

Standard Operating Procedure (SOP):

Dataset Generation Using Independent Dual Programming and SAS Comparison

1. Purpose

This SOP defines the standardized process for generating analytical datasets using **independent dual development**, followed by **Quality Control (QC)** and **Quality Process (QP)** verification using SAS file comparison procedures (e.g., PROC COMPARE). The objective is to ensure **accuracy, reproducibility, traceability, and regulatory compliance** of all datasets.

2. Scope

This SOP applies to:

- All Osteoarthritis Initiative (OAI)-related research datasets used for learning labs, data exploration, or submission to external or internal clients
- All personnel involved in dataset development, validation, or approval
- All datasets programmed using SAS or SAS-compatible environments

This SOP does **not** cover downstream statistical analyses or reporting outputs.

3. Definitions

- **Primary Programmer (PP):** First individual independently generating the dataset
- **Secondary Programmer (SP):** Second individual independently generating the dataset
- **QC:** Quality Control — verification of correctness and concordance
- **QP:** Quality Process — adherence to documented procedures and standards
- **Gold Standard Dataset:** The finalized, approved dataset released for analysis
- **Specification Document (Spec):** Document defining variables, derivations, formats, and rules
- **Independent Programming:** Dataset development performed without shared code, macros, or intermediate outputs between programmers.

4. Roles and Responsibilities

Role	Responsibility
Primary Programmer	Independently develops dataset per approved specifications
Secondary Programmer	Independently develops dataset using the same specifications
QC Reviewer	Reviews comparison outputs and investigates discrepancies
Lead Investigator	Approves final dataset and ensures SOP compliance

5. Required Documentation

- Client approved dataset specifications
 - Programming logs (PP and SP)
 - SAS comparison outputs
 - Discrepancy resolution log
 - Dataset approval documentation
 - Version control records
-

6. Dataset Development Procedure

6.1 Specification Finalization

- Dataset specifications must be:
 - Version-controlled
 - Reviewed and approved by client prior to programming
 - Locked before dataset development begins

No programming may begin without an approved specification from the client.

6.2 Independent Dataset Programming

- PP and SP must:
 - Program independently
 - Use separate code bases
 - Avoid sharing code or logic during development
- Both datasets must:

- Conform exactly to specifications
 - Use consistent naming conventions
 - Apply identical input data sources
-

6.3 Dataset Output Standards

Each programmer must produce:

- Final dataset (SAS dataset or txt file (tab-, pipe-, or comma-delimited))
 - Annotated program code
 - Execution log confirming successful run
-

7. Quality Control (QC) Procedures

7.1 SAS File Comparison

The QC reviewer (PP or SP) performs a comparison using SAS (e.g., PROC COMPARE) with:

- Variable-level comparison
- Observation-level comparison
- Metadata (labels, formats, lengths)
- Any use of comparison tolerances (e.g., numeric precision thresholds) must be pre-specified and justified.

Comparison outputs must be saved and archived.

7.2 Discrepancy Handling

- All discrepancies must be:
 - Documented in a discrepancy log
 - Classified (e.g., logic error, missing variable, formatting issue)
- Resolution steps:
 1. Identify root cause
 2. Correct code (PP or SP, as appropriate)
 3. Re-run comparison
 4. Confirm full alignment

No dataset may proceed with unresolved discrepancies.

8. Quality Process (QP) Procedures

8.1 Process Adherence Review

The Lead Investigator confirms:

- Independent programming was maintained
- Approved specifications were followed
- All QC steps were executed and documented
- Version control was properly applied
- Any identified process deviations must be documented per Section 10.

8.2 Traceability and Audit Readiness

The following must be traceable:

- Input data versions
- Specification version
- Programmer identities
- QC reviewer identity
- Final approval date

All materials must be stored in the designated repository.

9. Final Dataset Approval

- Upon successful QC and QP review:
 - One dataset is designated as the **Gold Standard Dataset**
 - Final approval is documented and signed
- The dataset is then:
 - Locked
 - Released to client
 - Protected from modification

10. Deviations

Any deviation from this SOP must:

- Be documented with justification
 - Be approved by the Lead Investigator
 - Include an assessment of impact on data integrity
 - Deviations and their resolutions must be retained with the dataset archive.
-

11. Training Requirements

All personnel involved must:

- Be trained on this SOP prior to participation
 - Complete refresher training annually
 - Training completion must be documented and retained.
-

12. Revision History

Version	Date	Description	Approved By
1.0	2025-12-31	Initial release	Jeffrey Driban