Enhanced Data Sharing Working Group Recommendation: The Cancer Data Ecosystem

What is the recommendation (1-3 sentences)?

Our ability to accelerate progress towards improving cancer outcomes demands that researchers, clinicians, and patients across the country collaborate in sharing their collective data and knowledge about the disease. The Vice President's National Cancer Moonshot Initiative provides an unprecedented opportunity to create a national infrastructure for sharing cancer data. This infrastructure will support the development of a Cancer Data Ecosystem that will enable all participants across the cancer research and care continuum to contribute, access, combine and analyze diverse data that will enable new discoveries and lead to lowering the burden of cancer in our country.

Where are we are now (2-3 paragraphs)?

- **Summary of the current state of the science/practice**

The current dramatic and important revolution in biomedical research, influencing our understanding of cancer and how to treat it, are fueled by large data sets and complex analyses. Under the current cancer research and care paradigm, many powerful sources of data and potential insight are generated but are not being fully leveraged. These data are critical for identifying and utilizing associations between molecular data (e.g., from patient samples or model systems), other patient data, treatment, and response; however there are technical and logistical challenges for researchers to locate, integrate, and translate these stores of data from existing resources and repositories.

Recognizing the need for large datasets, several alliances and collaborative efforts have been initiated by different stakeholders to create an environment for sharing and collaboration such as Platform for Engaging Everyone Responsibly (PEER), Oncology Research Information Exchange Network (ORIEN), Project Genomic Evidence Neoplasia Information Exchange (GENIE), and Learning Intelligence Network for Quality (CancerLinQ). These individual efforts differ in many ways regarding design, construction and access to data; these differences challenge our ability to integrate these data and make these initiatives interoperable and synergistic. Given the diversity of goals of numerous stakeholders who are creating and using large data sets, and in the spirit of the precision medicine maxim that "one size does not fit all," accommodating and building upon the many cancer programs and dataset collaboratives already in place is a crucial goal of the proposed Cancer Data Ecosystem.

In addition, these efforts differ in their data governance approach, with some using an "opt-in" approach and requiring patient consent with patients donating their clinical data, tissue and molecular data (e.g., ORIEN, Project GENIE), The Cancer Genome Atlas [TCGA]); while others use an "opt-out" approach and use de-identified data from patient records (e.g., CancerLinQ). Still others, like Surveillance, Epidemiology, and End Results Program (SEER) and the CDC National Program of Cancer Registries (NPCR), include all cancer patients in a given catchment area. This

highlights the need for a clear data governance and management model to address data quality and access, as addressed in the *policy recommendations* companion document.

- **Identify barriers to progress and/or emerging opportunities**

The establishment of the Cancer Data Ecosystem is an emerging opportunity that would support research across the spectrum from basic research, through patient engagement, to care delivery. The proposed Cancer Data Ecosystem will allow both public and private information resources to be readily discovered and connected through the use of a common information architecture. This need is evident in several of the recommendations from all the other Blue Ribbon Panel (BRP) working groups that require large scale data collection and integration across many sources. These include a variety of projects and repositories, such as national clinical trials, patient registries, a Pre-cancer Genome Atlas, a repository of pediatric-cancer-related data from model organism systems, and a Cancer Immunity Atlas. Without an infrastructure for sharing and integrating these new data, we risk building more data silos and missing important opportunities for new insights that would be available if these data were "born interoperable". Implementing and unifying these new repositories through the underlying data science infrastructure of the Cancer Data Ecosystem will ensure they can be linked with one another and with future information resources that adopt this common platform. By building the proposed technical capabilities and making appropriate changes to key policies and governance models, long-standing challenges such as those listed below are now addressable.

