by NCI is funded through the Informatics Technology for Cancer Research program (ITCR) and programs in the Division of Cancer Biology. Machine learning is not a key component of these programs but rather embedded in the scientific questions and drivers of the overall programs, so the research typically focuses on applying existing machine learning methods rather than developing new ones. Funding of machine learning research is further complicated by the fact that current NIH and NCI review panels generally do not have technical machine learning expertise to review proposals in this area, as measured by their publication record in machine learning conferences and journals. As a result, very few top machine learning researchers contribute to the NCI agenda since it is challenging to identify appropriate funding opportunities and to ensure appropriate review.

Recommendations

1. **Develop targeted machine learning methodology for cancer research and care:** While the NCI community can directly benefit from ongoing advances in deep learning research, there is a pressing need to develop customized technologies. For instance, models developed in the context of cancer biology would ideally utilize rich mechanistic knowledge about the underlying biological processing. Infusing human knowledge in a principled way is likely to improve model accuracy and decrease its dependence for large amounts of annotated data, which may not be available. Conversely, designing interpretable models can shed new light on the underlying biological mechanisms. Existing machine learning machinery is not sufficiently developed to address this need, therefore further research on the intersection of deep learning and cancer biology is required. Other areas of priority include:
    a. Transfer learning (across diverse populations, diseases, healthcare system);
    b. Biological modeling through integration of multiple datasets;
    c. Longitudinal modeling of disease progression over time;
    d. Algorithms for processing multimodal patient data (images, genomic, sequences, text);
    e. Robust methodology for confidence estimation;
    f. White-box architectures for deep learning, contextualized in the context of cancer research and care.

2. **Compile large, diverse datasets for training of machine learning algorithms:** To facilitate rapid algorithmic developments and support reproducible science, we recommend curating large datasets in key areas of cancer research. Specific areas of interest include diagnostic imaging, pathology slides, de-identified EHR data, and mRNA expression. In addition to uni-modal data, we recommend curating datasets that contain patient data in multiple of the modalities summarized above. This could be achieved in part through our interim recommendation: Investments to Leapfrog Data Sharing for

High-Value Datasets (Appendix A). Special effort should be taken to ensure diversity of the collected data — across populations, healthcare centers, etc. Ideally, these datasets will be augmented with outcomes data and patient reported data, supporting the use of data for multiple learning tasks.

3. **Incorporate principles of AI ethics into machine learning research at NCI**: The ethics of using AI to diagnose and treat patients with cancer is complex and largely unexplored. Multiple studies in other scientific fields have demonstrated that machine learning algorithms can be biased and systematically underperform for certain populations. This bias may result either from the way the data was collected and/or the way the machine learning model was developed and trained. We recommend that NCI research be informed by the best practices in this area established in the broader AI community. Some of these principles are already part of the NIH guidelines for the inclusion of women and minorities as subjects in research, but these guidelines have to be expanded in light of the evolving understanding of AI ethics. The potential impact of these biases on diagnoses and treatment, and any unintended consequences, should be evaluated by a group of relevant stakeholders. The group should also make recommendations on how to prevent bias and to avoid any unintended consequences.

4. **Build machine learning infrastructure for drug discovery:** Recent progress in applications of deep learning for modeling molecules and their properties opens exciting possibilities of using these methods for drug discovery. This includes generation of new molecules, safety profiling, prediction of drug interactions, etc. When combined with other applications, including modeling of biological processes at the cellular level and modeling in silico populations, these approaches have potential to transform the design of novel cancer therapies for personalized treatment. We recommend NCI dedicate resources to this important new area in cancer research.

5. **Support automation of data curation:** Effective data utilization is crucial for the development of clinical AI. To reduce the cost and delays associated with manual data curation, we recommend NCI encourage investigators to employ and validate NLP tools for the automation of data collection. Such tools include information extraction of patient characteristics from EHRs, document classification, and data retrieval. Broad utilization of these tools would also inform further development of automatic curation methodology, affecting NCI priorities in this area. As a first step, the NCI should convene a group of stakeholders to understand the current issues that are preventing researchers from using NLP more broadly and explore how data models can be improved to enhance data validity, as further expanded upon in the subsequent recommendation regarding real world data.

