Cancer Moonshot Data and Technology Team: Enabling a National Learning Healthcare System for Cancer to Unleash the Power of Data

ER Hsu[1], JD Klemm[1], AR Kerlavage[1], D Kusnezov[2] and WA Kibbe[1]

The Cancer Moonshot emphasizes the need to learn from the experiences of cancer patients to positively impact their outcomes, experiences, and qualities of life. To realize this vision, there has been a concerted effort to identify the fundamental building blocks required to establish a National Learning Healthcare System for Cancer, such that relevant data on all cancer patients is accessible, shareable, and contributing to the current state of knowledge of cancer care and outcomes.

The vision of a National Learning Healthcare System for Cancer has many ramifications, including our ability to identify factors contributing to disparities in the dissemination of standard of care and access to high-quality oncology services and to identify populations at risk for initial disease, recurrence, and nonresponse to treatment. Data need to be consistently captured and shared, regardless of whether the patient is a clinical trial participant and where in the healthcare system a patient receives care. Three initial priority areas have surfaced as important steps towards this vision:

- Enabling a seamless data environment for patients, providers, and researchers;
- Unlocking science through open computational tools and storage platforms;
- Developing a data science-aware workforce capable of using the connected data environment.

These three areas will lay the foundations for a National Learning Healthcare System for Cancer, where we can learn from the contributed knowledge and experience of every cancer patient.

This article highlights the federal activities launched as part of the Cancer Moonshot to begin building the foundation of a National Learning Healthcare System for Cancer, as illustrated in **Figure 1**. The exemplars here reflect some of the efforts that aim to impact cancer prevention, early detection, screening, treatment, and outcomes for cancer patients. The National Learning Healthcare System for Cancer will contribute to the scientific evidence base necessary to understand cancer, to design more effective strategies to reduce the burden of cancer, and to continuously improve and adjust cancer care.

ENABLING A SEAMLESS DATA ENVIRONMENT FOR PATIENTS, PROVIDERS, AND RESEARCHERS

A key component of a National Learning Healthcare System for Cancer is establishing a scalable, interoperable data infrastructure to access, connect, and analyze multimodal datasets. Two activities launched as part of the Cancer Moonshot highlight the spectrum of data that must be supported: the National Cancer Institute's (NCI) Genomic Data Commons (GDC) and the Department of Energy (DOE) and Department of Veterans Affairs (VA) Million Veteran Program-Computational Health Analytics for Medical Precision to Improve Outcomes Now (MVP-CHAMPION).

The GDC,[1] which officially launched in June 2016, is an interactive system for researchers to store, process, and access genomic and clinical data generated by NCI and other research organizations to enable data-driven discoveries that provide insights into cancer biology and mechanisms of cancer resistance to therapy, with the goal of improving the diagnosis and treatment of cancer. The GDC stores the raw genomic data along with the analyzed data and phenotype data, so information can be reanalyzed as new computational tools and analytical methods are developed. Over time, the power of the GDC to enable discoveries will grow as new data are added. For example, both Foundation Medicine and the Multiple Myeloma Research Foundation have committed to contributing data that more than doubles the total number of patients represented in the GDC. These are important examples of partnerships, with a for-profit company and a nongovernmental entity contributing directly to public knowledge.

[1]Center for Biomedical Informatics and Information Technology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA; [2]National Nuclear Security Administration, Department of Energy, Washington, DC, USA. Correspondence: W Kibbe (warren.kibbe@nih.gov)

Received 5 December 2016; accepted 21 January 2017; advance online publication 31 January 2017. doi:10.1002/cpt.636