Lack of searchable and interconnected data repositories with associated tools and services The many existing repositories of data are not always easy to find and are typically inconsistent with each other (e.g., use different nomenclatures, coding, data models, and definitions), making it difficult to integrate and analyze multiple datasets. In addition, they are generally not easily accessible via application program interfaces (APIs) and often reside within institutional boundaries. This creates a barrier for new discoveries by preventing integrative analysis of large amounts of data, which often requires co-localization of data and compute. While important advances have been made with regard to computing infrastructure for cancer genomics (e.g., Genomic Data Commons [GDC], NCI Cloud Pilots, Global Alliance for Genomics and Health [GA4GH] APIs etc), on the whole, this has not yet occurred for many important cancer data sources and data types.

Barriers preventing patients and researchers from contributing their data. A variety of sociological, technical, and policy issues make data sharing difficult for both researchers and patients. Most cancer researchers deposit their data in shared repositories when a study has been completed, rather than working in a data sharing environment while the research is ongoing. As a result, when researchers attempt to share data they are faced with multiple databases and multiple formats; often, they do not have the expertise or support structure needed to make their data sharable by ensuring it conforms to specific standards and formats. Data should be born digital and interoperable. Researchers also face conflicting incentives for meaningfully sharing data such as intellectual property/licensing, promotion and tenure incentives. Likewise, patients who want to share their medical records are faced with the

challenge of accessing these records, multiple sources of records, and the lack of appropriate user interfaces or repositories for securely managing and sharing their data. In addition, patients may have concerns related to privacy, potential downstream consequences of sharing their data, and lack of control over how their data is used. In many cases, participants may also feel disconnected from the research process as they do not always receive information in return for sharing data. Initiatives such as PCORNet (the National Patient-Centered Clinical Research Network), FDA's Sentinel Initiative, and Sync for Science have made great strides in accelerating patient-centered research and these efforts should be leveraged and expanded through this BRP recommendation. Policy related issues are discussed in detail in the companion policy document.

Standards and Interoperability: In addition to facilitating the sharing and ease of use of multiple existing data resources, a primary goal for the Cancer Data Ecosystem is to host contributions from different sources across the translational spectrum and provide a common data dictionary and tooling that promotes interoperability. The current lack of agreed upon ontologies, vocabularies, and data models severely impacts interoperability, integration, and analysis across multiple datasets, projects, and repositories. Agreed upon standards representing phenotypic data are particularly lacking. Evidence and provenance of the data must also be systematically recorded for algorithmic use and quality assurance. The ability to index and normalize definitions of data elements from all contributing sources will promote an "awareness of the possible" and encourage collaboration. Finally, there is a lack of tools to facilitate ease of use of adopting existing standards for researchers at the time of data generation as well as incentives for the use of such tools.

Consent and data-use agreements: The models for involving patients in research have not kept pace with the growth in the generation of biomedical data or with the growing desire to be directly involved and contribute to scientific knowledge. Uniform processes to increase and assess participant informed-ness at moment of enrollment are needed, as well as the ability for participants to manage their own data preferences in an ongoing and interactive way. In addition, there must be rules for handling participant data including ability to return data to the participant, as well as more streamlined ways of moving data out to researchers. Electronic, trackable, and machine-readable consents and terms-of-use agreements for data and other services would enable monitoring, computationally enforcing, and updating these agreements - tasks that are currently difficult. Ongoing efforts in these areas, such as those from Sage Bionetworks and Genetic Alliance, should be leveraged and built upon.

Recommendations for incentivizing the adoption of common data schemas and electronic consent are further detailed in the companion policy document.

Where do we need to be (in 1-5 years)?

The goal of this recommendation is to create a scalable and sustainable translational Cancer Data Ecosystem which will enable basic science researchers, clinicians, and patients to contribute, share, combine, and analyze cancer relevant data, in order to accelerate the pace of

discovery and lead to better patient outcomes and understanding of the underlying mechanisms of cancer. The Cancer Data Ecosystem will provide the data science infrastructure - including APIs, data schemas, consent templates, and service registration - necessary to connect repositories, analytical tools, and knowledge bases. Ultimately, knowledge held in the Cancer Data Ecosystem will enable the creation and evolution of new cancer treatment models, help initiate new clinical trials, and improve the overall quality of care for cancer patients.