6. **Explore research on effective translation of machine learning methodologies into clinical care:** Successful adoption of machine learning into cancer care goes far beyond algorithmic development. This involves finding effective mechanisms for clinical integration, educating medical personnel about the strengths and weaknesses of the

technology, and quantitatively assessing its benefits in terms of clinical outcomes, patient experience, and cost. These issues have to be systematically studied to develop effective integration strategies, and to collect feedback which will inform the development of the next generation of machine learning algorithms.

7. **Develop new funding opportunities for machine learning research:** Given the increasing importance of machine learning methodologies in cancer research, we propose developing a program dedicated to the development of machine learning approaches contextualized in cancer research and care. The program should address areas of priority identified above, along with other areas of machine learning research relevant to NCI. To make the program successful, it is crucial to identify reviewers who have technical expertise in this area (as measured by their publications in top machine learning venues).
8. **Attract a broad machine learning community to contribute to cancer research**: Since machine learning is not a traditional area of expertise for NCI, it is important to reach out to a broader data science and machine learning community to contribute to cancer research. These efforts may include clearly publicizing funding opportunities for this area on the NCI website, providing access to data sources, creating links between cancer researchers and machine learning researchers, and the specific recommendations to support and engage computational scientists highlighted in the earlier training recommendation.

Conclusion

AI and machine learning have the potential to transform cancer research and ultimately the care and outcomes of patients. In this recommendation, we have outlined the important steps for NCI take in paving the path forward to fully leverage the potential of machine learning and AI.

Real World Data

Introduction

Historically, advances in cancer diagnostics and therapeutics have been achieved by learning from the experiences of less than 5% of cancer patients who participate in clinical research. This approach has successfully led to a 27% decrease in the cancer death rate since 1991. Imagine the impact on cancer patient outcomes if we could learn from the experiences of every patient. Real World Data (RWD) implemented in the context of a Learning Healthcare System presents such an opportunity. Yet despite the promise of RWD, several barriers remain to broader application of RWD. This report outlines recommendations to the NCI for the creation of validated frameworks for RWD and its associated metadata, as well as the demonstration of their utility in Learning Healthcare System reference implementation projects.

RWD is patient health data collected outside of a clinical trial setting, and can come from a variety of sources, including Electronic Health Records (EHR), claims and billing, disease registries, patient-reported information, and patient-generated data such as biosensors and mobile devices. According to the FDA, RWD "holds potential to allow us to better design and conduct clinical trials and studies in the health care setting to answer questions previously thought infeasible."[4] RWD can provide valuable insights into large and small populations in the areas of understanding of treatment pathways in diagnosis and efficiency of treatment, as well as drug safety, and can help in the planning of clinical trials to extend eligibility criteria to a broader cohort. Most significantly, RWD can support a Learning Healthcare System, where data from all available sources contribute to a broad understanding of disease, treatment, patient behavior, and outcomes.

The FDA is investing significantly into RWD and the Real World Evidence (RWE) that can be derived from it, creating a multifaceted program that includes "demonstration projects, stakeholder engagement, internal processes to bring senior leadership input into the evaluation of RWE and promote shared learning and consistency in applying the framework, and guidance documents to assist developers interested in using real-world data (RWD) to develop RWE to support Agency regulatory decisions."[5] In April 2019, the FDA approved the first drug label expansion using only RWD as RWE of the safety and efficacy of palbociclib in men with hormone receptor positive metastatic breast cancer, a rare disease compared to woman with breast cancer.

Singh et. al. reviewed the use of RWD in a variety of publications. They found evidence that "novel uses of RWD can foster new understandings of disease associations and or comorbidities that would be particularly useful when trying to target new populations or indications for research" and "examples where the use of large datasets created novel approaches to

[4] "Real World Data." *FDA.* https://www.fda.gov/scienceresearch/specialtopics/realworldevidence/default.htm

[5] "Framework for FDA's Real World Evidence Program." *FDA*, December, 2018. https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf

stratification of patient populations."[6] Their research also showed the use of large datasets enabled researchers "to sift for rare indications, and develop unique algorithms to find therapeutic 'diamonds in the rough', as well as uncover previously missed or early indications of disease incidents that might have previously been undetectable."

