Supporting Statement A for

Generic Clearance for National Cancer Institute (NCI)

Resources, Software and Data Sharing Forms

OMB No. 0925-XXXX; Expiration Date xx/xx/20xx

Date: March 14, 2022

Check off which applies:

X New
- Revision
- Reinstatement with Change
- Reinstatement without Change
- Extension
- Emergency
- Existing Collection in Use Without an OMB Number

Federal Government Employee Address:

Name: Diane Kreinbrink

Address: 9609 Medical Center Drive, Rockville, MD 20850

Telephone: 240.276.7283

Email: Diane.Kreinbrink@nih.gov

Table of contents

List of Attachments

A. Justification

This is a new generic information collection request seeking approval for three years. The purpose of the "Generic for National Cancer Institute (NCI) Resources, Software and Data Sharing Forms" is to provide a cloud-based data science infrastructure to test ideas quickly, respond to the project's needs as they evolve, incorporate feedback from scientists for flexible, innovative research methods, and provide a foundation for the cancer research community to make new scientific discoveries. The cloud-based infrastructure will connect data sets with analytical tools and access to online workspaces, tools, and NCI resources to support data sharing. Cloud-based data sharing, and analysis is only possible when requests, or applications, are made to access, upload, store, and analyze cancer data and metainformation.

A.1 Circumstances Making the Collection of Information Necessary

The NCI plans to co-locate data, storage, and computing infrastructure in the cloud with tools for analyzing and sharing data to create an interoperable resource for the research community. A shift in the location of the data has begun with the creation of the NCI Cancer Research Data Commons and the Childhood Cancer Data Initiative. These two separate resources will allow a multitude of different types of data (genomics, proteomics, imaging, clinical, population science, and other (e.g., flow cytometry, videos, etc.) to be submitted, stored, accessed, and managed by NCI. Data sharing allows data generated from one research study to be used to answer questions beyond the original study. It reinforces open scientific inquiry, encourages diversity of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, and enables the exploration of topics not envisioned by the initial investigators. Biomedical researchers and data scientists can use the NCI cloud resources, web interface, and computational workspace to query, submit data, analyze, and visualize data.

The Public Health Law Title 42 of the United States Code provides the legal authority that allows the NCI to collect this information. NCI, was established under the National Cancer Act of 1937, and is the Federal Government's principal agency for cancer research and training and has a direct congressional mandate to disseminate information related to cancer to the public. The National Cancer Act of 1971 broadened the scope and responsibilities of the NCI and created the National Cancer Program. Over the years, legislative amendments have maintained the NCI authorities and responsibilities and added new information dissemination mandates as well as a requirement to assess the incorporation of state-of-the-art cancer treatments into clinical practice. The Health Omnibus Programs Extension of 1988 (Public Law 100-607, Nov. 4, 1988, 102 Stat. 3048) and its amendments require the NCI to establish an information and education program to collect, identify, analyze, and disseminate on a timely basis, through publications and other appropriate means, information on cancer research, diagnosis, prevention, and treatment (Sections 410 and 412 of the Public Health Service Act (42 USC § 285 and 285a-1)).

To disseminate information and data, the National Institutes of Health (NIH) created the NIH Data Sharing Policy and Implementation Guidance (Final NIH Policy for Data Management and Sharing) which will require investigators to submit a data sharing plan beginning January 25, 2023 (https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm). All data, particularly those generated through public funds, should be considered for data sharing. Furthermore, when the data are shared, they should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data. To facilitate data sharing for investigators, the NIH Data Sharing Policy applies to:

- Sharing of final research data for research purposes, especially unique data that cannot be readily replicated
- Basic research, clinical studies, surveys, and other types of research supported by NIH
- Research that involves human subjects and laboratory research that does not involve human subjects
- Applicants seeking $500,000 or more in direct costs in any year of the proposed project period through grants, cooperative agreements, or contracts

A.2 Purpose and Use of the Information Collection

In preparation for dissemination and sharing of data sets, forms requesting or applying for access, upload, share, and store data will be needed. The purpose of data sharing allows data generated from one research study to be used to answer questions beyond the original study. It reinforces open scientific inquiry, encourages diversity of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, and enables the exploration of topics not envisioned by the initial investigators. Biomedical researchers and data scientists can use the NCI cloud resources, web interface, and computational workspaces to query, submit data, analyze, and visualize data. The forms would be used to register a scientist's research data, apply for data storage, and submit a request to access and use the data. In addition to these forms, forms related to metadata information (i.e., related to the collection of the research data; how the data was collected) would be collected for some research (Attachment 1).

