# National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Worksheet for Drafting a Data Management and Sharing Plan for Scientific Data

The National Institutes of Health (NIH) Data Management and Sharing (DMS) policy ([NOT-OD-21-013](https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html)) requires investigators to submit a DMS Plan for research that will generate scientific data as described by NIH. The DMS Plan should describe how the data will be managed and appropriately shared and address the six elements of DMS Plans:

1) Data Type

2) Related Tools, Software, and/or Code

3) Standards

4) Data Preservation, Access, and Associated Timelines

5) Access, Distribution, or Reuse Considerations

6) Oversight of Data Management and Sharing

Complementary to NIH policy, NIDDK has established [Institute-specific guidance](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/guidance-writing-dms-plan) for each DMS element. This worksheet is designed to assist investigators as they compose their initial DMS Plan. The worksheet aligns with the NIH optional [DMS Plan Format Page](https://grants.nih.gov/grants/forms/data-management-and-sharing-plan-format-page) and provides links to the Institute-specific DMS guidance on the NIDDK webpage for each DMS element. Additional links to NIDDK tools designed to assist investigators with specific data elements are supplied in the relevant sections.

While drafting the DMS Plan, investigators are encouraged to review the NIDDK-specific guidance and refer to the instructions included in the current version of the [NIH Application Guide](https://grants.nih.gov/grants/how-to-apply-application-guide.html) ([research](https://grants.nih.gov/grants/how-to-apply-application-guide/forms-h/general/g.400-phs-398-research-plan-form.htm#11), [career development](https://grants.nih.gov/grants/how-to-apply-application-guide/forms-h/general/g.410-phs-398-career-development-award-supplemental-form.htm#17)) for developing a plan, and any additional guidance on [sharing.nih.gov](https://sharing.nih.gov/).

This worksheet is intended as a resource for investigators and does not impose requirements. It provides additional considerations that align with the essence of the 2023 NIH DMS Policy for each element and may be helpful when developing an initial DMS Plan or revising the Plan throughout the project period as information about the scientific data being generated, managed, and shared evolves. NIDDK-specific expectations are not included in this worksheet. NIDDK-specific expectations will be communicated in a Notice of Funding Opportunity (NOFO) when applicable.

Note that plan elements are not mutually exclusive and considerations for one element may be relevant for other elements. In this instance, information may be recorded for only one element, or if corresponding information is recorded under more than one element then the information must align. For example, use of metadata is a consideration for several elements although metadata should be listed under Element 1.c.

Examples provided in the [Appendix](#_Appendix:_Examples) are solely for explanatory purposes and expand on rationale for adopting data management and sharing best practices. In addition to this worksheet, NIDDK investigators are encouraged to review the four [DMS Plan Examples for common data types](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/dms-tools-examples#dmsp) to understand the minimum level of detail that investigators should provide. Additional plan examples are available from [NIH](https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/writing-a-data-management-and-sharing-plan#sample-plans).

## Data Management and Sharing Plan

### Element 1: Data Type

[NIDDK Guidance for Data Type](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/guidance-writing-dms-plan#data-type)

##### Types and amount of scientific data expected to be generated in the project:

###### Summarize the types and estimated amount of scientific data expected to be generated in the project.

* 1. Types of Scientific data:
     + Structured Data: Research specifics, coded survey responses, standardized interview responses.
     + Images: Microscopic images, photographic data, radiographic imaging.
     + Sensor Readings: Environmental data, physiological measurements.
     + Genetic Sequences: DNA/RNA sequences, genomic data.
     + Numerical Data: Statistical models, simulation data.
     + Textual Data: audio/video transcripts, Short Message System (SMS).
  2. Amount of Data – For each type of data mentioned, estimate the amount or volume that will be generated using metrics such as:
     + File Sizes: Estimate the size of data files (e.g., gigabytes, terabytes).
     + Number of Records: For database-oriented data, estimate the number of records or entries.
     + Frequency of Data Capture: Consider how often data will be collected (e.g., daily, throughout a specific phase of the project).
     + Data Complexity: Determine if the data require significant computational resources for processing or analysis.
  3. Data Formats and Quality
     + Consider the formats in which data will be recorded and stored, determining the choice of these formats based on data integrity, reusability, and compatibility with commonly used tools in the field.
     + Think through the measures that will be taken to ensure data quality and integrity throughout the data lifecycle.

##### Scientific data that will be preserved and shared, and the rationale for doing so:

###### Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.

Specify any data preservation and sharing policies or guidelines that will be followed. Consider data formats, repositories, access control, versioning, and any relevant licenses or agreements.

1. Data to be Preserved and Shared
   * Detail the specific types of data that will be preserved and shared (e.g., raw data, processed data).
   * Criteria used to determine which data will be preserved and shared might include the data's potential for reuse, its role in supporting the findings reported in publications, and its relevance to the broader field.
2. Rationale for Preservation and Sharing may include:
   * Supporting the transparency of research processes and enabling other researchers to reproduce and validate findings.
   * Fostering collaborations within the scientific community, leading to new research initiatives and innovations.
   * Funding agency requirements or industry standards that mandate data sharing. For NIH-funded research, reference the NIH Data Management and Sharing Policy.
   * Preserving and sharing data contributes to the advancement of science, potentially leading to breakthroughs or significant advancements in the field.

NOTE: DMS policy expects that sharing of scientific data will be maximized. For any data that will not be shared, justification must be included explaining why those data will not be shared, specifically any ethical, legal, or technical factors that may preclude sharing.

##### Metadata, other relevant data, and associated documentation:

###### Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

Metadata provides descriptive information about the data, facilitating its understanding, discovery, and usability.

* This could include documentation, research protocols, code, algorithms, data collection protocols, data processing workflows, data dictionaries, data cleaning procedures, or any other relevant documentation or supporting material.
* Consider the purpose and intended audience, ensuring metadata are comprehensive and easily understandable for both project members and potential secondary data users.
* Appreciate that some repositories may require specific metadata standards and formats in order to deposit data.

When defining a schema for metadata and relevant data in the context of scientific research, consider the following elements:

1. Identification and Descriptive Metadata:

* Title: The title or name of the dataset.
* Description: A concise summary describing the content and purpose of the dataset.
* Creator/Author: The individuals or organizations responsible for creating the dataset.
* Date Created: The date when the dataset was initially generated or compiled.
* Date Updated: The date of the most recent update or modification to the dataset.
* Keywords: Relevant keywords or phrases that capture the dataset's main topics or themes.