In addition to technical capacity building, developing this Cancer Data Ecosystem will require changes in policies that currently inhibit data sharing; these policy changes are detailed in the accompanying *policy recommendations* document.

A central tenet of the Cancer Data Ecosystem is to enable patients and healthy individuals to directly contribute their data, or request an institution to do so on their behalf, for scientific research. The public will directly participate and derive immediate value from their interactions and contributions. It will provide patients with useful knowledge, community, and options as they move through their cancer journey, such as understanding the prevalence of their disease and clinical presentation, their anticipated standard of care, and the availability of nearby clinical trials. Ultimately, it is envisioned that the Cancer Data Ecosystem will support the ability of patients and their caregivers to participate in personalized healthcare decisions made jointly with their oncologists.

The Cancer Data Ecosystem will provide the appropriate level of protection of patient privacy, based on informed patient preference and understanding of risk, while allowing the public to benefit from the fruits of scientific and medical advances and the experience of individual cancer patients and cancer survivors.

Cancer Data Ecosystem Infrastructure: The fundamental infrastructure connecting the components of the Cancer Data Ecosystem are a centralized API, data schema, and a common data dictionary necessary for interoperability of the federated repositories, analytical services, and portals of the Ecosystem. The infrastructure will also provide a set of machine-readable consent and terms-of-use templates that will allow participants to quickly and easily understand access requirements and act on recommendations and findings tailored to them. All of the participating services (repositories, analytical services, and portals) of the Cancer Data Ecosystem will be registered in a central index to provide a "Yellow Pages" where participants will be able to find services of interest including machine-readable information about the inputs, outputs, terms-of-use, and credentials required by a service.

Cancer Data Ecosystem Components: The Cancer Data Ecosystem will be comprised of a dynamic collection of interoperable repositories, analytical services, and interactive portals that will allow data to be queried, aggregated, analyzed, and visualized in unique and powerful ways by researchers, patients, and clinicians. The flagship service of the Cancer Data Ecosystem will be a public-facing portal that will enable patients and healthy individuals to contribute their data (clinical, genetic, imaging, etc.) for scientific research. This portal will provide methods to collect and integrate individual-level patient data from their entire life experience, cancer journey, and all interactions with the healthcare system and provide the results of research performed with their data back to them in understandable terms.

In addition to data contributed by patients, the Cancer Data Ecosystem will support data repositories, curated knowledge bases, standard nomenclatures and ontologies, tools and services from diverse research programs and care systems. Participation in the Cancer Data Ecosystem will allow these resources to be integrated with a broader collection of data and tools, maximizing their potential value to cancer treatment and discovery. Newly-developed information resources, including those proposed through the domain-specific BRP working groups, will provide the initial focus for the Cancer Data Ecosystem. Looking forward, the breadth of the Cancer Data Ecosystem is envisioned to include the following categories of existing and future components (see figure below):

Software Tools

- Research Tools and Services: Clinical research tools leveraging the data repositories and knowledge bases will identify outcomes to inform activities such as the design of future prospective trials, clinical trial recruitment feasibility analysis, or provide a retrospective cohort as a comparator arm. Basic research tools will support new discoveries into cancer biology as well as translational research such as the discovery of new drug-biomarker interactions that lead to further preclinical and clinical studies.

- Patient-centered Tools and Services: Patient-centered tools including dynamic consent, access to current information about specific conditions, clinical trials, research opportunities, and integration with the many cancer advocacy and disease-focused communities.

- Clinical Decision Support Tools and Services: Point of care tools leveraging the knowledge base and data repositories will be integrated into clinical workflows and clinical information systems, enabling healthcare providers and patients to engage in shared decision making for treatment prioritization for individual patients. The subsequent treatment decisions will inform the further refinement of the knowledge base and treatment prioritization algorithms, and the patient specific treatment outcomes will be incorporated into the patient-centric data repository.