Epidemiologist Cathy Critchlow, vice president and head of Amgen's Center for Observational Research, says, "there are important clinical questions that can't be answered by randomized clinical trials and crucial data available from patients in clinical practice settings. Real-world data won't supplant clinical trials or prospective observational studies…but it will play an important role in complementing them."[7] But, she adds, "It will be important to work on processes and methods that increase the validity of such evidence. We need to build consensus on topics such as data quality." According to Peter Embi, MD, president of the Regenstrief Institute, the idea that research and care are distinct activities needs to be challenged. "We need to leverage everyday healthcare activities to inform and generate evidence, and then use that evidence to improve care for individuals in the future. We need to align the way we implement and use our technology, how we collect information, and how we apply that information back to the system to learn."[8]

The NCI is uniquely positioned to collaborate across stakeholders including the FDA, industry, healthcare delivery systems, researchers and patients to accelerate the opportunities for RWD and RWE in cancer research and care. The following recommendations comprise several foundational activities that can contribute to addressing some of the challenges related to using RWD to create a learning healthcare system.

Recommendation Summary

As RWD and RWE are used increasingly for research, regulatory, and quality purposes, it is critical for the NCI to create a framework to enable cancer researchers and care teams to effectively and appropriately use RWD. The framework should seek to define high quality data for specific cancer research and clinical needs and to ensure data is treated consistently across treatment centers or that differences in care are noted in metadata for future evaluation. Development of a RWD/RWE framework for cancer research and care will require convening stakeholders to develop guidance and criteria for evaluating and populating key clinical concepts, particularly those unique to cancer (e.g., progression, recurrence, functional status, etc). Identifying, evaluating, and validating these key clinical concepts will be fundamental to being able to consistently capture RWD in cancer patients. These stakeholders should also develop a set of criteria for extracting data from RWD sources, such as the PRISSMM criteria

[6] Singh, et. al. "Real world big data for clinical research and drug development." *Drug Discovery Today*, March, 2018. https://www.sciencedirect.com/science/article/pii/S1359644617305950

[7] "The Future of Real World Data in Clinical Trials." *Clinical Informatics News, January 11, 2019.* http://www.clinicalinformaticsnews.com/2019/01/11/the-future-of-real-world-data-in-clinical-trials.aspx

[8] Kent, Jessica. *Health IT Analytics.* "Learning Health Systems, Big Data Key to Improving Outcomes," October 10, 2018. https://healthitanalytics.com/news/learning-health-systems-big-data-key-to-improving-outcomes

currently being developed, best practices for using RWD, and a metadata model for describing the completeness, quality, and interoperability of these RWD sets. This metadata model should be informed from current clinical care standards that may impact the quality and interoperability of the data as well as support future data consent needs. Finally, to demonstrate the utility and generalizability of the RWD framework, NCI should fund efforts to demonstrate reference implementations of a Learning Healthcare System that utilizes the developed RWD framework.

Background

RWD and RWE present a unique opportunity to learn from the experiences of every cancer patient to accelerate discovery and improve the quality of care. However, several barriers continue to inhibit the use of RWD/RWE in cancer research and care.

Lack of transparency regarding the quality, completeness, interoperability, and curation process for existing public and private RWD sources

There are many disparate groups investing in RWD data platforms for oncology which are being increasingly promoted to provide RWE. However, few of these platforms have transparent data models or data normalization/curation practices. There are no agreed upon standards for describing these data with regard to representativeness or generalizability of their populations or the underlying sources of information. There are also no commonly accepted metadata standards that improve the comparability or interoperability across data types or different sites of care.

Metadata can improve the usability of any dataset, but few metadata models have been constructed with the goal of RWD/RWE in mind. Metadata, data that describes other data, can support RWD in several ways. For example, a data element may be a patient’s blood pressure value. The metadata could include things like the position of the patient (standing, sitting, lying), the position of the blood pressure cuff (arm, forearm, thigh, leg), the size of the blood pressure cuff, the size of the patient’s arm (or site of measurement), the number in the sequence of blood pressure measurement, the method of blood pressure measure (manual, automated), and so on. The metadata helps us understand the unique data element more completely and ensures that we can tie it back to the clinical context in which the data was collected. A metadata model can support RWD/RWE in several ways. It can help clarify areas where clinical care variation may occur (such as the blood pressure example) which improves our understanding of the clinical context. Next, metadata can help us understand features of the data (e.g., number of staging events) to ensure that combined data sets treat data elements the same way. Metadata can also be used to understand how the data element can be used by attaching patient consent for various uses at the element level. A metadata knowledge model can also provide the necessary context for computers to interpret the structure of the data elements, which will ultimately be necessary to leverage such RWD for development of algorithms that can support a learning healthcare system. There are many more potential uses for metadata, and the lack of a metadata model for oncology RWD will limit our ability to fully recognize the potential of RWD in the future.