1. Technical Metadata:

* Format: The file format(s) used to store the dataset, such as CSV, JSON, XML, or specific domain-specific formats.
* Size: The size of the dataset, typically measured in bytes or another appropriate unit.
* Structure: The structure or organization of the dataset, including the number and types of fields or columns.
* Data Dictionary: Definitions and descriptions of the fields or columns in the dataset, explaining their meaning and format.
* Units of Measurement: If applicable, the units of measurement used in the dataset.

1. Provenance and Contextual Metadata:
   * Data Source: The origin or source of the dataset, including information about data collection methods or instruments used.
   * Experimental Conditions: Details about the experimental setup or conditions under which the data was collected, if applicable.
   * Sampling Methodology: Information on the sampling approach used to collect or select the data.
   * Ethical Considerations: Documentation of any ethical considerations, such as informed consent or human subjects' protections.
   * Related Publications: References to published papers or articles related to the dataset.
2. Usage and Access Metadata:

* Access Rights: The level of access and any restrictions on data usage, such as open access, restricted access, or data use agreements.
* Licensing: The type of license under which the dataset is released, such as Creative Commons or proprietary licenses.
* Versioning: Keep track of different versions of the dataset, including version numbers, release dates, and descriptions of changes.

1. Relationship Metadata:
   * Related Datasets: Any other datasets that are related or linked to the current dataset, along with the nature of their relationship.
   * Data Dependencies: Any dependencies between datasets, including prerequisites or requirements for using the current dataset.

### Element 2: Related Tools, Software and/or Code

[NIDDK Guidance for Related Tools, Software, and/or Code](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/guidance-writing-dms-plan#tool)

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

In addition, recognize the importance of good documentation practices and specify where the documentation for the software and tools can be found.

Enumerate the tools, software, and code necessary to access or manipulate the shared scientific data. If appropriate, provide a brief description of each, specifying its role and importance in your research.

* For example, include specific names and versions to avoid ambiguity. For instance, list software like Python 3.8 for data analysis, R for statistical processing, or specialized bioinformatics tools like BLAST for sequence analysis.

For specialized tools, software, and/or code listed, briefly describe where and how they can be accessed:

1. Distinguish between open-source tools (which can be freely accessed and used) and proprietary tools (which require licenses or purchases).
   * For open-source software, specify the type of license (e.g., MIT, GPL).
   * For proprietary software, outline the terms under which the software is shared within and potentially outside the research community.
   * To ensure that other researchers can replicate your setup without significant barriers, when appropriate, briefly mention any guidance documents or available technical support for the installation and configuration of the tools and software.
2. Provide specifics for download pages, repositories for code (e.g., GitHub), or instructions for accessing proprietary software (e.g., through institutional licenses or direct purchase).
   * If your project involves complex tools or software, consider providing information about available training resources when appropriate. This could include workshops, online tutorials, or manuals that can help other investigators use the tools effectively.

Consider how the use of these tools complies with legal and regulatory standards, particularly those pertaining to data privacy and security.

### Element 3: Standards

[NIDDK Guidance for Standards](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/guidance-writing-dms-plan#standards)

[NIDDK Resource - Standards Examples](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/dms-tools-examples#metadata)

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist.

1. Identification of Relevant Data Standards
   * Identify and list all data standards that will be applied to the scientific data and metadata generated by the project. This can include industry-specific standards, open data standards, and any other relevant protocols.
     + For example, for genomic research, you might list standards such as FASTQ for raw sequence data, BAM for aligned sequencing data, and VCF for variant calls. For some clinical data (e.g., real world data, electronic health record data), standards like HL7 or FHIR might be relevant.
2. Description of Data Standards Application
   * Specify how each data standard will be applied to different types of data and metadata in the project, and when appropriate, detail the processes or tools that will be used to format, store, and share data according to these standards.
   * Recognize how applying these standards will facilitate interoperability with other datasets and resources, both within and outside the research community.
3. Implementation Strategy
   * Consider tools or software that will be used to implement the data standards and how these tools support the standards and any configuration required.
   * Think through the procedures for ensuring that all data collected or generated in the project conforms to the specified standards. This may include data validation checks or quality assurance processes.
   * Recognize training needs that should be provided to research team members to familiarize them with the data standards and the importance of compliance.
   * Consider potential challenges or limitations in applying these data standards, including technological, logistical, or financial constraints, along with any solutions or alternatives for overcoming these challenges.
4. If No Consensus Standards Exist
   * If no consensus standards exist for some types of data or metadata in the project, clearly state this.
   * Describe any alternative approaches you plan to use to ensure data quality and interoperability in the absence of standard protocols.

### Element 4: Data Preservation, Access, and Associated Timelines

[NIDDK Guidance for Data Preservation, Access, and Associated Timelines](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/guidance-writing-dms-plan#preservation)

[NIDDK Resource - Repository Selection Considerations Tool](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/dms-tools-examples#preservation)

##### Repository where scientific data and metadata will be archived:

###### Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived (see [Selecting a Data Repository](https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository)).

1. Selection of Repository
   * When evaluating a repository, consider its relevance to the field of study and its credentials (e.g., compliance with standards, certifications).
   * Criteria to select a repository include data type specialization, reputation, accessibility, sustainability, and compliance with funder (NIH) requirements.
   * The selected repository should ideally reflect the [desirable characteristics of data repositories](https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository#desirable-characteristics-for-all-data-repositories) and any other relevant federal regulations.
   * Consider processes involved in preparing data for archiving, such as data cleaning, anonymization (if necessary), and formatting according to the repository’s specifications.
2. Access Conditions
   * When selecting a repository, based on the nature of the data, determine whether the data will be open access (available to all) or controlled access (restricted to certain users), and the rationale for this choice and any privacy considerations, commercial interests, or other sensitivities.
   * If applicable, be familiar with any Data Use Agreements (DUAs) that outline the terms and conditions for external use of the data.
   * Consider any licensing agreements under which the data will be shared. This might include Creative Commons licenses for datasets.

##### How scientific data will be findable and identifiable:

###### Describe how the scientific data will be findable and identifiable (i.e., via a persistent unique identifier or other standard indexing tools).

The following practices support the “Findable” aspect of the FAIR (Findable, Accessible, Interoperable, Reusable) principles.