Curated knowledge Bases and Data Repositories

The above components would serve to connect, extract data from, and enable analysis of the following two components, which consist both of currently existing efforts as well as future efforts:

- Multimodal Patient Data Repositories: These data repositories consist of multi-modal data derived from patient-centric or pre-clinical sources and include data donated by patients, healthcare systems, laboratories, payors, registrars, researchers, and data collaboratives. Patient-centric data sources include linked data generated as part of patient care, patient experiences, or clinical research. Examples of existing data

repositories include the Genomic Data Commons (GDC), the Cancer Imaging Archive (TCIA), SEER, datasets from the Multiple Myeloma Research Foundation (MMFR), and the Million Veterans Program.

- Curated Knowledge Bases: Curated knowledge bases consist of computable assertions regarding the clinical utility of germline, somatic, epigenetic and imaging (including pathology and radiology images) biomarker alterations across the cancer care continuum. They include information derived from the biomedical literature, clinical practice guidelines, and open or completed clinical trials, as well as aggregated assertions from the patient-centric databases. Examples of existing knowledge bases include the Pharmacogenomics Knowledgebase (PharmGKB), Reactome, the Monarch Initiative, and My Cancer Genome.

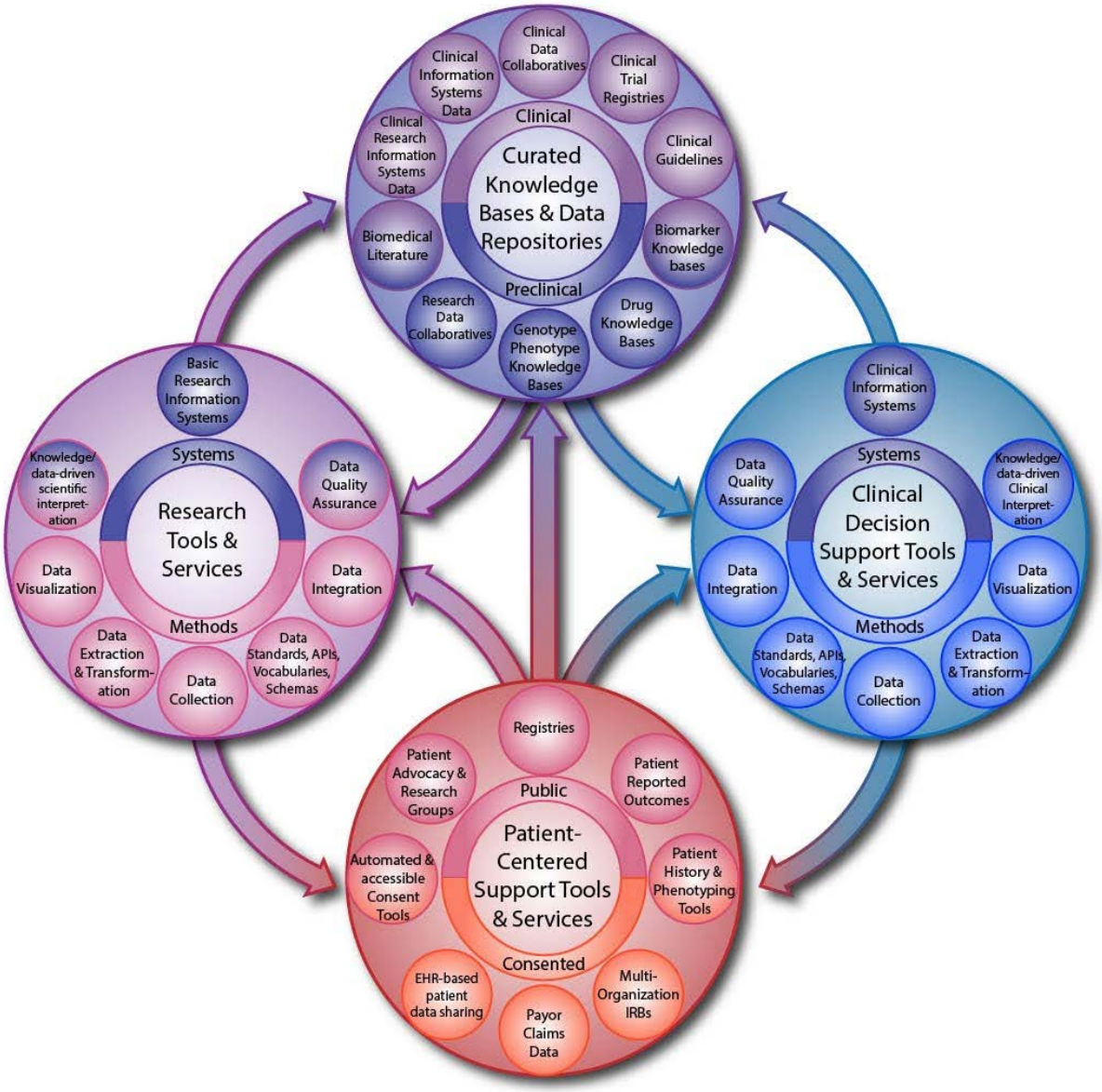

Figure 1: Components of the proposed Cancer Data Ecosystem. Circles represent different types of services in the Cancer Data Ecosystem, including data services, tools and portals. Arrows represent exchange of data, models and analysis results. The provenance, evidence, as well as quality metrics of each of the data elements will be recorded. A centralized registry of all services (ie. "Yellow Pages") will enable discovery of services as well as their inputs, outputs and terms-of-use in a machine-readable form.

- **Rationale for investing (Why is this priority ripe for accelerating?)**

An ecosystem approach is essential to ensuring interoperability, access to, and analysis of information resources that will be launched as part of the larger National Cancer Moonshot Initiative as well as future resources implemented through this platform. Developing the fundamental and enabling data science infrastructure necessary to connect these many resources and to make them findable and accessible will facilitate knowledge management and discovery.

