Duke Superfund Data Chain of Custody Guidance Document

Introduction

This document provides guidance for maintaining data chain of custody within the Duke Superfund project teams. Its aim is to establish a uniform framework for the handling, processing, storing, and sharing of research data across all labs and projects. A clear and well-maintained chain of custody for data management is crucial for several key reasons. It supports the integrity and transparency of research outcomes, facilitating verification of findings by the broader scientific community. It facilitates regulatory compliance with funders and sponsors, particularly within clinical trial environments. Moreover, it strengthens operational ability to manage laboratory and center data assets in a secure and efficient manner. This document outlines the best practices and procedures across the data life cycle as well as additional considerations for storage, security, and compliance.

Objectives

- To standardize data management practices.
- To ensure data reproducibility and reliability.
- To enhance data security and access control.
- To facilitate compliance with industry best practices and regulatory

Data Collection

Good data chain of custody practices start at data collection to ensure that all data collected in a research environment is accurately recorded, traceable, and handled according to established standards.

Standard Operating Procedures (SOPs)

Procedures resulting in data generation should be written through data collection activities. Data collection procedures should be designed to ensure consistency and reliability in the data collection process through each method of generation utilized across labs and projects. Assign clear responsibilities for data collection to specific individuals or teams.

Documentation

Each data collection instance must be accompanied by documentation through traditional lab notebooks, electronic lab notebooks (ELN), or data management systems. This documentation should include key information such as:

- Date and time of data collection.
- Environmental conditions (if relevant).

- Equipment and tools used.
- Identity of personnel involved in the collection process.
- Specific methodology or protocol followed.
- Any deviations from standard procedures, along with justifications.

Data Labeling and Tracking

All collected data should be appropriately labeled with unique identifiers to facilitate tracking through subsequent stages. This labeling can be done through file naming conventions or within embedded metadata but should include a unique experiment or process ID or the date of collection, data generator, and the type of data. This ensures the original data files can describe or be linked to the information that describes how, when, and by who the data was generated.

Data QC, Processing, and Analysis

The key to good data chain of custody is how data is tracked through quality control (qc), processing, and analysis, where the data will undergo iterative transformation to a final product. It is important at this stage to set up processes and controls that ensure traceability and a transparent decision-making process.

Version Control

Implementing version control will allow for the tracking of data from its original form through cleaning, down selection, transformation, and further down stream data generating steps that produce analyzed or reportable data. There are several approaches to version control depending on the needs of the lab(s).

Manual Versioning with Local Storage

How it Works: Manually saving different versions of a dataset with clear naming conventions (e.g., dataset_v1, dataset_v2) to network stores with structured folders. Separate documentation or code should be maintained for describing transformations and stored with the data. Locking and archival of data should be performed to ensure data can not be retroactively modified after study close out.

Benefits: Low to no cost. No additional system management.

Trade-Offs: This method is simple and requires no special software, but it relies heavily on user discipline and is prone to human error. It's not efficient for large datasets or projects with multiple collaborators. Doesn't automatically log changes and rollback options are limited.

Requirements: Implementation of procedures for storing data according to pre-determined naming conventions and structures, as well as verify adherence. Labs should work with I.T. personnel to ensure back up and retrieval, as well as set access permissions for locked data.

Manual Versioning with Cloud Services (AWS, Google Drive, GitHub, etc.):

How it Works: These services require the same manual practices described previously but often include basic version control, allowing users to revert to previous versions of a file and see who made which changes.

Benefits: Low to no cost. No additional system management. Additional features for version and access control.

Trade-Offs: These platforms are user-friendly and provide better tracking than manual methods. However, they might not be suitable for very large datasets, and their version control capabilities can be limited compared to dedicated systems. Some cloud solutions such as GitHub require experience or training.

Requirements: In addition to manual practices described previously, labs should ensure cloud storage services are supported by institutional policies and do not contain any PHI or other restricted or confidential information.

Certified Off the Shelf Solutions (Electronic Lab Notebook or Data Management System)

How it Works: These are tools specifically designed for version control, widely used in research and healthcare environments. They can handle complex projects with many contributors and files and often provide rich features for data chain of custody. Some institutions provide these tools to research laboratories for free or discounted rates. Duke currently supports LabArchives.

Benefits: Rich data chain of custody features. A standard tool or platform for data management and recording keeping. Reduces human error and provides additional security controls. Can provide a compliant environment out of the box.

Trade-Offs: These platforms can vary in cost and will require direct or indirect management through a support role via laboratory staff or I.T. personnel. Depending on the research environment, they might not be suitable for all types of data. Additional training and onboarding are required.

Requirements: Policies for system usage and system outage should be implemented. For regulated environments, system owners will need to take additional measures to ensure systems compliance and validation will need to be confirmed with the vendor and I.T. personnel.

Choosing the Right Option

The choice of version control method depends on several factors: the size and complexity of the data, the technical expertise of the team, the number of collaborators, and the specific needs of the research project. Smaller projects might do well with low-tech solutions, while large-scale projects with multiple collaborators might benefit from more sophisticated systems. It's crucial to balance ease of use, scalability, and the level of control required.

Process Documentation

Process documentation for the qc, processing, and analysis state involves capturing methods, decision points, and rationale for actions on a dataset. This is critical to ensuring reproducibility and transparency in addition to providing utility for laboratory operational asset management. This can be achieved through developing procedures for manual data processing, the capture of code utilized for programmatic transformation, or saving the files or logs associated with data analysis tools.

Investigators, analysts, statisticians and other decision makers in the data qc and analysis pipeline should be able to provide documentation on key decisions made by responsible persons and provide resources or rationale as to why. Often this can boil down to domain specific best practices and expertise of subject matter experts. In addition, models, formulas, and calculations performed should be able to be verified in the case where proprietary or open-source tools are utilized.

