



Combining Large Datasets

Cancer Moonshot Task Group Final Summary

September 27, 2024

Version 2.0



Choose **VA**

VA



U.S. Department
of Veterans Affairs

Contents

1	Introduction to the CoLD Working Group.....	2
1.1	Formation of the Working Group	2
1.2	Importance of Combining Large Datasets	3
2	Task Group Activities.....	4
2.1	Overview	4
2.2	CoLD Challenges and Opportunities.....	4
2.2.1	Challenges	4
2.2.2	Opportunities	5
2.2.3	Discussion	6
2.3	Federal Data Landscape Assessment	7
2.3.1	Motivation	7
2.3.2	Approach	7
2.4	Individual Use Cases Explored	8
2.4.1	Radon Data	8
2.4.2	NASA CaRE	9
2.5	Next Steps.....	10
3	Appendix.....	10
3.1	Working Group Contributors.....	10
3.2	Full Text of the Originally Proposed RFI	11

1 Introduction to the CoLD Working Group

1.1 Formation of the Working Group

In February 2022, President Biden re-ignited the Cancer Moonshot with bold new goals: to reduce the cancer death rate by half within 25 years and improve the lives of people with cancer and cancer survivors. To achieve these ambitious goals, the White House convened the first-ever Cancer Cabinet, bringing together departments and agencies from across the federal government to end cancer as we know it.

The Cancer Cabinet convened three task forces and supporting task groups (Figure 1), including the Data and Innovation Task Force, which supported the Cancer Moonshot priority to “Deliver innovation to patients and communities.” In early 2023, the “Combining Large Datasets” (CoLD) Task Group was created within the Data and Innovation Task Force, including subject matter experts at the Department of Veteran Affairs (VA), Environmental Protection Agency (EPA), National Aeronautics and Space Administration (NASA), Office of the National Coordinator for Health Information Technology (ONC), United States Department of Agriculture (USDA), Centers for Medicare & Medicaid Services (CMS), National Cancer Institute (NCI), Advanced Research Projects Agency for Health (ARPA-H), and the Food and Drug Administration (FDA).

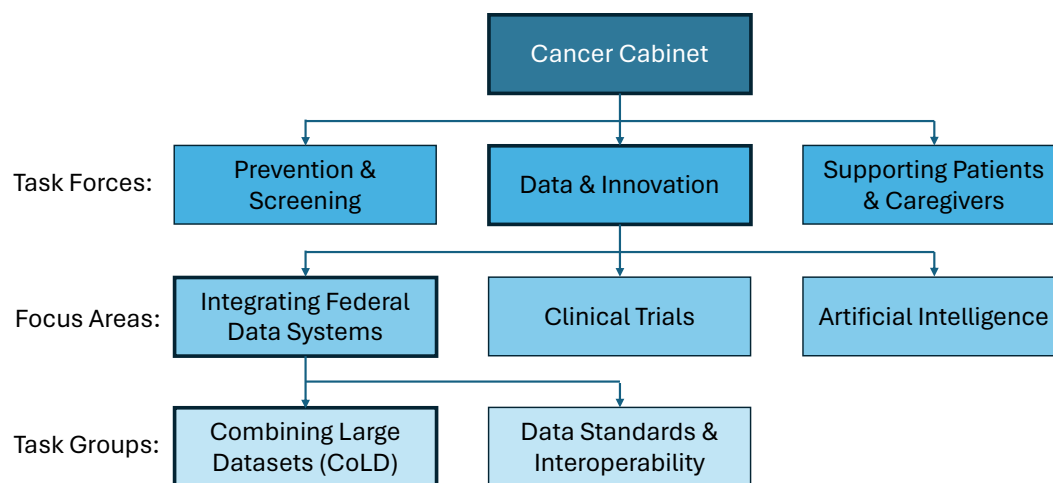


Figure 1: Cancer Moonshot Organizational Chart

The scope of the CoLD Task Group was how federal agencies combine large datasets for broad applications across cancer prevention and control, including nutrition, epidemiology, and military/Veteran health. Within this scope, the group sought to better leverage the immense potential of data and power of data tools to increase our understanding of cancer incidence, causes, mortality, treatments, prevention, outcomes, costs, and all other aspects of the burden of cancer. In this document, we use “Cancer Research” as shorthand for all efforts to enhance our understanding of and response to the burden of cancer.