1. Implementation of Persistent Unique Identifiers (PIDs)
   * Specify the types of persistent identifiers that will be assigned to the datasets. Common PIDs include DOIs (Digital Object Identifiers), URNs (Uniform Resource Names), and ARKs (Archival Resource Keys).
   * Describe the process by which PIDs will be assigned. This could involve collaboration with data repositories that automatically assign PIDs upon data submission.
   * Integration with metadata: Explain how PIDs will be linked to the metadata to ensure that each dataset can be easily located and referenced.
2. Metadata Standards
   * Consider the metadata standards or schemas that will be used (e.g., Dublin Core, DataCite, MODS) and how these contribute to the findability of the resources. These schemas should be appropriate for the data types and disciplines involved in the research.
3. Indexing and Cataloging
   * Mention any catalogs or databases in which your data will be listed. These might include institutional repositories, specialized data registries, or national/international data indexes.
   * List any tools or platforms that will be used to index the data, making it searchable across multiple platforms. Examples might include CrossRef, Scopus, or subject-specific platforms like PubMed for biomedical data.
   * Mention if any academic libraries or data repositories are being considered to aid in the indexing and cataloging of the data.
   * Note any steps taken to ensure that datasets are discoverable via search engines, such as the use of structured data and metadata that are favorable to search algorithms.
4. Training and Guidance
   * Consider, and mention when appropriate, any training or resources that will be provided to researchers to help them understand and implement these practices effectively.

##### When and how long the scientific data will be made available:

###### Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

1. Timing of Data Availability
   * Consider the timeline for when data will be deposited in the repository during the project. This timeline should align with the project milestones and NIH requirements (e.g., GDS submission/release expectations).
   * When data will be made available no later than the time of publication of the associated research findings, detail the process for synchronizing data release with publication, such as including data availability statements or links within published articles. Recognize that a repository may allow for an embargo period in limited situations.
   * If data will be released before publication or made available prior to the end of the performance period, explain the procedural steps and preparations required to meet this timeline.
2. Duration of Data Availability
   * Refer to the guidelines of the selected data repository, which may dictate the duration of data preservation and accessibility.
   * Outline the minimum period during which the data will be available. Data should generally be accessible for a certain number of years after the conclusion of the project or publication or longer if warranted by the potential scientific value.
   * Mention any agreements or policies that ensure long-term data availability, such as transferring data to a long-term archival service or maintaining it within institutional archives.
   * Mention any procedures for periodically reviewing the data's relevance and the need for continued availability, including who will be responsible for this review and how decisions on extended preservation will be made.

### Element 5: Access, Distribution, or Reuse Considerations

[NIDDK Guidance for Access, Distribution, or Reuse Considerations](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/guidance-writing-dms-plan#access)

##### Factors affecting subsequent access, distribution, or reuse of scientific data:

###### NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See [Frequently Asked Questions](https://sharing.nih.gov/faqs#/data-management-and-sharing-policy.htm) for examples of justifiable reasons for limiting sharing of data.

Consider the following factors that could affect the subsequent access, distribution, or reuse of scientific data.

1. Informed Consent
   * Recognize how the informed consent process impacts data sharing based on what participants were told about the uses of their data, including future research and potential sharing with other researchers.
   * Justify any limitations on data sharing based on the specifics of the consent agreements. For example, if participants only consented to their data being used for a specific study, explain how this affects the ability to share data for other purposes.
2. Privacy and Confidentiality Protections (NOTE: Ensure consistency with information provided in section 5-C below.)
   * Outline the measures taken to protect participant privacy and confidentiality, such as data anonymization or pseudonymization. Mention the extent to which these data de-identification measures enable or restrict data sharing. Refer to the NIH supplemental information to the DMS Policy on protecting privacy: [NOT-OD-22-213](https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-213.html) and consider how the approach aligns with these principles and best practices.
   * Note any legal or regulatory requirements (e.g., The Health Insurance Portability and Accountability Act (HIPAA) that influence how data can be shared, and how these requirements impact the access, distribution, and reuse of data.
3. Intellectual Property and Proprietary Concerns
   * State any intellectual property rights (e.g., patents) that might limit data sharing. Note any circumstances under which data might be withheld until patent applications are filed or other intellectual property considerations are resolved.
   * Consider any agreements with collaborators or third-party entities and ensure that these agreements comply with NIH policies and do not impose unnecessary restrictions on data sharing.
4. Other Considerations
   * Mention any issues related to cultural sensitivity or the rights of indigenous communities that may impose restrictions on how data is shared. Refer to [NOT-OD-22-214, Supplemental Information on Responsible Management and Sharing of American Indian/Alaska Native Participant Data](https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-214.html).
5. Strategies to Maximize Access
   * Consider how Data Sharing Agreements (DSAs) will be used to safely share data while respecting the constraints identified.
   * Consider any plans to implement tiered access to data, where different levels of data access are granted based on user credentials or the sensitivity of the data.

##### Whether access to scientific data will be controlled:

###### State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

It is important to recognize the conditions under which data will be restricted and the mechanisms for managing controlled access. Controlled access is often necessary for sensitive data where unrestricted availability could compromise participant privacy, proprietary information, or compliance with regulatory requirements. Consider the following when structuring your DMS plan:

1. **Rationale for Controlled Access**
   * **Appreciate the types of data considered sensitive and why they require controlled access. This may include personal health information (PHI), identifiable human data, or data from vulnerable populations.**
   * **Cite specific regulatory or ethical guidelines that necessitate controlled access, such as HIPAA (for health information in the U.S.) or other relevant national/international regulations.**
2. **Mechanisms for Controlled Access**
   * **Understand the process for obtaining access to the data, including who will review access requests (e.g., an internal review board, data access committee), the criteria for approval, and the expected timeline for decision-making.**
     + Investigators may not retain custodianship or exclusively control access to data once those data should be shared via a public repository.
   * **Consider how Data Use Agreements will be employed to ensure that data users agree to comply with terms and conditions that protect data confidentiality and usage restrictions.**
   * **Assess the systems in place for verifying user identities and authorizing access based on user qualifications or the intended use of the data.**
3. **Data Repository Role**
   * **Consider any specific features of the repository that support controlled access, such as secure data environments or user management systems.**
   * **Evaluate the services provided by the repository related to controlled access, such as secure data transfer mechanisms, encryption, and user activity logging.**
4. **User Responsibilities and Compliance**
   * **Recognize guidelines that users must follow when accessing controlled data, including data security practices, restrictions on data sharing, and obligations to protect data privacy.**
   * **Consider how compliance with the terms of data use will be monitored and enforced by either the data repository or your institution.**
5. **Limitations and Conditions of Use**
   * **Understand any specific conditions or limitations on the use of the data, such as prohibitions on attempting to re-identify anonymized data or limitations on the types of analyses that can be conducted.**
   * **Be aware of how long approved users will have access to the data and under what conditions access might be revoked or expire.**