It is important to note that documenting and verifying all methods and decision making throughout a process can be extremely labor intensive, and so description must be utilized to determine where reactionary approaches to inquiry or pre-emptive approaches to documenting and validating methodologies are needed.

They key aspects of version control include:

- Tracking of Changes: Every modification, whether it's an addition, deletion, or alteration, should be logged with timestamps and details of the person making the change. This creates a chronological record of how the data evolves over time.
- Revertability: Version control allows for the restoration of previous data states. This is crucial in cases where changes need to be undone or when tracking the history of the data's evolution.
- Collaboration and Consistency: In environments where multiple researchers or teams are working on the same data set, version control ensures that everyone has access to the latest version of the data, thereby maintaining consistency across the project.
- Documentation of Rationale: Alongside each change, a brief rationale should be documented. This helps in understanding the reasons behind specific modifications or analytical choices, contributing to the transparency of the research process.

Reporting

When data are reported externally, it is important to record information associated with data transfer and acceptance.

1. **Maintaining Records**: Keep comprehensive records of all data reported externally, including dates, versions of datasets used, and details of the reporting medium (e.g., journal articles, conference presentations).

2. **Documenting Data Requests**: Document any requests for data from external parties, including how the data was prepared and shared. Many public data repositories have systems for managing this automatically.

3. **Manifest Generation**: Creating a manifest when sending or reporting data that is comprised of a statistical summary on the counts and levels of data ensures agreement between the sender and recipient on what was sent and what was received.

Data Security

Safeguarding data is an essential component to maintain chain of custody as it ensures the integrity of the data through security against unauthorized access, breaches, and data loss.

Access Control

User Authentication: Implement strong user authentication methods for accessing data. This can include passwords, digital tokens, biometric verification, or multi-factor authentication. In most cases, institutional policies and frameworks provide the tools to comply with authentication requirements, however consideration should be given in the case of shared equipment or open systems. If multiple individuals can access data without unique identification, then controls and protocols should be implemented to prohibit data handling or otherwise implement measures to identify and log activity.

Roles: The definition of roles provides external visibility as to where in the data pipeline various actors should be handling the data. Compartmentalization of data access through roles provides additional security measures that reduce unauthorized access. Define and enforce role-based access controls (RBAC) to ensure that team members can only access data relevant to their role in the project.

Audit Trails: Maintain audit trails that record who accessed or modified data where possible. Modern scientific equipment and systems can include features for capturing user activity. This helps in tracking unauthorized or inappropriate access and modifications. These logs should be restricted and be backed up to ensure they cannot be manipulated.

Other Security Practices

- When data is in transit or at rest using local storage systems such as hard drives or local file servers, data encryption should be utilized to prevent access in the case of loss, theft, or accidental access.
- Utilize appropriate tools for data transfer to ensure adequate security of vulnerable data. It is not good practice to send sensitive data through email and is usually forbidden in regulated environments.
- Ensure data is backed and archived. Data not stored on redundant systems such as institutional network shares or cloud services are vulnerable to data loss.
- Adhere to institutional policies on data security and ensure staff complete training requirements in a timely manner.
- When depositing data to a repository, select a repository with appropriate access restrictions and controls according to the sensitivity of your data.

Compliance & Policy

Data chain of custody frameworks should be designed around legal requirements as well as sponsor, regulatory, and institutional policies for the handling of data.

Regulatory Compliance

1. Understanding Legal Requirements: Stay informed about and comply with all applicable laws and regulations related to data management in research. This includes data protection laws (e.g., GDPR in the EU, HIPAA in the US), intellectual property laws, and any sector-specific regulations.
2. Documentation for Compliance: Maintain comprehensive documentation of data management practices as evidence of compliance. This should include data management and sharing plans, SOPs, data processing logs, consent forms (where applicable), and data security measures.
3. Data Sharing and Publication: Ensure that data sharing and publication adhere to legal and ethical standards, including respecting confidentiality agreements and the rights of data subjects. More information can be found Superfund Best Practices Library.

Ethical Standards

1. Ethical Review Boards: Where applicable, submit research protocols to ethical review boards or institutional review committees for approval.
2. Informed Consent: In research involving human subjects, ensure that informed consent is obtained and documented, explaining how data will be used and stored.
3. Respect for Privacy: Uphold high standards of privacy, especially when handling sensitive or personal data. Implement and maintain robust de-identification procedures where necessary.

Regular Audits and Reviews

1. Internal Audits: Conduct regular internal audits to assess compliance with legal requirements and internal data management policies.
2. External Audits: Engage external auditors as necessary to provide an independent assessment of compliance and data management practices.
3. Continuous Improvement: Use findings from audits and reviews to continuously improve data management practices. Address any identified gaps or weaknesses promptly.

Training and Awareness

1. Regular Training: Provide ongoing training to all team members on legal requirements, ethical standards, and data management best practices.
2. Updates on Laws and Regulations: Keep the team updated on any changes in laws and regulations that may affect data management practices.

Incident Reporting and Management

1. Incident Reporting Procedures: Establish clear procedures for reporting and managing incidents that may involve non-compliance or ethical breaches.
2. Corrective Actions: In the case of identified non-compliance or ethical issues, take prompt corrective actions and document these actions.

Periodic Policy Review

1. Review Cycle: Establish a regular review cycle for this guidance document and all related data management policies.
2. Stakeholder Feedback: Involve key stakeholders in the review process to gather diverse perspectives and insights.
3. Adaptation to Change: Ensure that policies remain relevant and effective in light of new technologies, changing legal landscapes, and evolving research practices.