Several BRP working groups have identified the need for new data, new analysis methods, and methods to access and integrate these new data with existing data. Defining and building the essential underlying data science infrastructure, standards, methods, and portals for the Cancer Data Ecosystem will serve as a central unifying structure for these new data and projects, providing tools to enable access, analysis, and interoperability across multiple data types and sources.

- **What would be needed for success?**

Critical to the success of the Cancer Data Ecosystem are changes to governance and policies that impede data sharing and data contributions from the public, research, and clinical care communities. The specific recommendations to address these barriers are outlined in the companion *policy recommendations* document. The technical recommendations are closely intertwined with the policy and governance recommendations and technical progress cannot be accomplished without accompanying changes to policy. In summary, the recommendations detailed in the policy document are to:

- Build easy ways for patients to contribute their data, whether directly or through a third party. This requires the elimination of policy barriers and addition of motivations to share data and consent for the use of the data for research purposes.
- Establish sustainable data governance to ensure long-term health of the ecosystem
- Develop standard electronic, trackable consent frameworks
- Develop standard data access frameworks for researchers and other stakeholders
- Develop standards and tools such that data is born interoperable.

Strategy: What will it take to get there?

The Cancer Data Ecosystem will necessarily develop through a phased approach, starting with core components of the infrastructure, the patient portal, and pilots to test different models for seeding the interconnected components. These projects will establish the foundation necessary for data, services, and definitions to flow through the Cancer Data Ecosystem and enable immediate engagement by the community.

Patient Portal: The flagship service of the Cancer Data Ecosystem will be a public-facing portal that enables patients and healthy individuals to contribute their data (clinical, genetic, imaging, etc.) for scientific research. This portal will provide methods and tools to collect and integrate individual-level patient data from their medical records, personal health data, and patient insights into disease progression, and provide the results of research performed with their data back to them in understandable terms. Network for direction patient engagment, described earlier in this report, envisions patients and providers directly sharing cancer outcome information and would serve as one of the initial pilots of this flagship service alongside other initiatives - such as PEER, Sync for Science and the Medicare Blue Button Initiative - that enable patient-directed contribution of data.