Lack of validated data models and best practices for manual curation, automated extraction, and use of RWD

There are several methodology frameworks or guidance documents for using RWD (e.g., ISPOR, National Academy of Sciences, Sox's work in comparative effectiveness research, PCORI). However, to date, there has not been parallel agreement or investment in any frameworks for RWD outside of methodology. Data directly impacts the study design and methods, and there is a current lack of frameworks or guidance around the critical elements of RWD which impact study designs and meaningful use.

There are several ongoing efforts to develop and validate data models and curation standards for RWD. Two key oncology-focused efforts, mCODE (minimal Common Oncology Data Elements) and PRISSMM, are taking first steps at defining standards for RWD in cancer. mCODE is assembling a core set of structured data elements for oncology EHRs and is planning pilot implementation projects to evaluate the generalizability and completeness of the models and demonstrate utility. PRISSMM is developing a standard taxonomy for classification and communication of structured information about cancer status and treatment outcomes including evaluation of the reproducibility of standard operating procedures for curation across multiple institutions. The manually curated data linked to primary EHR source documents can further be used to train and evaluate automated methods to extract the data. The current lack of publicly available training sets for such use cases limits access to the data science community in further developing such systems.

Limitations in the use of EHRs as a source for RWD

EHRs are often viewed as a ubiquitous source of RWD, but we must recognize that there are limitations to the data that can be extracted from EHRs.

Variance in oncology EHR implementations. Multiple EHR vendors provide support for the oncology community. Clinical standards for how data are captured in EHRs are lacking, resulting in each EHR vendor developing their own approach to documentation and workflow, which often vary even among implementations of the same EHR system. Different implementations of the EHR often capture desired information, such as cancer staging information, in different places in the EHR, sometimes in structured forms, sometimes in unstructured text. Frequency standards for time-varying data elements, such as disease status, are not defined. There is a lack of understanding around metadata for uniquely oncological data (e.g., time between biopsy and processing for tissue samples, type of assay used for genomics studies, number of genomics analysis run per tissue sample). As such, there is uncertain validity and reliability of the data captured in EHRs, even when captured in structured data fields. Finally, the development and implementation of clinical documentation standards generally does not include consideration for clinical workflows, nor does it involve the secondary users of the data (i.e., researchers). These large variances in oncology EHR implementations present real and practical challenges for RWD/RWE use cases.

Incompleteness of data inherent to the cancer patient journey across multiple locations. At best, data residing inside a local EHR is a partial representation of the cancer patient's journey.

Within a local EHR, data may be missing due to migration from legacy EHR systems. Furthermore, an estimated 20% - 30% of oncology patients migrate between clinics and across EHRs during their clinical care. Claims data have demonstrated how often people change payers, but people also move residences or have multiple homes and seek other opinions. The documented “snowbird” or “sunbird” populations of retirees will impact RWD for oncology as these highly mobile populations move between states and health systems. Exchange of information among the EHR instances of these different care settings is lacking, but is required in order to fully capture a patient’s care experience. Conversations reintroducing the concept of personal health records (PHRs) are once again surfacing, and may represent an opportunity to empower patients to control and own their own health data and ultimately decide whether and how to contribute it for research purposes. In addition, PHRs may provide patients more control and transparency over how their data is used. As more types of data are integrated, the patient’s privacy may be negatively impacted. PHRs offer the potential for patients to grant permission for uses they deem appropriate and beneficial.

In addition, while many of the efforts around RWD today focus solely on the EHRs, we must also recognize that there is a wealth of RWD beyond the EHR. EHRs are limited in only capturing information from clinical encounters; the time periods between these clinical encounters and the context and environment in which patients live also serve as important covariates in the determination of health status. More and more, data from the lived experience of patients is not captured in the EHR. RWD should seek to include data from biomonitoring, sensors, home monitoring, and other systems to better understand the impact of treatment on patients and their caregivers. Claims and registry data are also valuable sources of RWD. Recently proposed federal laws against data blocking, and the increasing importance of patient advocacy groups driving research and policy mean that patient-generated or patient-collected data will be of increasing importance. Frameworks or guidance on standards should consider the broader landscape of current and future potential RWD sources.