1.2 Importance of Combining Large Datasets

Large datasets relevant to cancer include patient medical records, clinical trial databases, population-based cancer registries in various countries, and The Cancer Genome Atlas, to name a few. New insights into cancer also come from various large non-medical datasets, such as environmental (e.g., earth, water, atmosphere, solar) observations, personal and population histories, and many others. Combining these large datasets for enhanced insight can have a profound impact on cancer research, treatment, and prevention, including:

- i. **Improved Understanding of Cancer Biology:** large datasets enable researchers to analyze vast amounts of genetic, genomic, proteomic, and clinical data from thousands to millions of patients. This helps in identifying patterns, mutations, biomarkers, and molecular signatures associated with different types and subtypes of cancer. This deeper understanding of cancer biology can lead to the development of more precise diagnostic tools and targeted therapies.
- ii. **Personalized Medicine:** cancer is not a single disease but a collection of diseases with diverse genetic and molecular characteristics. Large datasets allow for the identification of genetic mutations and biomarkers that can predict how a patient will respond to specific treatments. This enables health care providers to tailor treatment plans to individual patients, improving outcomes and potentially reducing adverse reactions.
- iii. **Drug Development and Clinical Trials:** by analyzing large datasets, researchers can identify potential drug targets and develop new therapies. Large-scale genomic studies can uncover novel mutations or pathways that can be targeted with new drugs. Additionally, large datasets contribute to more efficient and effective clinical trials by identifying suitable patient populations and endpoints for evaluating new treatments.
- iv. **Epidemiological Studies and Risk Assessment:** large datasets provide valuable insights into the epidemiology of cancer, including identifying risk factors, geographic and demographic trends, and other environmental influences. This information is crucial for developing strategies for cancer prevention, informed risk mitigation or interventions, and early detection.
- v. **Predictive Modeling and Prognostic Tools:** machine learning and artificial intelligence techniques applied to large datasets can build predictive models for cancer risk, prognosis, and treatment response. These models can help clinicians make more informed decisions and improve patient outcomes by predicting outcomes based on individual patient characteristics and biomarkers.
- vi. **Data Sharing and Collaboration:** large datasets facilitate collaboration among researchers, institutions, and even across countries. Shared data allows for validation of findings, replication of studies, and pooling of resources to accelerate discoveries and advancements in cancer research and treatment.

In short, large datasets have begun to transform cancer research and treatment by providing unprecedented insights into the molecular basis of cancer, enabling personalized approaches to treatment, fostering innovation in drug development, and improving outcomes for cancer patients worldwide.

Despite the promise and initial impact of combining large datasets to reduce the burden of cancer, many challenges and tremendous unrealized potential impact remain. Mitigating these challenges and further realizing the potential of combining large datasets was the motivation for the formation of the CoLD Task Group.

2 Task Group Activities

2.1 Overview

As the CoLD team embarked on our goal – to help realize the full promise of combining large datasets for improved cancer research, prevention, and treatment – we understood that this work must reflect the current state and light a path to the envisioned future. The three primary activities of the CoLD team were designed to enable this, while remaining within the capacity of the CoLD team to implement. These activities were:

- discussions within the team and with invited experts regarding the challenges and opportunities in combining large datasets for cancer research.
- work towards a survey of the relevant federal data landscape, to uncover and enable new research opportunities.
- highlight use case studies that demonstrate the power and potential of combining large datasets for cancer research.

The outcomes of these activities are summarized in the following sections.

2.2 CoLD Challenges and Opportunities

In a series of discussions within the team and with invited experts, the CoLD team surfaced common challenges in combining large datasets for the various cancer stakeholders (the “CoLD challenges”), as well as ongoing efforts at various agencies and organizations to address these challenges (which we termed “opportunities”). We found that knowledge about both the challenges and the solution efforts is often siloed. Thus, the team determined that a key contribution of our work would be to collect and share information about the CoLD challenges and opportunities. The results of those discussions are summarized in this section.