Gathering information on the researcher and research project from which data will be submitted is important so that submitting researchers can be assisted throughout the multi-year process of data submission and to notify them of certain expectations for task completion based on data sharing terms. While NCI will define the conditions under which users can access and use data from the NCI Cancer Research Data Commons and the Childhood Cancer Data Initiative data repositories, this generic recognizes the shared need throughout NIH. Information on the proposed research questions or scientific research development will be needed by each respective IC committee within NIH to review and adjudicate data access requests. Requests that are not approved may result in recommendations to support the study data in alternative repositories.

Information on the forms may include a description of the research, how the data was collected, research tools, requestor's credentials, proposed use of the data, and questions to assess whether the data are appropriate for sharing. The data catalogues and file repositories would cover various topics including molecular, cellular, social, and population-based sciences. In addition, ICs may ask questions related to the management and storage of the information so they can provide additional guidance, resources, and support for their respective studies. For NIH funded investigators, all requests would abide by the NIH Data Sharing Policy and Implementation Guidance.

While not inclusive, this is a sampling of the types of forms we anticipate:

- Request Data Access/Use Form
- Data Submission/Storage Form
- Request Access to Use NCI Resources/Software
- Project Renewal or Close-out Form

A.3 Use of Information Technology and Burden Reduction

NCI has developed The NCI Cancer Research Data Commons (CRDC), a cloud-based data science infrastructure that connects data sets with analytics tools to allow users to share, integrate, analyze, and visualize cancer research data to drive scientific discovery. This cloud-based data science infrastructure connects data sets/repositories with analytical tools. This online system was developed to register studies, submit data, and request access to data. Using an online system allows investigators to submit the required information directly, thereby minimizing burden not only for investigators and institutions, but also for NIH staff. The online system uses time-saving features, such as the use of pull-down and scrolling menus to fill data fields, "find as you type" (or "type ahead") functionality, and text fields that allow investigators and requesters to cut and paste information from other sources.

The NCI Privacy Act Coordinator was consulted, and it was determined that a Privacy Impact Assessment (PIA) will be required when personally identifiable information is being uploaded, stored, or accessed. For each sub-study requiring a PIA, a copy of the PIA that has been submitted to the NIH Privacy Act Coordinator will be included in the submission.

A.4 Efforts to Identify Duplication and Use of Similar Information

An extensive search to identify duplication and similar information collections as is proposed here was undertaken. There were three recent submissions found. The NCI Genomic Data Commons (GDC) Data Submission Request Form – Center for Cancer Genomics (CCG) (OMB No. 0925-0752, Expiration Date 3/31/2023) submission has a GDC Data Submission Request Form to provide a mechanism in which investigators can submit data for studies into the NCI GDC . This form determines in a structured fashion whether the study is applicable to the cancer genomics research only and whether the GDC has the available resources to support the management of study data. The GDC Data Submission Review Committee also uses the form to review and assess the data submission request for applicability to the GDC mission.

The National Institute of Mental Health Data Archive (NDA), NIMH (OMB No. 0925-0667, Expiration Date 01/31/2024) have two forms. The Data Submission Agreement and Data Use Certification forms collect information about researchers submitting data and requesting access to shared data in the NIMH Data Archive (NDA). These forms collect research data and information about the researchers that is solely focused on the prevention, cause, diagnosis, and treatment of mental health, not cancer-related data.

In addition, there is the NIH Information Collection Forms to Support Genomic Data Sharing for Research Purposes (OMB No. 0925-0670, Expiration Date 11/30/2022) which has an access request to a Database for Genotypes and Phenotypes (dbGaP), study registration, renewal and close-out forms related to the sharing of human and non-human genomic data from large-scale genomic research studies. The infrastructure that NCI has created expands beyond genomics and includes clinical, imaging, and population-science data. Thus, neither of two full submissions would encompass the broad scope of NCI's proposed submission.