Lack of incentive alignment for point of care collection of high quality RWD

Finally, there are few incentives for clinical staff to ensure complete and accurate data capture at the point of care. Clinical billing remains the main driver for clinical documentation practices. Accreditation standards and external quality reporting have influenced documentation practices, but are often downstream from point of care impacts, challenging their adoption and sustainability. Most RWD/RWE use cases are far downstream from clinical care. Ultimately, what we lack is a true Learning Healthcare System, in which the data from all aspects of a patient’s journey is leveraged and used to improve both the care of the patient being seen today, as well as the care of the patient being seen tomorrow. The incentives of the clinical care team for high quality collection of RWD must be aligned with the incentives for high-quality, real-time management of the local cancer patient population. RWD is the foundation for developing Learning Healthcare Systems that enable measurement and management of the quality of cancer care in real time to improve outcomes. These can be implemented as systems for real-time quality improvement or as pragmatic studies comparing interventional approaches.

Recommendations

While much work remains to develop a worldwide Learning Healthcare System for cancer leveraging RWD, we believe that the following recommendations present immediate and directed opportunities for the NCI to assist in accelerating the path towards that vision.

1. **Convene a group of stakeholders around a RWD metadata framework.** The NCI is well positioned to facilitate convening a group of stakeholders to address key issues of a metadata model framework for RWD in cancer. We recommend that this group define components of RWD metamodel with the goals of improving the link to clinical context, to improve the interchangeability of data elements, and to understand how data elements are being used and should be used. This should include constructing RWD metadata model(s) for oncology specific uses, considering clinical care standards, randomized controlled clinical trial standards, and unique issues around commonly used data elements. These RWD metadata model(s) should be tested to understand their validity, reliability, and generalizability. Furthermore, the NCI should identify both intramural and extramural initiatives that should deploy the RWD metadata frameworks, and advocate for the use of RWD metadata model(s) by other groups who work with RWD.

2. **Develop a framework and criteria for evaluating and populating key concepts from EHRs and other key sources of RWD.** There are several methodology frameworks or guidance documents for using RWD. However, to date there has not been parallel agreement or investment in any frameworks for RWD outside of methodology. Because data directly impact study design and methods, we recommend that NCI convene stakeholders to develop a framework or guidance around the critical elements of RWD which impact study designs and meaningful use. Such a framework should build off of existing efforts such as PRISSMM, and existing candidate frameworks should be evaluated to determine suitability and identify gaps. Finally, NCI should appropriately fund the steps needed to achieve a RWD framework, including development, validation, evaluation, and adoption. While not exhaustive, critical elements for RWD for inclusion in an RWE framework include the following:
    a. *Selection and generalizability of study population:* Provide a framework or defined guidance to enable systematic comparisons between populations drawn from different sources of RWD. Develop recommendation for how to describe cohorts developed from RWD relative to source populations or relevant diseased population.
    b. *Exposure and outcomes:* The framework should provide guidance on how exposures and outcomes need to be described and align on expectations around transparent reporting of data quality elements (e.g., provenance, conformance, validity, etc).
    c. *Key confounders:* The framework should provide guidance on how investigators can evaluate and describe the many data elements available for confounding control and how to justify selection or development of key variables for the specific question. It should also provide some guidance on how to assess and

transparently report missing confounders and how to describe how the study design or methodology will be assessing the impact.

   d. *Content and training.* RWD are primarily collected for other purposes and defined by other use cases which impact their utility for re-use. Ideally the framework would identify ways in which to describe this context and guidance on how to facilitate training/knowledge transfer regarding the underlying processes and mechanisms.
   e. *Agility and time-varying components.* RW datasets are not static and are constantly changing with regarding to patients, content, and structure due to rapid changes and increased investments in technology. The framework needs to be sensitive and responsive to this reality.