2.2.1 CHALLENGES

- i. Data Access and Availability
 - a. Fragmentation: cancer data often resides in fragmented databases across institutions, countries, and research consortia, making it challenging to access comprehensive datasets.
 - b. Data Silos: institutional and legal barriers can restrict data sharing, leading to data silos that hinder collaborative research efforts.
 - c. Data Privacy and Security: concerns about patient privacy and data security can limit access to sensitive clinical and genomic data.
- ii. Data Quality and Variability

- a. Data Curation: ensuring data quality is often the most labor-intensive aspect of using large data for cancer research, with few shortcuts for existing datasets.
 - b. Variability: variations in data collection methods, assay platforms, and data processing techniques can lead to inconsistencies in data quality.
 - c. Missing Data: some datasets may lack comprehensive clinical annotations, longitudinal follow-up data, or metadata crucial for interpreting genomic findings.
- iii. Interoperability and Integration
 - a. Data Formats and Standards: lack of standardized data formats, terminologies, and metadata schema complicates data integration and interoperability.
 - b. Technical Heterogeneity: integration of heterogeneous data types (genomic, clinical, imaging, longitudinal Earth observation, etc.) from different sources requires sophisticated computational methods and interoperability frameworks.
 - c. Cross-domain Integration: integrating genomic data with clinical outcomes, environmental factors, and patient-reported data poses technical and semantic challenges.
- iv. Ethical and Legal Considerations
 - a. Data Governance: establishing governance frameworks for ethical data use, consent management, and responsible data sharing practices.
 - b. Regulatory Compliance: adhering to regulatory requirements (e.g., General Data Protection Regulation in Europe, Health Insurance Portability and Accountability Act in the US) while promoting data sharing and collaborative research.
 - c. Informed Consent: ensuring informed consent procedures that accommodate evolving data uses, patient reach-back to revoke consent, and secondary research purposes.
- v. Resource and Infrastructure Constraints
 - a. Computational Resources: processing and analyzing large-scale cancer datasets require substantial computational resources and expertise.
 - b. Data Storage: managing and storing large volumes of data, including raw sequencing data and clinical records, requires scalable and secure infrastructure.
 - c. Training and Capacity Building: ensuring researchers and the supporting workforce have the necessary skills and training in data science and bioinformatics to effectively utilize large-scale cancer datasets.

2.2.2 OPPORTUNITIES

Table 1 lists examples discussed of efforts to solve the above CoLD challenges. We call these “opportunities” because their full potential will be realized through further development and refinement, sharing them broadly, and integrating multiple solution elements.

Table 1: Example efforts to solve the CoLD challenges.

Name	Description	Data Access and Availability	Data Quality and Variability	Interoperability and Integration	Ethical and Legal Issues	Resources and Infrastructure
FHIR: Fast Healthcare Interoperability Resources	International standard for exchange of health care data, including patient information, treatment details, and research findings.	X	X	X	X	
USCDI/USCDI+: United States Core for Data Interoperability	Establishes core set of data to support patient care and facilitate patient access while maintaining privacy. USCDI+ extends USCDI to allow case specific data such as cancer.	X	X	X	X	
mCODE: Minimal Common Oncology Data Elements	Builds on USCDI+ to standardize data elements for oncology data managed by electronic health records (EHRs) and shared across different health systems.	X	X	X	X	
CodeX: Clinical Oncology Data Exchange	Leverages mCODE to increase the amount of high-quality, shareable data for all cancer types.	X	X	X	X	
EA: Enterprise Architecture	Potential cornerstone for management of cancer data sets, providing a structured approach to leveraging resources and ensuring that data is accessible, integrated, and primed for future advancements in cancer treatment and research.	X		X		
AI: Artificial Intelligence	Techniques such as Retrieval-Augmented Generation (RAG) optimized Large Language Models (LLMs) could reduce data curation effort, support data interoperability and integration, and ensure ethical and legal standards.		X	X	X	
BDF: Biomedical Data Fabric Toolbox	ARHA-H effort to simplify connecting biomedical research data from thousands of sources while overcoming barriers caused by incompatible data dialects. BDF seeks to support: high-fidelity and timely data collection in computer-readable forms, multi-source data analysis at scale, advanced and intuitive data exploration, improved stakeholder access with privacy and security, and generalize biomedical data fabric tools across disease types.	X	X	X	X	X
NIH Strategic Plan for Data Science	Describes NIH's approach to interoperability, integration, and other challenges to data-enabled biomedical research.	X	X	X	X	X
Privacy Enhancing Standards and Techniques	Technologies such as synthetic datasets, homomorphic encryption, DataSHIELD, TripleBlind, and others could help maintain privacy while enabling data analysis.	X		X	X	
STRIDES: Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability	NIH initiative to partner with cloud service providers to allow NIH-supported researchers to affordably access cloud services and environments.					X

2.2.3 DISCUSSION

Addressing the CoLD challenges, and sharing and integrating appropriate CoLD solution elements, will require sustained, concerted efforts from stakeholders. The team considered, but decided against, compiling an example architecture that would integrate some of these elements into a more comprehensive, end-to-end solution. Taking this additional step would have ignored that most of the opportunities listed above are in development, and that there are many potential integrations. The best combination of current elements depends on various

factors and constraints, such as the funding agency, knowledge of the participants, solution requirements, and specific data being used.