The change request approved by OMB on 9/28/2021 for the Generic Clearance for NIH Citizen Science and Crowdsourcing Projects (NIH) (OMB No. 0925-0766, Expiration Date 4/30/2023) allows for submitting applications and forms to access data. NCI proposal goes beyond requests to access data to includes forms for uploading, submitting, storing, and sharing data and access to data-related tools such as workspaces and analytic tools and software.

A.5. Impact on Small Businesses or Other Small Entities

No small businesses or other small entities will not be impacted.

A.6 Consequences of Collecting the Information Less Frequently

Following the initial request and approval to use controlled-access data, some requesters must provide annual updates on their research progress and renew access to the dataset(s) for another year or close-out access to the dataset(s). The consequence of not submitting the required information annually is a reduction in NIH oversight in the use of data, possibly leading to an increased rate of adverse data management incidents.

A.7 Special Circumstances Relating to the Guidelines of 5 CFR 1320.5

This collection fully complies with 5 CFR 1320.5.

A.8.1 Comments in Response to the Federal Register Notice

The 60-Day Federal Register Notice was published on December 20, 2021, Vol, 86, Page 71901 and allowed 60 days for public comment. No public comments were received.

A.8.2 Efforts to Consult Outside Agency

We have not consulted with any outside agency on this project.

A.9 Explanation of Any Payment of Gift to Respondents

No incentives (neither payments nor gifts) will be distributed to individuals or institutions.

A.10 Assurance of Confidentiality Provided to Respondents

Prior to sharing, data should be redacted to strip all identifiers, and effective strategies should be adopted to minimize risks of unauthorized disclosure of personal identifiers. Stripping a dataset of items that could identify individual participants is referred to by several different terms, such as "data redaction," "de-identification of data," and anonymizing data. In addition to removing direct identifiers, e.g., name, address, telephone numbers, and Social Security Numbers, researchers should consider removing indirect identifiers and other information that could lead to "deductive disclosure" of participants' identities. Deductive disclosure of individual subjects becomes more likely when there are unusual characteristics of the joint occurrence of several unusual variables. Samples drawn from small geographic areas, rare populations, and linked datasets can present challenges to the protection of subjects' identities.
Investigators may use different methods to reduce the risk of subject identification. One possible approach is to withhold some part of the data. Alternatively, an investigator may restrict access to the data at a controlled site, sometimes referred to as a data enclave. Some investigators may employ hybrid methods, such as releasing a highly redacted dataset for general use but providing access to more sensitive data with stricter controls through a data enclave.

Researchers who seek access to individual level data are typically required to enter into a data-sharing agreement. Data-sharing agreements, which come by many terms, including "license agreements," and "data distribution agreements," generally include requirements to protect participants' privacy and data confidentiality. They may prohibit the recipient from transferring the data to other users or require that the data be used for research purposes only, among other provisions, and they may stipulate penalties for violations.

For access to and submission of data, researchers are both NIH-funded and non-NIH-funded investigators. Making these researcher's names available is an important ethical underpinning of the NIH GDS Policy as it allows NIH to be transparent in informing research participants, the scientific community, and the public on how data are being shared, with whom, and for what research purpose in addition to fostering future research collaborations.

The names and institutional affiliations of the researchers (both data submitters and data requesters) may be posted publicly on a website, and thus there is no assurance of confidentiality afforded to the researchers. However, it is important to emphasize that no personal information is requested from researchers submitting or accessing data beyond their name and institutional affiliation. Data submitters are largely NIH-funded investigators whose names and institutional affiliations are already a matter of public record. All information will be kept secure to the extent allowable by law.

The Privacy Act is applicable as determined by the NIH Privacy Officer in the Privacy Act Memo (Attachment 2). This data collection is covered by the following Privacy Act System of Records:

- 09-25-0200, "Clinical, Basic and Population-based Research Studies of the National Institutes of Health (NIH), HHS/NIH/OD."
- 09-25-0036, "Extramural Awards and Chartered Advisory Committees (IMPAC 2)"
- 09-90-1401, "Records About Restricted Dataset Requesters"

A.11 Justification for Sensitive Questions

No questions of a sensitive nature are included in this data collection.