3. **Demonstrate the utility of RWD in a series of Learning Healthcare System reference implementations.** Reference implementations of Learning Healthcare Systems that demonstrate the utility of RWD frameworks to support clinical research and continuous clinical quality improvement at the point of care are needed in order to identify the practical strengths and limitations of these RWD frameworks. It is essential that these RWD frameworks continuously evolve as their utility and generalizability is evaluated in practical applications. Given the large variance in documentation standards of oncology EHR vendors, evaluation of the generalizability of the RWD frameworks needs include implementations that span EHR vendors and hospital instances of the same EHR vendor. Use cases both intramural and extramural to the NCI are needed to evaluate the implementation of the RWD framework. Use cases could include but are not limited to:
   a. Creation of a set of reusable clinical trial eligibility criteria (disease and disease state specific) using RWD with the goals of improving the percentage of the cancer patient population who are eligible for studies, and improving the successful accrual and completion of clinical trials.
   b. Use RWD as paired (or partially paired) arm in a clinical trial through NCI's National Clinical Trial Network (NCTN) or the NCI Community Oncology Research Program (NCORP) (e.g., Friends of Cancer Research).
   c. Demonstrate the use of RWD in pragmatic clinical trials. Examples include but are not limited to pragmatic clinical trials that aim to decrease cancer disparities, improve the quality of standard of care treatment, decrease toxicities and complications associated with treatment, or decrease costs of care.
   d. Demonstrate the use of RWD to improve the quality of cancer care and clinical research efficiency using real time decision support and/or population management.

Conclusion

RWD/RWE has the potential to significantly impact the pace clinical research, and transform the daily management of patients with cancer to achieve better outcomes. Significant challenges remain to achieving this vision for a Learning Healthcare System for cancer. We believe these recommendations provide discrete and actionable steps towards improving the quality, transparency, and best practices for use of RW data sources and demonstrating the impact of their utilization to advance cancer research and care.

Data Sharing

Data Sharing Statement

We hold that for research that is publicly funded or funded by nonprofit research foundations, data should be broadly and routinely shared. Data sharing improves the scientific process through independent verification, decreases barriers to current knowledge, improves data reuse, uptake, and promotes reproducible science. Data, code, pipelines and other sharable information that promotes reuse and scientific reproducibility all need to be shared as part of the research process. Conforming to the FAIR principles and releasing data and code in a FAIR-compliant resource is crucial for continued innovation and efficient use of research funding.

Value of Data Sharing

In order to maximize the utility and value of data sharing, data needs to be well characterized, collected consistently through a well-documented, validated process with all data elements and methods documented and recorded. Using well-defined, common data elements, data dictionaries, and ontologies greatly aids in creating sharable, reusable data. Having a well-described protocol governing data collection is perhaps the most important component of data sharing, affecting data quality, consistency, and usability. The use of standardized collection protocols and data definitions enables future use and reuse. Measuring conformance with the protocol is also important. For instance, for clinical research demonstrating compliance with the protocol from consent to biorepository collection to outcome reporting is part of the standard clinical research process. Consistent data collection also needs to incorporate disparate types of data that can be collected for research and clinical care, including clinical, imaging, pathology, genomic, proteomic, and outcome data. We should strive to create and adopt protocols that put the patient first in decision-making and respect, honor, and adhere to their desires for sharing their data throughout the process. In addition, we should honor their gift of data by shepherding it through the process consistently and with expediency.

When data is collected consistently, with well-defined elements, researchers can use the data with confidence and utilization rates increase. When they query the data, they will be assured of dependable quality that will allow them to reproduce existing studies to disprove or confirm, further interrogate data around those studies for novel discoveries, and combine the data in new ways for enhanced understanding beyond the original collection's scope. It makes the entire scientific process more reliable. As pointed out in a letter to the editor in Nature Genetics, "Journals, although they conduct peer review, do not validate each experimental result or claim."[9] Well-annotated data enables detailed documentation and validation.

In addition, high-quality data can be leveraged to flip our existing academic model and be a reward within itself. Researchers of rare cancers don't see enough cases at their institutions to

[9] https://www.nature.com/articles/ng.3830.epdf?author_access_token=FpxejEzLz7hhQzjNXOww-9RgN0jAjWel9jnR3ZoTv0PfV-r4OPX6-oeq4sefK7lcN5bVG1B5qFB7_k_8fISqClZXqTAZMow3grVUktzXNV2dQ8x4b0-1was8C4MQCO_t

use big data analytics on their data, but by combining data, they will gain greater insight to their work as well as identify similar patients. This is the centerpiece of the Undiagnosed Disease Network and as we apply precision medicine to cancer, cancer takes on attributes of rare disease. Furthermore, institutions like the NCI could reward the positive behavior of timely data sharing instead of just at time of publication, making data sharing a more attributable and credited activity.