2.3 Federal Data Landscape Assessment

2.3.1 MOTIVATION

While the focused discussions above were intended to uncover promising elements of the CoLD challenges, the CoLD team determined that one area of cancer research would benefit from a more comprehensive survey and data product: the data landscape.

Access to accurate, complete, well-curated information and suitable data analytics capabilities are essential for productive cancer research. Researchers and other stakeholders face challenges in receiving, using, combining, producing, and distributing cancer-relevant information, with a particular focus on potentially cancer-relevant datasets and their dataset-specific barriers to use or integration.

Our interagency discussions surfaced that there is a significant informational and awareness gap of existing and emerging datasets potentially relevant for impactful cancer use cases. Some of the biggest information gaps were about opportunities to bring together datasets from multiple agencies that might uncover new insights regarding cancer incidence, causation, and more. The reasons for this were many, including the lack of knowledge of data existence, access, relevance to cancer, and more. The team felt that publishing a comprehensive survey of cancer-relevant data and access would immediately allow creative researchers to imagine and pursue innovative research goals.

Given the potentially substantial efforts, time, and fiscal burdens required by organizations and research teams to combine (and provision) large datasets, it is critically important to ensure a positive value proposition for any such activities. Typically, this requires an evaluation of the potential specific utilizations of combined datasets, with an emphasis on the unique synergistic potential of the combined data beyond the individual utilization of component datasets. However, it was determined through numerous team deliberations that there is a significant informational and awareness gap of existing and emerging datasets potentially relevant for impactful cancer use cases. The task group therefore identified a relevant need for comprehensive and searchable descriptors of data across the federal landscape.

2.3.2 APPROACH

In February 2024, the CoLD team completed a draft Request for Information (RFI) to solicit responses from federal, commercial, and academia organizations regarding cancer-relevant datasets. The full text of the draft RFI is found in Section 3.2. It requests a high-level description of each dataset, as well as dataset details such as contents, size and scope, timeframe covered, granularity (e.g., resolution if geospatial data), primary data source(s), methodology to acquire/generate, ownership, availability and use restrictions, and encoding format(s). In addition, RFI responses could include a brief explanation of how this data has been or might be useful, either alone or in combination with other datasets, to address specific use case(s)

related to cancer incidence, causes, mortality, treatments, prevention, outcomes, costs, and/or all other aspects of the burden of cancer.

While preparing this RFI, the CoLD team came to recognize that we had neither the bandwidth nor expertise to conduct a data landscape assessment that would likely include hundreds or thousands of RFI responses, nor did we have the enduring presence to form those responses into a user-friendly, accessible, curated, enduring, online resource that could be used by thousands of researchers. We concluded that such a task would have to be undertaken by a team specifically funded for that work. The draft RFI in Section 3.2 is therefore our deliverable.

The CoLD team also explored potential mechanisms for conducting the RFI. The traditional approach of requesting spreadsheets and/or narrative documents would create a massive data integration challenge to integrate the responses into a single database for easy analysis, search, and display. We noted that a dynamic, sustained, web-based data collection approach could entirely avoid the data integration challenge.

2.4 Individual Use Cases Explored

To demonstrate the concrete potential of combining datasets, we sought to nominate cancer-specific use case(s) meeting several pragmatic criteria: 1) component datasets already exist across two or more federal agencies, 2) data sharing/combination could be performed without legal, privacy, or other policy/regulatory restrictions, 3) existing platform(s) for data exchange, and most importantly, 4) clear cancer relevance with the potential for clinical actionability.