Infrastructure of the Cancer Data Ecosystem: The infrastructure that will allow data to be exchanged among participating services includes:

- The API, data schemas, and a common data dictionary for initial components of the Cancer Data Ecosystem. The API will be developed in collaboration with existing efforts such as GA4GH (for genomic data) and the newly developed NCI API for cancer clinical trials data (https://clinicaltrialsapi.cancer.gov).
- A registry of components that will state in a machine-readable format for each service its inputs, outputs, terms-of-use and credentials.
- A collection of data services that will link disparate information across samples, including clinical data, image data, and molecular data. Whenever possible, the Cancer Data Ecosystem will contain links to biospecimen repositories, patient-contributed samples, and model organisms that can be further analyzed to generate additional data.
- Enhanced cloud-computing platforms to enable tool developers to easily host their tools and provide them as a service in the Cancer Data Ecosystem. In the short term, the components and data should be redundantly hosted in different clouds. In the long term, as data grows, it will be more efficient to to store data on dedicated systems and to use smart caching of key data assets across clouds to allow for robust computational resources and choice in the marketplace.
- (Long term) Benchmarking tools and crediting mechanisms to evaluate the quality of the services, such as the DREAM Challenges.. This should be done jointly with efforts from the National Institute of Standards and Technology (NIST) and the US Food and Drug Administration (FDA).

Components of the Cancer Data Ecosystem: Priorities for the initial components of the Cancer Data Ecosystem will be driven by the needs identified by the BRP domain working groups. In addition, major ongoing efforts (of NCI and others) may also seed the Ecosystem. These ongoing efforts will need to be mapped to common vocabularies and data schema in order to interoperate in the Cancer Data Ecosystem. It is anticipated that the GDC will be a key initial component of the Cancer Data Ecosystem and will serve as the primary repository for genomic datasets.

- Tools for managing patient consents and handling the consenting process: To facilitate institutions' ability to share data in both the research and care settings, a small number of machine-readable, harmonized consent and access templates will be developed to be used by the data repositories. In the long term, the Cancer Data Ecosystem will include tools for managing patient consents and handling the consenting process. The Sage Bionetworks Participant Centered Consent Toolkit is a system in use today that is a model for such a service. The end result would drive towards sharing of outcomes and increasing the quality of care across healthcare providers.
- Center of Excellence for Cancer Data Harmonization: The Center of Excellence for Data Harmonization will have a mandate to provide tools, guidance, and training for harmonization for use of multi-modal research and clinical data. The Center of Excellence will support data curators and data shepherds to help upload and transform the data to common vocabularies and data schema. It will also provide similar utilities, guidance and training for tool developers to ensure that their tools adhere to the standards.

What does success look like?

Building upon the convergence of high throughput technologies, cloud computing and big data analytics, a focused national investment to develop a Cancer Data Ecosystem to dramatically accelerate discovery that will ultimately affect patient care. The creation of the basic infrastructure, the patient portal, and key initial services of the Cancer Data Ecosystem, will enable researchers to create new knowledge by accumulating and integrating much larger and diverse datasets. The patient portal will immediately affect the public and fundamentally change the interaction and engagement of patients and researchers. We anticipate that these initial components will be available and fully functioning within two years of initiating this program. The patient portal will allow all cancer patients to contribute their data to this widely accessible system and by Year 2, thousands of patients will have done so, providing access to dozens of researchers. By providing researchers, both basic and clinical, with the capability of exploring, analyzing and visualizing data across different tumor types and different populations, the cancer data ecosystem will grow, adapt, and evolve, attracting new data sets and new analytical methodologies.

As the participation in the Cancer Data Ecosystem grows, the goal is that by Year 5 the infrastructure and tools needed to integrate and analyze the data representing more than 1,000,000 patients will be supported by the Cancer Data Ecosystem. Utilizing the tools and services made available by the infrastructure of the Cancer Data Ecosystems to integrate and analyze a variety of data types and sources, researchers will be developing new treatments, screening and intervention strategies, decision support tools, and initiating more effective clinical trials.

Overall, an investment in this data science infrastructure will provide a sustainable Ecosystem to meet the growing demand of patients, clinicians and researchers for data access, data integration and analytical tools such as those articulated by the other working groups.