And of course, high-value, well-annotated, standardized, and well-characterized cancer data sets may be generated outside of NCI-funded research. Enabling the generation of high-quality datasets by other funders and by commercial interests that conform to these same rigorous standards is of value for all cancer researchers, and to the extent feasible, these data should be brought into the same, consistent data sharing framework. Data from BeatAML (acute myeloid leukemia), the MMRF CoMMpass (Multiple Myeloma Research Foundation Relating Clinical Outcomes in Multiple Myeloma to Personal Assessment of Genetic Profile) study and Foundation Medicine have all been made available through the NCI Genomic Data Commons, and these activities are to be commended and extended.

Incentives for Data Sharing

For data sharing to be effective and ubiquitous, there will need to be both positive incentives that encourage and reward data sharing and penalties for projects and individuals unwilling to participate in meaningful and appropriate data sharing. Some of the positive incentives can include citations for datasets that have been shared, consistent and well validated, recognized and valued mechanisms for attribution and measurement of dataset, code, and pipeline usage - perhaps leading to an ‘S-factor’ for measuring the type and quality of data sharing attributed to an individual researcher. Support for data management, curation, vocabularies, data models, and annotation is necessary to create high-quality, useful datasets and code that can contribute to generalized knowledge and reproducible science. Having a reasonable, measurable, and clear data sharing plan for all research projects is vital. Support for and use of standardized vocabularies like the NCI Thesaurus, the OBO (Open Biological and Biomedical Ontology) Foundry ontologies, and resources like PhenoPackets are to be encouraged. Ensuring that research funding includes dedicated budgets for data management, curation, mapping and submission of research data and results into appropriate resources needs to be mandatory for all federally funded research. Penalties for a poorly written or hard to measure data sharing plan should include reduced reviewer enthusiasm for new proposals. Likewise, for seasoned investigators, being able to measure and confirm conformance with past data sharing plans, including data completeness, quality, and uptake by the research community, will result in improved data sharing and create an expectation (and peer enforcement) that data sharing is an important, valued, and supported activity in biomedical research.

Recommendations to Ease the Barriers to Data Sharing and Access

To encourage the cultural shift towards data sharing, it is essential that NCI do as much as possible to ease the burden on investigators by addressing barriers and streamlining the process for both sharing and accessing data as much as possible. There are several areas in which NCI can begin the work of moving all toward a culture of inherent data sharing.

Consent/Regulatory Issues

Issues around consent and regulations may impede data sharing. Consent forms differ on what types of data can be shared, and what types of downstream research can be done. These differences can lead to challenges around data sharing and aggregating data from multiple sources. Consent needs to be consistent with and enable broad use. Solving important problems in cancer research and applying data to maximizing the benefit and creating positive outcomes for cancer patients is very much aligned with the wishes of participants in cancer research. Respecting and realizing these goals is crucial. Toward that end, defining what can easily be shared, what is sharable as a limited dataset, and what requires detailed permission to access is critical for maximizing the utility of cancer data particularly for new applications of data science like deep learning. Consistent consent and simplifying consent documents to increase participant understanding and decrease both patient and care team burden must continue as a shared goal for all cancer researchers. Providing information back to participants about their personal trajectory, the trajectory of their disease, and incorporating patient reported outcome data to provide feedback to patients about their cancer journey are all crucial for next generation clinical research.

> ***Recommendation****:* NCI should convene stakeholders to develop best practices for consent that enable data reuse; this would be helpful to investigators in developing data sharing language for consent forms. This group should also define and broadly disseminate examples of simplified, robust, and acceptable consent documents, whether as templates or boilerplate language. Likewise, new modes of consent and sharing such as those used by the Global Alliance for Genomics and Health (GA4GH) that focus on the 'rights of individuals to participate in and benefit from research' should be evaluated and incorporated into NCI guidance. In addition to broad, consistent, simplified consent, dynamic consent that engages patients and their families in cancer research should be examined and encouraged. Finally, providing a consistent set of guidelines for how and when to recontact research participants with clinical actionable recommendations is very much needed and collaboration with the NIH ELSI (ethical, legal, and social implications) recommendations would benefit all research participants.