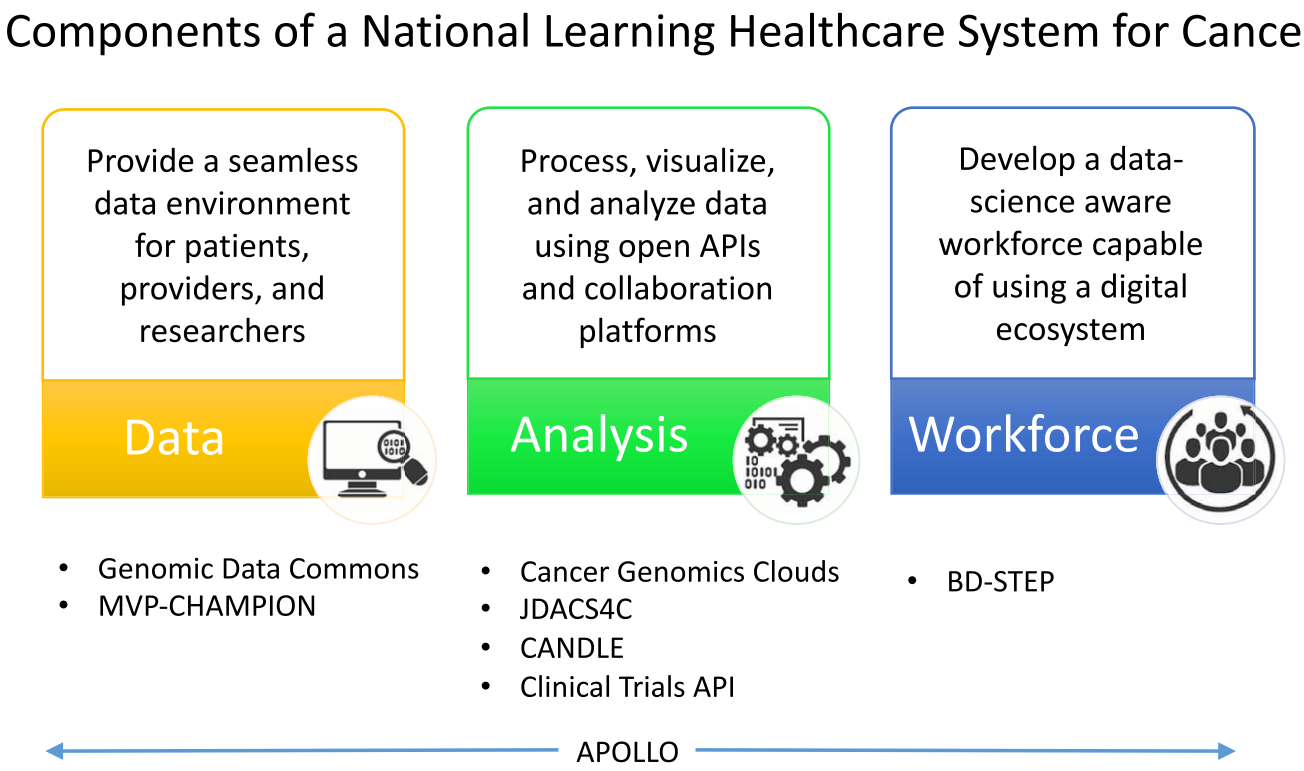

Figure 1 The critical components of a National Learning Healthcare System for Cancer are a seamless data environment; powerful computational tools and collaboration platforms; and a workforce trained in the use of these resources. Federal activities that have been launched as part of the Cancer Moonshot are listed for each of components. MVP-CHAMPION, Million Veteran Program-Computational Health Analytics for Medical Precision to Improve Outcomes Now; JDACS4C, Joint Design of Advanced Computing Solutions for Cancer; CANDLE, Cancer Distributed Learning Environment; API, application programming interface; BD-STEP, Big Data-Scientist Training Enhancement Program; APOLLO, Applied Proteogenomics OrganizationaL Learning and Outcomes Consortium.

MVP-CHAMPION (Computational Health Analytics for Medical Precision to Improve Outcomes Now) is a joint program between the VA and the DOE that combines the VA's clinical and genomic data from the Million Veteran Program (MVP) with DOE's national computing capabilities to simultaneously push the frontiers of precision medicine and computing. MVP-CHAMPION will accelerate the promise of precision medicine by developing better tools to prevent, detect, treat, and cure disease with the goal of transforming the practice of medicine and improving the lives of our nation's veterans and the public. In the area of cancer, the collaboration will be creating tools to help predict the best treatments for specific types of prostate cancer that will maximize benefit and minimize risk. Mental illness and cardiovascular disease will be the other initial areas of focus; eventually, the program plans to consider other medical ailments of concern to the VA.

UNLOCKING SCIENCE THROUGH OPEN COMPUTATIONAL TOOLS AND STORAGE PLATFORMS

In addition to access to data, a National Learning Healthcare System for Cancer must provide open access to the computational tools that process, analyze, and visualize these data. Open application programming interfaces (APIs) and platforms for tool-sharing are critical for enabling teams of researchers to leverage and draw insights from the data. The volume of biomedical data provided through resources like the GDC requires new models for access to compute and storage, designed with collaboration and sharing as central features. To meet this demand, NCI embarked on the Cancer Genomics Cloud Pilot[2] program to explore the feasibility of making data from the GDC available and computable in commercial cloud platforms, colocated with the elastic compute resources. In this model, researchers bring their tools to the data, along with their own datasets, eliminating the need to download, store, and protect these petabyte-scale datasets locally. These platforms facilitate collaborative analysis by distributed researchers, enabling team science approaches that are critical to modern cancer research.