2.4.1 RADON DATA

We nominated the combination of radon geospatial data with the Department of Veterans Affairs' (VA) clinical data as an optimal use case, satisfying all criteria:

- 1) **Existing Component Datasets Across Federal Agencies:** nationwide radon testing data are compiled by the Centers for Disease Controls' (CDC) Environmental Public Health Tracking Program at county-wide geospatial resolution. Whereas both residential history and longitudinal clinical data (including cancer diagnoses) are collected by VA as a fundamental component of its mission to provide healthcare for Veterans.
- 2) **Data Sharing and Legal Compliance:** CDC actively provisions radon data for public consumption without restrictions, enabling VA to ingest data without legal, privacy, or other regulatory barriers. Note that the reverse direction of data sharing in this case would be infeasible due to multiple existing legal, privacy, and policy constraints on VA's capacity to release relevant data to external entities, even within the federal ecosystem.
- 3) **Existing Platforms for Data Exchange:** CDC hosts an application program interface (API) enabling query and retrieve access to radon data among multiple other environmental health datasets. VA has access to this API, as well as existing platforms to ingest external datasets for internal integration and provisioning.

- 4) **Cancer Relevance and Clinical Actionability:** Radon gas is a naturally occurring radioactive hazard, representing a significant public health concern due to its proven association with lung cancer. It is the leading cause of lung cancer among non-smokers, and radon exposure accounts for >20,000 lung cancer deaths annually in the United States alone. The synergy between radon exposure and smoking drastically increases lung cancer risk, underscoring a critical area for cancer prevention and public health intervention. There are at least two key areas of potential clinical actionability leveraging radon data inside VA's clinical enterprise: 1) targeted radon testing and mitigation (cancer prevention) and 2) radon risk adjustment in lung cancer screening eligibility requirements (cancer screening).

2.4.2 NASA CARE

The NASA Satellite Analysis for Cancer Risk Evaluation (NASA CaRE) project also meets all required use case criteria. Its primary objective is to develop an integrated data platform that utilizes NASA's satellite imagery and databases from various federal agencies to identify environmental factors contributing to cancer. This platform leverages NASA's advanced hyperspectral imaging to correlate geospatial data, pollution measurements, cancer incidence, and biomarkers, employing AI to explore potential causal relationships.

- 1) **Existing Component Datasets Across Federal Agencies:** NASA CARE will integrate datasets from NASA, Environmental Protection Agency (EPA), United States Geological Survey (USGS), National Oceanic and Atmospheric Administration (NOAA), Centers for Disease Control and Prevention (CDC), Food and Drug Administration (FDA), and the National Institutes of Health (NIH). This includes satellite imagery from the NASA Earth Exchange (NEX) database, pollution and radiation data from EPA's RadNet, geological and atmospheric data from USGS, cancer epidemiology from the Surveillance, Epidemiology, and End Results (SEER) cancer registry program managed by NCI, and genotoxic measurements from EPA's ECOTOX database. These datasets collectively provide a comprehensive basis for analyzing environmental impacts on cancer incidence.
- 2) **Data Sharing and Legal Compliance:** collaboration involving databases designed with data privacy, licenses, and regulatory constraints ensures compliant data usage. The project will primarily use public data under the Creative Commons Zero (CC0) license to minimize privacy concerns. Where necessary, SEER data will be de-identified according to HIPAA standards. Aggregated cancer incidence rates at geographical levels (e.g., county or state) will be used to maintain privacy, aligning with the public availability of other datasets such as ECOTOX and RadNet.
- 3) **Existing Platforms for Data Exchange:** the NASA Open Science Data Repository (OSDR) (accessible via osdr.nasa.gov) will serve as the primary platform for deploying this project. Although SEER, ECOTOX, and RadNet lack APIs, relevant data will be downloaded and integrated into the OSDR, which features an API and a public S3 bucket on Amazon Web Services (AWS) for rapid analytical tool deployment and AI application considering the large array of data. Additionally, NASA's Earthdata Search

offers APIs that provide access to various datasets, including those from NEX, facilitating programmatic access.

- 4) **Cancer Relevance and Clinical Actionability:** Identifying strong correlations between satellite-detected environmental signatures and known carcinogens or rapid physical outcomes (e.g., DNA damage in various species) alongside local cancer incidence in large human populations can offer early warnings. This insight enables local and federal governments to implement protective measures and develop countermeasures, potentially revolutionizing cancer prevention in the U.S. and significantly reducing long-term healthcare costs.

NASA CARE's innovative approach to integrating diverse data sources aims to boost our understanding of environmental determinants of cancer, offering substantial public health benefits and informing policy decisions with its clinically actionable findings.

2.5 Next Steps

Each of the above primary activities of the CoLD team have clear, impactful follow-up actions, which would be overseen by an interagency group and carried out by a funded team of cancer and data experts.