Data Sharing Requirements

Funders are still discerning best policies and practices for data sharing, particularly since strong evidence regarding best practices for data sharing is lacking. While there is a need for funders to try different approaches to push the concept of data sharing forward as we collectively drive to make data sharing the norm, this results in a multitude of data sharing policies that may differ based on the funder, leading to confusion and increased administrative work for investigators.

> ***Recommendation****:* Consistent recommendations, requirements and examples for good data sharing plans and mechanisms for data sharing across NCI and NIH institutes would simplify requirements for investigators as well as the NIH and facilitate data sharing. Providing mechanisms for machine readable consents, machine readable data sharing

plans, and validated, well described APIs for data submission would standardize and simplify data sharing and reduce the burden on both funders and investigators.

Funding/Resource Needs

Requiring data sharing without explicitly recognizing and funding crucial activities such as data management, data cleaning, quality control, and adequate and standardized annotation and documentation is not sustainable. Likewise, including infrastructure costs for on-premise or in-the-cloud data storage and compute needs to be an anticipated and supported cost. Delays between the time data is uploaded and data is available for use by the community presents challenges that may discourage investigators from sharing data. Ensuring that data repositories have sufficient resources and staff to easily accept and share data in timely way would help to alleviate these barriers.

> ***Recommendation****:* Budgets and budget guidance need to include funding guidelines and expectations that data management, data cleaning, quality control, adequate and standardized annotation, and documentation are all written into the budget and clarify the guidance for justifying these activities.

Education and training

There is a role for institutions and funders in educating investigators and other staff regarding data sharing policies, how to comply, and how to contribute to data sharing resources. Data sharing requires resources and expertise that many investigators don't have. Needs include resources for infrastructure, including a secure computing environment, as well as data cleaning, quality control, data management, and documentation. This also includes the need for team members with expertise in data cleaning, quality control, and data management, which may require hiring additional staff and/or training existing staff. Finally, support for education and training on how to use and contribute to data sharing resources like the NCI Cancer Research Data Commons must also be developed and provided to create the next generation of cancer researchers who can efficiently and effectively contribute to and benefit from these resources. Additional specialized topics related to data sharing may also benefit from additional training and guidance, for example, regarding proprietary data. Investigators using clinical trial data from pharmaceutical companies would benefit from additional training in how to best adhere to restrictions around proprietary data while still being able to share data.

> ***Recommendation****:* To provide the needed support to investigators, NCI should develop training regarding data management processes and policies. For example, NCI could develop a training module to teach investigators about data sharing policies and how to appropriately share data, including topics such as:
>
> - What data is expected to be shared
> - Reputable repositories for sharing data
> - Tools and methods for sharing data
> - Sharing data safely (e.g., appropriate de-identification and protection of privacy)
> - Appropriate consent for downstream use of data
> - Appropriate annotation of data (e.g., metadata, standardized data models)

- Short- and long-term data management

Recognition of Data Sharing

Career advancement is a concern especially for early career investigators, and the traditional research culture has fostered a fear that data sharing will enable scientific competitors to use data before primary investigators generate publications that are valued in grant applications and tenure consideration. The whole community will need to work together to shift the culture regarding data sharing, requiring equal commitment from institutions along with funders such as the NCI. Together, funders and institutions should develop and implement systems that appropriately credit investigators for sharing data. A critical component of such a system is funders and institutions considering shared data as a valuable research product (e.g., giving shared data equal weight as publications in tenure consideration or grant applications). NCI may be able to start such a conversation with the cancer centers.

> ***Recommendation****:* NCI should encourage systems that credit investigators for data sharing. Systems that appropriately provide attribution to investigators for sharing data might include:
>
> - Establishing digital object identifiers (DOIs) for datasets and shareable assets, to allow for citation and acknowledgement of secondary use
> - Encouraging researchers to cite data in their publications
> - Offering opportunities for co-authorship when data is used, where appropriate/feasible

Conclusion

The Working Group's statement on data sharing is aspirational but should serve as a guiding principle as NCI implements data sharing policies and considers how to allocate resources to support data sharing and data access. The recommendations we have put forward are framed under the spirit of the statement on data sharing to help NCI lead the way in the cultural and philosophical shift that will be needed by funders, institutions, and researchers to fully realize the aspirations of sharing data in a way that will enable the research to affect the outcomes and lives of patients.

Appendix A: Interim Recommendations

Data Science Opportunities for the National Cancer Institute

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

August 14, 2018

Accepted by NCAB, August 2018

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 B: 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.