A.12.1 Estimates of Hour Burden Including Annualized Hourly Costs

The total estimated annualized burden hours are: 5,775 and the number of responses per respondent is estimated at 13,500 (Table A.12-1). The burden hours for the requested 3 years will be 17,325 with 40,500 responses.

Table A.12-1 Estimated Annualized Burden Hours

Form Name	Type of Respondents	Number of Respondents	Number of Responses per Respondent	Average Burden Per Response (in hours)	Total Annual Burden Hours
Request Data Access/Use					
Data Access Request - Submitter	Individuals	1,500	1	45/60	1,125
Institutional Certification	Individuals	1,500	1	30/60	750
Data Submission/Storage					
Data Submission/ Storage Request	Individuals	1,500	1	30/60	750
Institutional Certification	Individuals	1,500	1	30/60	750
Request Access to/Use NCI Resources/Software					
Data Resources	Individuals	1,500	1	30/60	750
Project Renewal or Project Close-out					
Project Renewal or Project Close-out form	Individuals	1,500	2	15/60	750
Institutional Certification	Individuals	1,500	2	18/60	900
Totals		**10,500**	**13,500**		**5,775**

A.12-2 Annualized Cost to respondents

The total estimated annual cost to respondents is $290,550.75. Table A12-1 illustrates the measures respondents will be answering and the number of hours to complete each type (Table A.12-2).

Table A.12-2 Annualized Cost to the Respondents

Type of Respondents	Total Annual Burden Hours	Hourly Respondent Wage Rate*	Respondent Cost
Request Data Access/Use			
Individuals - Submitter	1,125	$48.45	$54,506.25
Individuals - Institutional Official	750	$52.93	$39,697.50
Data Submission/Storage			
Individuals - Submitter	750	$48.45	$36,337.50
Individuals - Institutional Official	750	$52.93	$39,697.50
Request Access to/Use NCI Resources/Software			
Data Resources	750	$48.45	$36,337.50
Project Renewal/Project Close-out Process			
Individual - Submitter	750	$48.45	$36,337.50
Individuals - Institutional Official	900	$52.93	$47,637.00
Total			**$290,550.75**

*The median wage rate was calculated using the most recent data from Bureau of Labor Statistics for occupation code "19-1040" occupation title "Medical Scientists", rate of $48.45 and Occupation Code "19-2099" occupation title "Physical Scientists, All Other", rate of $52.93, https://www.bls.gov/oes/current/oes_nat.htm.

A.13 Estimate of Other Total Annual Cost Burden to Respondents or Record Keepers

There are no other costs to respondents other than their time.

A.14 Annualized Cost to the Federal Government

The annualized cost to the Federal Government for the collection of data is $16,900.54. The tasks performed by Federal Personnel include reviewing the data submission, access/use, storage, and resources request forms, evaluating the requests against the requirements and resources available, determining whether the request should be accepted or not, and overseeing the contractors.

Table 14.1 Annualized Cost to the Federal Government

Staff	Grade/Step	Salary**	% of Effort	Fringe (if applicable)	Total Cost to Gov't
Federal Oversight					
Project Officer	14/10	$164,102	2%		$3,282.04
Director	SES	$284,625	2%		$5,692.50
Program Director	15/10	$176,300	2%		$3,526.00
Contractor Cost					$4,400.00
Travel					$0
Other Cost					$0
Total					**$16,900**

**The Salary in the table above is cited from https://www.opm.gov/policy-data-oversight/pay-leave/salaries-wages/salary-tables/22Tables/html/DCB.aspx

A.15 Explanation for Program Changes or Adjustments

This is a new generic information collection.

A.16 Plans for Tabulation and Publication and Project Time Schedule

There are no plans for tabulation or publication for the information submitted on the requests. Below is the list of activities and a timeline.

A.16 - 1 Project Time Schedule	
Activity	Time Schedule
Data request forms available	1 week after OMB approval
Completion of review and response to forms	2-3 weeks after investigator submits request

A.17 Reason(s) Display of OMB Expiration Date is Inappropriate

We are not requesting exemption from the display of the OMB expiration date.

A.18 Exceptions to Certification for Paperwork Reduction Act Submissions

This information collection will comply with the requirements in 5 CFR 1320.9.