##### Protections for privacy, rights, and confidentiality of human research participants:

###### If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

1. De-identification Techniques
   * Describe the specific techniques and processes used to de-identify data, such as removing personal identifiers, masking parts of the data, or using advanced techniques like differential privacy.
   * Explain any tiered levels of de-identification, if applicable, and how these correspond to different levels of data sensitivity and intended use.
2. Certificates of Confidentiality
   * Describe situations in which Certificates of Confidentiality will be obtained, particularly if the research involves sensitive information that could potentially harm participants if disclosed.
   * Consider how these certificates protect against compelled disclosure of identifiable research information, and any limitations.
3. Additional Protective Measures
   * Specify any systems and protocols for controlling access to data, such as secure data environments, use of strong authentication practices, and role-based access controls.
   * Note use of encryption and privacy enhancing technology for data storage and transmission to protect data confidentiality.
   * Mention the use of DSAs to legally bind other parties to adhere to confidentiality and privacy standards.
4. Compliance with Regulations
   * For studies involving PHI under U.S. jurisdiction, describe how the project complies with HIPAA requirements for the protection of health information.
   * If applicable, discuss compliance with Tribal consultation and agreements or international regulations, which may involve additional protections or rights for participants, such as the right to data erasure.
5. Oversight and Monitoring
   * Understand the role of the Institutional Review Board (IRB) or equivalent in reviewing and approving the methods for protecting participant privacy and data confidentiality.
   * Think through the procedures for ongoing monitoring of data protection measures, including regular audits or assessments.
6. Training and Education
   * Consider training requirements for researchers and staff on handling sensitive data, including training on ethical considerations and specific technologies or protocols used in the project. Confirm documentation and policies on privacy protections are easily accessible to all relevant stakeholders.

### Element 6: Oversight of Data Management and Sharing

[NIDDK Guidance for Oversight of Data Management and Sharing](https://www.niddk.nih.gov/research-funding/research-resources/data-management-sharing/guidance-writing-dms-plan#oversight)

Describe how compliance with this Plan will be monitored and managed, the frequency of oversight, and by whom at your institution (e.g., titles, roles).

The DMS plan must ensure that all stakeholders are aware of the mechanisms in place to enforce the plan and maintain high standards of data management and sharing. Consider the following for your DMS plan:

1. **Oversight Responsibilities**
   * **Identify the individuals, committees, or departments responsible for overseeing the DMS plan. This might include roles such as the Data Manager, Principal Investigator, Institutional Review Board (IRB), or a specific Data Oversight Committee.**
   * **Describe the specific responsibilities of each party. For example, the Data Manager might oversee daily data handling practices, while the IRB might focus on compliance with ethical standards.**
2. **Monitoring Mechanisms**
   * **Describe the routine monitoring mechanisms in place, such as regular audits of data storage locations, data access logs, and compliance with data sharing agreements. This includes** mechanisms to monitor compliance with data sharing and preservation requirements and that will be reported to NIH.
   * **Mention how and when compliance reviews will be conducted. Note what aspects of the DMS plan are reviewed during these checks, such as adherence to data encryption standards, proper use of data identifiers, and timely data deposition.**
   * **Consider steps that may be taken in the event of non-compliance with the DMS plan, including the process for investigating issues, who is responsible for deciding corrective actions, and how these actions are implemented and monitored.**
   * **Recognize that feedback from oversight activities is necessary to improve data management and sharing practices over time.**
3. **Frequency of Oversight**
   * **Define the frequency of regular oversight activities, such as quarterly or bi-annual reviews.**
   * **Explain any circumstances that would trigger additional reviews, such as the initiation of a new collaboration involving data sharing, reported compliance issues, or updates to regulatory requirements.**
4. **Reporting and Documentation**
   * **Outline the procedures for reporting the findings of oversight activities. Specify who receives these reports (e.g., Principal Investigator, funding agencies, institutional officials) and how often.** State how and where the data availability details will be documented, such as in annual reports to NIH’s RPPR
   * **Determine how documentation related to oversight activities is maintained, where this documentation is stored, and who has access to it.**
5. **Training and Support**
   * **Consider any training programs provided to staff involved in data management and sharing to ensure they understand their roles and the importance of compliance, including any resources available to support staff in implementing the DMS plan effectively, such as technical support or access to data management tools.**

## Appendix: Examples

### Element 1: Data Type

1. **Types and amount of scientific data expected to be generated in the project:**

*Summarize the types and estimated amount of scientific data expected to be generated in the project.*In this project, we anticipate generating several types of scientific data, primarily consisting of the following:

1. **Genomic Sequencing Data:**
   * Type: Whole-genome sequencing (WGS) and RNA-seq data.
   * Volume: Approximately 3 terabytes (TB) of raw sequencing data will be generated from 500 human samples. Each sample is expected to yield around 6 gigabytes (GB) of WGS data and 1 GB of RNA-seq data.
   * Format: Data will be stored in FASTQ, BAM, and VCF formats, with associated metadata in JSON and CSV formats.
2. **Clinical Data:**
   * Type: De-identified clinical data including participant demographics, laboratory results, and treatment outcomes.
   * Volume: About 200 megabytes (MB) of data will be generated, covering 500 participants.
   * Format: Data will be stored in electronic health record (EHR) compatible formats such as HL7 FHIR and CSV.
3. **Imaging Data:**
   * Type: MRI and CT scan images related to disease progression and treatment response.
   * Volume: Approximately 5 TB of imaging data will be generated from 100 participants, with each MRI scan producing about 50 GB of data.
   * Format: Data will be stored in DICOM format, with metadata in XML and JSON.
4. **Proteomics Data:**
   * Type: Mass spectrometry-based proteomic profiles.
   * Volume: Expected to generate around 500 GB of data from 200 samples.
   * Format: Data will be stored in raw format (vendor-specific), with processed data in MGF and mzML formats.
5. **Behavioral Data:**
   * Type: Survey responses and observational data on participant behaviors.
   * Volume: About 50 MB of data will be collected from 300 participants.
   * Format: Data will be stored in CSV and Excel formats.

In total, the project is expected to generate approximately 8.75 TB of data across various scientific domains. This data will be managed and stored securely following NIH guidelines and will be made accessible to the broader research community through public repositories, ensuring compliance with data sharing and preservation standards.