- **CoLD Challenges:** a team of experts should convene to show how the activities and tools listed in Section 2.2 might integrate to accelerate research; identify the remaining gaps in these efforts; outline next steps for agencies to close those gaps.
- **Data Landscape:** one or multiple agencies should perform and/or solicit a data landscape assessment like the draft in Section 3.2. Any solicited responses should be curated and published in a user-friendly, accessible, curated, enduring, online resource that could be used by thousands of researchers.
- **CoLD Use Cases** – the use cases described in Section 2.4, or similar ones, should be funded by agencies to completion, to further illuminate CoLD challenges, opportunities, and productive approaches.

Implementing these next steps will further illuminate paths that enable researchers and healthcare professionals to collaborate more effectively, and for cancer research efforts to proceed more rapidly, leading to reduced cancer occurrence as well as better cancer treatments and outcomes.

3 Appendix

3.1 Working Group Contributors

Name	Organization	Role
Dr. Reid Thompson	VA	Co-Lead
Dr. Bryan Biegel	NASA	Co-Lead
Sylvain Costes	NASA	Member
Ioana Singureanu	VA	Member
Christopher Shawn	VA	Member

Name	Organization	Role
Catherine Lerro	VA	Member
Alison Krajewski	VA	Member
Kerry Rowe	VA	Member
Ziyad Al-Ali	VA	Member
Jennifer Roberts	ARPA-H	Member
Jennifer Hafterson	Fred Hutch Cancer Center	SME
Castine Clerkin	North American Association of Central Cancer Registries (NAACCR)	SME
Elizabeth Ragan	VA	Member
Marvin Rydberg	VA	Member

3.2 Full Text of the Originally Proposed RFI

AGENCY:

Office of Science and Technology Policy (OSTP) (or other agency TBD – e.g. VA).

ACTION:

Notice of Request for Information (RFI).

SUMMARY:

The Office of Science and Technology Policy (OSTP) seeks information on data-enabled approaches to advance the Cancer Moonshot. Specifically, OSTP seeks information to identify existing or aspirational datasets that might be integrated to increase understanding of cancer incidence, risks and causes, mortality, treatments, prevention, outcomes, and all other aspects of the burden of cancer, as well as barriers to the successful provisioning or integration of these datasets.

DATES:

Interested persons and organizations are invited to submit comments on or before 5:00 p.m. ET on May 28, 2024.

ADDRESSES:

Interested individuals and organizations should submit comments electronically to email_address@ostp.eop.gov and include “RFI for Data-Enabled Cancer Moonshot Advances” in the subject line of the email. Due to time constraints, mailed paper submissions will not be accepted, and electronic submissions received after the deadline cannot be ensured to be incorporated or taken into consideration.

Instructions

Response to this RFI is voluntary. One response should be submitted for each dataset description. Each response should include a high-level description of the dataset, as well as details such as contents, size and scope, timeframe covered, granularity of data (e.g. resolution if geospatial data), primary data source(s), methodology to acquire / generate, ownership, availability and use restrictions, encoding format(s). Each response

should also include a brief explanation of how this data has been or might be useful, either alone or combined with other datasets, to address specific use case(s) related to cancer incidence, causes, mortality, treatments, prevention, outcomes, costs, and/or all other aspects of the burden of cancer. It would be especially helpful if you could identify what data that you might uniquely have access to or otherwise produce or prepare in any way.

Please be concise with your submissions. There is no minimum or maximum length limit or format requirement, but each submission should ideally not exceed five pages (including graphics). If desired, responses should include the name and email of the person(s) or organization(s) filing the comment to enable recontact for future discussion, as relevant.

OSTP invites input from all stakeholders including members of the public, representing all backgrounds and perspectives. In particular, OSTP is interested in input from cancer researchers, clinicians, patients, caregivers, policy makers, healthcare providers, cooperative clinical trials groups and medical specialty societies, insurers and other commercial entities, federal agencies, and any other interested parties. Please indicate which of these stakeholder types, or what other description, best fits you as a respondent. If a comment is submitted on behalf of an organization, the individual respondent's role in the organization may also be provided on a voluntary basis.

Comments containing references, studies, research, and other empirical data that are not widely published should include copies or electronic links of the referenced materials. No business proprietary information, copyrighted information, or personally identifiable information should be submitted in response to this RFI. Please be aware that comments submitted in response to this RFI may be posted on OSTP's website or otherwise released publicly (names and contact details would not be included with any public release).