In support of the Cancer Genomics Cloud activities, the NCI has initiated collaborations with two major commercial cloud providers, Amazon Web Services and Microsoft Azure, to each host a copy of the genomic data maintained by the GDC at no cost for 2 years. The NCI will work with these cloud partners to understand data usage patterns to develop a sustainable strategy for optimizing data storage and utility, while limiting costs.

The complexity of cancer requires the development and application of advanced machine learning and artificial intelligence approaches to promote the creation of predictive models for cancer treatments and outcomes, including therapeutic models. To advance this goal, the NCI and the DOE launched the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C),[3] a

collaborative effort that simultaneously advances precision oncology and computing. The initial scientific goals include:

- Identifying promising new treatment options through the use of advanced computation to rapidly develop, test, and validate predictive preclinical models for precision oncology.
- Deepening understanding of cancer biology using molecular, functional, and structural data from the NCI RAS gene family initiative through improved computer simulations and predictive models.
- Transforming cancer surveillance by applying advanced computational capabilities to population-based cancer data to understand the impact of new diagnostics, treatments, and patient factors.

The Cancer Distributed Learning Environment (CANDLE),[3] a partnership between NVIDIA, DOE, and the NCI, complements JDACS4C. CANDLE is focused on machine learning and building a single scalable deep neural network code that can be used to address all three challenges. By making the data and tools developed in these initial areas available to the broader research community, these efforts will catalyze efforts beyond JDACS4C in areas such as predictive therapeutic models or *in silico* drug development.

While these efforts are laying the groundwork for the computational tools for researchers who are developing new potential therapies, the NCI has also been enabling patients to better locate therapies by making NCI-supported cancer clinical trials available through an open API. This cancer clinical trials API will enable the community—advocacy groups, academia, and others in the cancer clinical trials ecosystem—to build applications, integrations, search tools, and digital platforms tailored to individual communities that bring clinical trial information to more providers, patients, and their family members.[4]

DEVELOPING A DATA SCIENCE AWARE WORKFORCE CAPABLE OF USING THE CONNECTED DATA ENVIRONMENT

A multipronged approach is needed to address the skills and workforce gap in biomedical data science, from early education exposure to data science, through undergraduate and graduate education, to educating established biomedical and clinical investigators on the application of computation to biomedical research questions. In addition, those in the computational disciplines should be introduced to the interesting challenges in the biomedical space.

To truly support biomedical data science, both the federal government and academia must signal that the development of tools and algorithms for data analysis is a valued discipline. This will require dedicated effort from both the public and private sectors in areas including: the incorporation of data science into biomedical/clinical training and biological/cancer problems into computational curriculum; development of career paths for biomedical data scientists; support of collaborative research that brings together biomedical and data science experts; and interagency fellowship programs such as the VA-NCI Big Data-Scientist Training Enhancement Program.

CONCLUSION

The Applied Proteogenomics OrganizationaL Learning and Outcomes (APOLLO) Consortium, described in a separate article, is an exemplar that ties these three pieces together. APOLLO serves as a pilot for the integration of a seamless data environment and computational tools, as well as a training ground for the next generation of researchers, such that research can be more rapidly translated into care in the context of a learning healthcare system.

The efforts outlined here represent a few of the foundational components of a National Learning Healthcare System for Cancer. These activities also align with the National Cancer Data Ecosystem recommended by the Cancer Moonshot Blue Ribbon Panel, which is envisioned to enable new insights into cancer initiation, progression, and metastasis and to inform new cancer treatments.[5] The vision of a National Learning Healthcare System for Cancer can only be achieved through the contributions and collaborations among all government agencies, the public sector, and the private sector.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

1. National Cancer Institute. Genomic Data Commons. <https://gdc.cancer.gov/> (2016).
2. National Cancer Institute. NCI Cancer Genomics Cloud Pilots. <https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots/nci-cloud-initiative> (2014).
3. National Cancer Institute and Argonne National Laboratory. Joint Design of Advanced Computing Solutions for Cancer; Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer. <https://cbiit.nci.nih.gov/ncip/hpc/jdacs4c> and <http://candle.cels.anl.gov/> (2016).
4. National Cancer Institute. Cancer Clinical Trials API. <https://www.cancer.gov/syndication/api> and <https://clinicaltrialsapi.cancer.gov/v1/> (2016).
5. Cancer Moonshot Blue Ribbon Panel. Cancer Moonshot Blue Ribbon Panel Report 2016. <https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/blue-ribbon-panel-report-2016.pdf> (2016).