1. **Scientific data that will be preserved and shared, and the rationale for doing so:**

*Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.*  
  
In this project, the following scientific data will be preserved and shared:

1. **Genomic Sequencing Data:**
   * Data to be Preserved and Shared: The processed whole-genome sequencing (WGS) and RNA-seq data, including variant call files (VCF) and aligned sequence data in BAM format.
   * Rationale: The processed genomic data are critical for understanding genetic variations and gene expression profiles associated with the studied condition. Sharing these data will enable other researchers to validate findings, conduct meta-analyses, and explore new hypotheses, thereby accelerating advancements in precision medicine. Raw sequencing data will also be preserved but will be shared upon request due to its large volume and the need to protect participant privacy.
2. **Clinical Data:**
   * Data to be Preserved and Shared: De-identified clinical data, including participant demographics, laboratory results, and treatment outcomes.
   * Rationale: Clinical data provide essential context for interpreting genomic and other omics data, allowing for the development of integrated models of disease. Preserving and sharing these data will support efforts in translational research, particularly in validating clinical biomarkers and treatment protocols. The data will be shared in compliance with HIPAA and other relevant regulations.
3. **Imaging Data:**
   * Data to be Preserved and Shared: Selected MRI and CT scan images associated with key findings, with associated metadata.
   * Rationale: Imaging data offer critical insights into disease progression and treatment response. By sharing these data, we aim to support the development of imaging biomarkers and facilitate cross-study comparisons. Only images that are directly relevant to published findings will be shared to ensure a focus on scientifically valuable data, while the rest will be preserved in an institutional repository.
4. **Proteomics Data:**
   * Data to be Preserved and Shared: Processed proteomics data, including peak lists in mzML format and protein identification data.
   * Rationale: Proteomic data are essential for understanding protein expression and modifications in the studied condition. Sharing these data will allow researchers to compare protein expression patterns across studies and develop new therapeutic targets. The raw data will be preserved but shared upon request, as the processed data are more immediately useful for most researchers.
5. **Behavioral Data:**
   * Data to be Preserved and Shared: Aggregated survey data and de-identified observational data.
   * Rationale: Behavioral data contribute to understanding patient adherence and response to treatment, which are crucial for improving health outcomes. Sharing these data will enable other researchers to analyze behavioral trends and their impact on treatment efficacy. Individual-level data will be preserved with access controls to protect participant privacy.
6. **Metadata, other relevant data, and associated documentation:**

*Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.*  
  
To ensure that the scientific data generated by this project are interpretable and reusable by the broader research community, the following metadata, relevant data, and associated documentation will be made accessible:

1. **Genomic Data Metadata:**
   * Details: Metadata will include detailed descriptions of sequencing protocols, sample preparation methods, quality control metrics, sequencing platform details, and data processing workflows (e.g., alignment, variant calling).
   * Documentation: A comprehensive data dictionary describing each file format (FASTQ, BAM, VCF) and the specific fields within those files will be provided. Additionally, the software and versioning information for all bioinformatics tools used in data processing will be documented.
2. **Clinical Data Metadata:**
   * Details: Metadata will encompass data collection protocols, coding schemes (e.g., ICD-10 codes for diagnoses), and data harmonization processes.
   * Documentation: The project will share study protocols, informed consent forms, and a detailed codebook that defines all variables, data ranges, and units of measurement. This documentation will ensure that the clinical data can be accurately interpreted and linked with other datasets.
3. **Imaging Data Metadata:**
   * Details: Metadata will include imaging acquisition parameters (e.g., scanner model, resolution, imaging sequences), imaging processing details (e.g., software used, filtering techniques), and annotations for regions of interest (ROI).
   * Documentation: An imaging protocol manual will be provided, outlining the standardized procedures followed during image acquisition and processing. Additionally, DICOM header files containing technical metadata will be included with the shared images.
4. **Proteomics Data Metadata:**
   * Details: Metadata will describe sample preparation steps, mass spectrometry settings, data processing pipelines, and quality control measures.
   * Documentation: A protocol document will detail the mass spectrometry workflow, including specifics about reagents, instrument calibration, and data normalization procedures. The data will also include a cross-reference table linking protein identifiers to known databases such as UniProt.
5. **Behavioral Data Metadata:**
   * Details: Metadata will include survey design details, data collection methods (e.g., time points, participant instructions), and coding frameworks for qualitative data.
   * Documentation: The project will share the survey instruments, interview guides, and a detailed codebook outlining the categorization of responses. Additionally, a methodological appendix will explain the approaches used for behavioral data collection and analysis.

Overall, the project will provide extensive metadata and associated documentation to facilitate the interpretation and reuse of the scientific data. This includes not only technical details like data acquisition protocols and processing pipelines but also study-specific documentation such as data dictionaries, codebooks, and methodological appendices. All metadata and documentation will be made available through the same repositories as the primary data (e.g., dbGaP, TCIA, PRIDE), ensuring that future researchers have all necessary information to accurately interpret and build upon the project’s findings.

### Element 2: Related Tools, Software and/or Code

*State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.*   
  
The scientific data generated in this project will require specialized tools, software, and code for access, analysis, and manipulation. Below is a list of the essential tools, along with details on how they can be accessed:

1. **Genomic Data Analysis Tools:**
   * Tools and Software:
     + - **BWA (Burrows-Wheeler Aligner)**: Used for aligning sequence reads to the reference genome.
       - **GATK (Genome Analysis Toolkit)**: Employed for variant discovery and genotyping.
       - **Samtools**: Utilized for manipulating and analyzing BAM files.
   * Access: These tools are open-source and freely available. BWA and GATK can be downloaded from their respective GitHub repositories, while Samtools is available through the HTSlib project. Detailed installation guides and user manuals are available on the respective websites.
2. **Clinical Data Management Software:**

* Tools and Software:
  + **REDCap**: Used for data collection and management, particularly for handling de-identified clinical data.
  + **R**: Applied for statistical analysis and data visualization.
* Access: REDCap is available through institutional licensing at many academic and research institutions. The R programming language is open-source and can be downloaded from the Comprehensive R Archive Network (CRAN) website. REDCap documentation and R packages used in this project will be shared in a GitHub repository linked to the project.

1. **Imaging Data Processing Software:**

* Tools and Software:
  + **FSL (FMRIB Software Library):** Used for analyzing and visualizing functional MRI data.
  + **3D Slicer:** Employed for processing and visualizing CT and MRI images.
* Access: FSL and 3D Slicer are both open-source tools available for download from their respective project websites. Custom scripts used in this project for specific image processing tasks will be shared in a public GitHub repository.

1. **Proteomics Data Analysis Software:**

* Tools and Software:
  + **MaxQuant:** Utilized for analyzing mass spectrometry-based proteomics data.
  + **Perseus:** Used for downstream statistical analysis and data visualization of proteomics data.
* Access: Both MaxQuant and Perseus are freely available software tools that can be downloaded from the MaxQuant project website. The specific configurations and scripts used for data analysis in this project will be documented and made accessible via GitHub.