In accordance with FAR 15.202(3), responses to this notice are not offers and cannot be accepted by the Federal Government to form a binding contract. Additionally, those submitting responses are solely responsible for all expenses associated with response preparation.

FOR FURTHER INFORMATION CONTACT:

For additional information, please direct questions to [POC] at 202-456-???? or email_address@ostp.eop.gov.

SUPPLEMENTARY INFORMATION:

The Cancer Moonshot is a White House initiative to bring renewed leadership to the fight against cancer, facilitate new collaborations, and drive progress across the cancer journey utilizing all facets of the oncology community – federal agencies and departments, private companies, healthcare providers, patient groups, philanthropies,

and all Americans. The Cancer Moonshot is mobilizing efforts toward achieving two ambitious and compelling goals: To prevent more than 4 million cancer deaths by 2047 and to improve the experience of people who are touched by cancer.

Achieving these goals will require an integrated, multi-objective approach. One objective is to better leverage the immense potential of data and power of data tools. In pursuit of this objective, OSTP (in partnership with other Executive Office of the President components) is seeking information on ways that data could be better used to increase our understanding of cancer incidence, causes, mortality, treatments, prevention, outcomes, costs, and all other aspects of the burden of cancer.

The various cancer stakeholder communities will bring diverse and complementary perspectives and insights to the pursuit of this objective. Key stakeholder communities include:

- Patients: Motivate our effort to reduce the burden of cancer. Patients serve as the source of much of the data in cancer patient records, which are key to advancing our understanding of cancer. The accuracy and completeness of patient records can dramatically impact their value in developing insights.
- Caregivers: Provide daily support for cancer patients and may also provide information for patient records.
- Clinicians: Provide primary care for cancer patients, including prescribing tests, diagnosing, and developing a treatment plan. Clinicians may participate in research studies, use research results based on large-scale data, and search the vast array of patient records for relevant cases.
- Healthcare providers: These individual health professionals and health facility organizations provide health care to cancer patients, and may consult research results to set policy and practices for care is.
- Researchers: Envision and implement research, including experimental, clinical, and epidemiological studies, to advance our understanding of cancer. This research often involves data analytics to develop new insights using new and existing large-scale cancer-relevant data, such as experimental results, anonymized patient records, and epidemiologic data.
- Insurers: Use clinical and epidemiological studies, among other data, to decide what cancer care provides greatest value, to provide support for data-informed preventative measures and cancer screening and risk assessments.
- Industry: Perform and support basic and translational research studies to identify and realize new therapeutic interventions and/or indications, as well as develop novel testing methodologies for diagnosis and/or risk stratification
- Policy Makers: Develop laws, regulations, and (for business policy makers) business decisions related to cancer.

For all stakeholders, access to accurate, complete, well-curated information and suitable data analytics capabilities are essential. Most stakeholders also seek to augment that information, whether at an individual or epidemiological level. This RFI seeks

information about the challenges and gaps faced by stakeholders in receiving, using, combining, producing, and distributing cancer-relevant information, with a particular focus on potentially cancer-relevant datasets and their dataset-specific barriers to use or integration.

To envision and work towards a more efficient, enabled, and effective future for cancer-related data, we must understand the current cancer data landscape, with emphasis on novel and untapped opportunities.

Information Requested: OSTP invites input from all interested parties as outlined in the instructions. Respondents may provide information for one or as many items as they choose. Be as specific as possible. Remember to indicate your stakeholder categor(ies).

Here are several motivating questions that may help to frame your response(s):

What are the most valuable cancer-related data sources and datasets that you think others in your or other stakeholder communities should know about?

What data that are not traditionally used for cancer purposes would be valuable in your or other stakeholder communities to provide potential insights into cancer?

What is an aspirational dataset (i.e. that does not currently exist or is otherwise not accessible) that you wish was available for providing insights into cancer?

What are the technical, legal, policy, regulatory, ethical, fiscal, and/or other constraints that might limit access to or use of this dataset?

Dated:
[POC],
[Role].

Footnotes

1. Example RFI #1 (to be removed in final draft): Request for Information on Data Collection for Emergency Clinical Trials and Interoperability Pilot, Office of Science and Technology Policy, 10/28/2022.
2. Example RFI #2 (to be removed in final draft): Request for Information to Support the Development of a Federal Environmental Justice Science, Data, and Research Plan, Office of Science and Technology Policy, 10/13/2023.