1. **Behavioral Data Analysis Tools:**

* Tools and Software:
  + **NVivo:** Used for qualitative data analysis, particularly for coding and analyzing interview transcripts and survey responses.
* Access: NVivo is a commercially licensed software, available through institutional licenses or individual purchase. The NVivo project files, coding schemes, and analysis protocols will be shared in a publicly accessible repository, along with guidance on how to use them.

The project will rely on a variety of specialized tools, software, and code to process, analyze, and interpret the scientific data. Most of the required tools are open-source and freely accessible, ensuring that other researchers can replicate and build upon our work. Where commercial software is used, such as NVivo, all necessary project files and documentation will be shared to facilitate access. Detailed instructions and custom scripts will be made available through a public GitHub repository, which will be linked to the data repositories (e.g., dbGaP, TCIA, PRIDE) where the primary scientific data are stored.

### Element 3: Standards

*State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist.*To ensure interoperability and facilitate the integration of datasets and resources across the research community, the following common data standards will be applied to the scientific data and associated metadata generated by this project:

1. **Genomic Data Standards:**
   * Standard: Global Alliance for Genomics and Health (GA4GH) Data Use Ontology (DUO)
   * Application: All genomic data, including whole-genome sequencing (WGS) and RNA-seq data, will be annotated using the GA4GH DUO standard to specify data use conditions. Variant data will be formatted in compliance with the Variant Call Format (VCF) 4.3 specification, ensuring compatibility with widely used bioinformatics tools and databases. Metadata describing sequencing methods, sample processing, and data quality metrics will adhere to the Minimum Information About a Microarray Experiment (MIAME) guidelines.
2. **Clinical Data Standards:**
   * Standard: Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR)
   * Application: De-identified clinical data will be structured using HL7 FHIR standards to facilitate data sharing and integration with electronic health records (EHRs). Variables such as participant demographics, laboratory results, and treatment outcomes will be encoded using standardized vocabularies, including the Logical Observation Identifiers Names and Codes (LOINC) for lab results and the International Classification of Diseases (ICD-10) for diagnoses. This approach will enable interoperability with other clinical datasets and compliance with regulatory standards.
3. **Imaging Data Standards:**
   * Standard: Digital Imaging and Communications in Medicine (DICOM)
   * Application: All MRI and CT imaging data will be stored in DICOM format, which is the standard for handling, storing, and transmitting medical imaging information. This will include the use of standardized DICOM tags for participant information, imaging parameters, and study details. Imaging metadata, such as acquisition parameters and processing steps, will follow the Neuroimaging Data Model (NIDM) guidelines to ensure consistency and facilitate cross-study comparisons.
4. **Proteomics Data Standards:**
   * Standard: mzML (Mass Spectrometry Markup Language)
   * Application: Proteomics data will be captured and stored using the mzML format, a standard developed by the HUPO-PSI (Human Proteome Organization - Proteomics Standards Initiative) for mass spectrometry data. This ensures that the data can be readily shared and reanalyzed using a wide range of proteomics tools. The metadata associated with proteomics experiments, including sample preparation protocols and instrument settings, will adhere to the Minimum Information About a Proteomics Experiment (MIAPE) guidelines.
5. **Behavioral Data Standards:**
   * Standard: Data Documentation Initiative (DDI)
   * Application: Behavioral and survey data will be documented using the DDI standard, which provides a framework for describing data from social, behavioral, and economic sciences. This will include detailed metadata on survey instruments, question design, coding schemes, and data collection methods. The DDI standard will enable interoperability with other social science datasets and facilitate data discovery and reuse.

Overall, the application of these common data standards will enhance the interoperability, reusability, and long-term value of the scientific data generated by this project. By adhering to established standards such as GA4GH DUO, HL7 FHIR, DICOM, mzML, and DDI, the project ensures that data and metadata are consistent, well-documented, and compatible with other datasets and tools. This approach aligns with NIDDK’s commitment to promoting data sharing and integration across the research community. In cases where no specific consensus standards exist, best practices and guidelines from relevant scientific communities will be followed to maintain high data quality and usability.

### Element 4: Data Preservation, Access, and Associated Timelines

1. **Repository where scientific data and metadata will be archived:**

*Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived (see* [*Selecting a Data Repository*](https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository)*).*

**Repository Selection:** The scientific data and metadata generated by this project will be archived in the following repositories, selected based on their suitability for the types of data produced and their alignment with NIDDK and NIH data sharing policies. Please note NIH encourages researchers to select the repository that is most appropriate for their data type and discipline:

1. **Genomic Data Repository:**
   * Repository Name: Database of Genotypes and Phenotypes (dbGaP)
   * Rationale: dbGaP is a well-established repository for archiving and sharing genomic and phenotypic data. It is specifically designed for storing human genomic data, with strong mechanisms for controlled access to protect participant privacy. All genomic data generated from whole-genome sequencing (WGS) and RNA-seq analyses will be submitted to dbGaP, along with associated phenotypic and clinical metadata. This repository was chosen because it meets NIH requirements for data sharing, provides robust support for secure data access, and facilitates the linkage of genomic data with clinical and phenotypic information.
2. **Clinical Data Repository:**
   * Repository Name: NIH provides a list of generalist repositories that accept all data types. Additionally, NIH-supported Scientific Data Repositories may be suitable if they align with the focus of your data. NIDDK Central Repository may be an option if the project is eligible to submit data or biospecimens.
   * Rationale: The proposed repository provides a standardized platform for sharing de-identified clinical data, enabling interoperability across studies. Clinical data, including participant demographics, laboratory results, and treatment outcomes, will be deposited in this repository. The repository was selected due to its focus on standardizing clinical data elements, which enhances data discoverability and reuse across multiple research initiatives.
3. **Imaging Data Repository:**
   * Repository Name: The Cancer Imaging Archive (TCIA)
   * Rationale: TCIA is a leading repository for medical imaging data, widely used by the research community for storing and sharing imaging datasets such as MRI and CT scans. Imaging data generated in this project, along with associated DICOM metadata, will be archived in TCIA. This repository was chosen because of its strong support for DICOM standards, its extensive use in imaging research, and its capabilities for linking imaging data with other clinical and genomic datasets.
4. **Proteomics Data Repository:**
   * Repository Name: Proteomics Identifications Database (PRIDE)
   * Rationale: PRIDE is an established repository for archiving and sharing mass spectrometry-based proteomics data. All proteomics data from this project, including mzML files and associated metadata, will be submitted to PRIDE. This repository was selected due to its compliance with HUPO-PSI standards, its widespread use in the proteomics community, and its support for data sharing and reanalysis.
5. **Behavioral Data Repository:**
   * Repository Name: Inter-university Consortium for Political and Social Research (ICPSR)
   * Rationale: ICPSR is a well-respected repository for social and behavioral science data, offering comprehensive services for data archiving, preservation, and access. Behavioral and survey data collected during this project will be deposited in ICPSR. This repository was chosen for its focus on social and behavioral data, its support for DDI standards, and its commitment to making data widely accessible to the research community.

The project’s scientific data and metadata will be preserved and made accessible through carefully selected repositories that align with the nature of the data and NIDDK’s data sharing goals. Timely submission and public access to these data will ensure compliance with NIH data sharing policies and will support the broader research community in conducting secondary analyses, validating findings, and developing new research hypotheses.

1. **How scientific data will be findable and identifiable:**

*Describe how the scientific data will be findable and identifiable (i.e., via a persistent unique identifier or other standard indexing tools).*

To ensure that the scientific data generated by this project are findable and identifiable, the following strategies will be employed, aligning with the FAIR (Findable, Accessible, Interoperable, Reusable) principles:

1. **Persistent Unique Identifiers (PUIDs):**
   * DOIs and Accession Numbers:
     + All datasets deposited in selected repositories will be assigned persistent unique identifiers (PUIDs) such as Digital Object Identifiers (DOIs) or accession numbers. For example, genomic data submitted to dbGaP will receive accession numbers that uniquely identify the dataset, while proteomics data in PRIDE will be assigned DOIs.
     + Rationale: These PUIDs ensure that the datasets can be reliably cited and located by other researchers, enhancing the findability of the data over time. The use of PUIDs aligns with the FAIR principles by making the data easily findable through standardized indexing systems.
2. **Metadata Standards and Indexing:**
   * Comprehensive Metadata Records:
     + Each dataset will be accompanied by detailed metadata that adheres to the relevant standards (e.g., GA4GH DUO for genomic data, DICOM for imaging data). These metadata records will include information about the data's origin, methodology, and processing, ensuring that the data are fully described and easily searchable within the repositories.
   * Repository-Specific Indexing:
     + The repositories selected for data deposition (e.g., dbGaP, TCIA, PRIDE) provide robust indexing tools that allow datasets to be discovered based on key metadata attributes such as study design, data type, and research focus. For example, in dbGaP, data can be searched by genomic features, phenotypic characteristics, and study identifiers, making it easier for researchers to find relevant datasets.
     + Rationale: The use of standardized metadata and repository-specific indexing enhances the findability of data by allowing precise and targeted searches, ensuring that datasets can be located by researchers with specific interests.
3. **Integration with Public Data Catalogs:**
   * NIH Data Catalog and Other Public Indexes:
     + Data deposited in NIH repositories will be listed in the NIH Data Catalog, which aggregates metadata from multiple repositories and makes it searchable through a single interface. Additionally, the data will be registered with public data catalogs such as FAIRsharing.org and re3data.org, which index research data repositories and datasets globally.
     + Rationale: Listing the datasets in these public catalogs increases visibility and ensures that the data can be discovered by a wide audience, including those outside the immediate research community. This approach supports the "Findable" aspect of the FAIR principles by providing multiple avenues for researchers to locate the data.
4. **Use of Rich Keywords and Ontologies:**
   * Controlled Vocabularies and Ontologies:
     + Metadata will include rich keywords and tags derived from controlled vocabularies and ontologies, such as MeSH (Medical Subject Headings) for clinical data and the Gene Ontology (GO) for genomic data. This structured approach allows for the consistent categorization of data and enhances the ability to find datasets through semantic searches.
     + Rationale: Employing controlled vocabularies and ontologies ensures that the data are not only findable through exact keyword matches but also through related terms, enabling more comprehensive searches and facilitating the discovery of related datasets.

**Alignment with FAIR Principles (Focus on "Findable"):**

The strategies outlined above ensure that the scientific data generated in this project will be highly findable. By assigning PUIDs, utilizing standardized metadata, integrating with public data catalogs, and employing rich keywords and ontologies, we align with the FAIR principles, particularly the "Findable" aspect. These efforts will ensure that the data are easily discoverable by other researchers, thereby promoting data reuse and accelerating scientific discovery.

1. **When and how long the scientific data will be made available:** *Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.* **Availability Timeline:**
2. **Genomic, Clinical, and Proteomics Data:**
   * When Available: The genomic, clinical, and proteomics data will be made available to other researchers no later than the time of publication of the primary research findings or at the end of the project’s performance period, whichever comes first. Typically, these data will be submitted to the respective repositories (dbGaP, NIH-supported or Generalist repository, PRIDE) within six months after final quality control checks are completed.
   * Rationale: This timeline ensures that the data are accessible to the broader research community as soon as the results are published, promoting transparency and enabling other researchers to validate findings or conduct secondary analyses. The early release of data, even before publication, will be considered for key datasets that could significantly benefit the research community.
3. **Imaging Data:**
   * When Available: Imaging data, including MRI and CT scans, will be made available upon publication of the associated results or at the end of the performance period, whichever is sooner. The data will be deposited in the Cancer Imaging Archive (TCIA) once image processing and necessary annotations are complete.
   * Rationale: The timeline for imaging data should allow for the necessary processing and annotation to ensure high-quality and interpretable datasets. However, making the data available at the time of publication remains a priority to maximize their utility.
4. **Behavioral Data:**
   * When Available: Behavioral and survey data will be made available within six months after the completion of data analysis and no later than the publication of the main study findings or end of the performance period. These data will be deposited in the Inter-university Consortium for Political and Social Research (ICPSR).
   * Rationale: Timely availability of behavioral data supports the reproducibility of research findings and allows for the immediate application of insights in related studies.

**Duration of Data Availability:**

**Long-term Preservation and Access:**

* All scientific data and associated metadata will be preserved and made available for a minimum of 10 years after the end of the project’s performance period. This long-term preservation will be ensured by the chosen repositories (dbGaP, TCIA, PRIDE, ICPSR), each of which has robust data preservation policies in place.
* Rationale: The 10-year availability period ensures that the data remain accessible for an extended period, supporting long-term research initiatives and enabling data reuse well beyond the initial publication. In cases where the data continue to be in high demand or relevance, the availability period may be extended further in collaboration with the repository.

### Element 5: Access, Distribution, or Reuse Considerations

1. **Factors affecting subsequent access, distribution, or reuse of scientific data:**

*NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See* [*Frequently Asked Questions*](https://sharing.nih.gov/faqs#/data-management-and-sharing-policy.htm) *for examples of justifiable reasons for limiting sharing of data.*In accordance with NIH expectations to maximize the appropriate sharing of scientific data, the following factors may affect the subsequent access, distribution, or reuse of the data generated by this project. These considerations will help to protect participant privacy, ensure compliance with ethical standards, and address other relevant factors:

1. **Informed Consent and Data Use Limitations:**
   * Considerations: The project involves the collection of genomic, clinical, and behavioral data from human participants. Informed consent obtained from participants includes specific provisions regarding the types of data that can be shared, the scope of data use, and any restrictions on data distribution. For instance, some participants may have consented to the use of their data solely for research related to specific diseases or within certain types of institutions.
   * Impact on Data Sharing: To respect these consent provisions, data sharing will be restricted in cases where participants have not consented to broad data sharing. For example, genomic data linked to identifiable health information will be shared through controlled-access repositories like dbGaP, which require data users to apply for access and agree to use the data only for approved purposes.
   * Rationale: These limitations are necessary to uphold the ethical obligations to participants and to comply with the terms of informed consent, ensuring that the sharing of scientific data does not violate participant rights or expectations.
2. **Privacy and Confidentiality Protections:**
   * Considerations: The protection of participant privacy and the confidentiality of sensitive health information are paramount. This project involves data types that, even when de-identified, may carry a risk of re-identification, particularly when combined with other datasets.
   * Impact on Data Sharing: To mitigate these risks, all shared data will be de-identified in accordance with the HIPAA Safe Harbor method or through expert determination. Additionally, certain data elements, such as precise geographic information or detailed genetic variants, may be further restricted or withheld from public datasets to reduce re-identification risk. Data shared through repositories like TCIA or ICPSR will be accompanied by agreements that outline the responsibilities of data users to maintain confidentiality.
   * Rationale: These measures are essential to safeguard participant privacy and to ensure that data sharing practices align with ethical and legal requirements, such as HIPAA.
3. **Intellectual Property and Proprietary Considerations:**
   * Considerations: Although not a primary factor, there may be instances where the data generated in this project are subject to intellectual property (IP) rights. This could apply to certain aspects of the proteomics data or novel methodologies developed during the research.
   * Impact on Data Sharing: In cases where data are subject to IP considerations, data sharing may be delayed until patent applications are filed or other IP concerns are addressed. However, once these issues are resolved, the data will be shared in accordance with NIH guidelines and timeframes.
   * Rationale: Balancing the need for data sharing with IP rights ensures that the data can be disseminated broadly without compromising the rights of the researchers and their institutions.
4. **Sensitive Data Considerations:**
   * Considerations: Some of the data collected in this project may be considered sensitive due to the nature of the health conditions studied (e.g., mental health, substance use disorders) or because they involve vulnerable populations (e.g., minors, individuals with disabilities).
   * Impact on Data Sharing: Data deemed sensitive may be subject to additional access controls, including the requirement for data users to provide justification for their use or to agree to specific conditions regarding data handling and security. In some cases, sensitive data may be available only to qualified researchers through controlled-access repositories, rather than through open-access platforms.
   * Rationale: Controlling access to sensitive data is necessary to protect the interests of participants and to comply with ethical guidelines for research involving vulnerable populations.
5. **Whether access to scientific data will be controlled:**

*State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).*

* **Genomic Data (dbGaP):**
  + Controlled Access: Yes. Data will be available through dbGaP, requiring approval via a Data Access Committee to ensure alignment with informed consent and privacy protections.
* **Clinical Data (named Generalist Repository):**
  + Controlled Access: Yes. Sensitive clinical data will be shared through named Repository, with access controlled to protect confidentiality.
* **Imaging Data (TCIA):**
  + Controlled Access: No. De-identified imaging data will be openly accessible in TCIA, as it poses minimal privacy risks.
* **Proteomics Data (PRIDE):**
  + Controlled Access: No. Anonymized proteomics data will be openly accessible in PRIDE, supporting transparency and reuse.
* **Behavioral Data (ICPSR):**
  + Controlled Access: Yes. Access to sensitive behavioral data will be controlled via ICPSR to protect participant privacy.

Controlled access will be applied to sensitive genomic, clinical, and behavioral data to protect privacy and comply with ethical standards. Open access will be provided for de-identified imaging and proteomics data to maximize accessibility and reuse.

1. **Protections for privacy, rights, and confidentiality of human research participants:**

*If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).*These protective measures will be implemented to ensure the privacy, rights, and confidentiality of human research participants are rigorously protected throughout the data sharing process.

1. **De-identification:** All human data will be de-identified according to HIPAA Safe Harbor standards or through expert determination to remove personally identifiable information (PII).
2. **Certificates of Confidentiality:** A Certificate of Confidentiality will be obtained to protect sensitive data from forced disclosure in legal proceedings, further safeguarding participant privacy.
3. **Controlled Access:** Sensitive datasets, including genomic, clinical, and behavioral data, will be shared through controlled-access repositories (e.g., dbGaP, named clinical repository) to ensure that only qualified researchers with approved projects can access the data.
4. **Informed Consent:** Informed consent forms will explicitly address data sharing, including any limitations or conditions, ensuring that participants are fully aware of how their data will be used and shared.

### Element 6: Oversight of Data Management and Sharing

*Describe how compliance with this Plan will be monitored and managed, the frequency of oversight, and by whom at your institution (e.g., titles, roles).*

1. **Institutional Oversight:**
   * Compliance with the Data Management and Sharing Plan will be overseen by the *Data Governance Committee and the Institutional Review Board (IRB)* at the institution. The institution will ensure that appropriate oversight mechanisms are in place to monitor data management practices.
   * The institution will take appropriate oversight actions to support and enforce compliance with the Data Management and Sharing Plan, providing resources and guidance as needed.
2. **Roles and Responsibilities:**
   * The *Principal Investigator (PI)* will be primarily responsible for ensuring that data management and sharing practices adhere to the plan, with support from the *Data Manager and the Research Compliance Officer*.
   * The *Data Manager* will monitor data handling, storage, and sharing practices monthly, ensuring that data is managed according to the outlined procedures.
   * The *Research Compliance Officer* will conduct quarterly reviews to ensure adherence to ethical guidelines, regulatory requirements, and the plan's stipulations.
3. **Frequency of Oversight:**
   * Monitoring will occur monthly by the Data Manager, with quarterly compliance reviews by the Research Compliance Officer, supported by the institution's oversight mechanisms.

The Data Governance Committee, IRB, PI, Data Manager, Research Compliance Officer, and other designated parties at the institution will collaboratively ensure regular and thorough oversight of data management and sharing. Monitoring will be conducted monthly, with quarterly compliance reviews, and the institution will provide appropriate oversight to support adherence to the plan.