**OBI Study Quality Checklist - Template**

Version: 1.2 - Date: July 17, 2024

|  |
| --- |
| ***Document overview:***  This checklist is a reorganized version of the DAQCORD guidelines[[1]](#footnote-1). DAQCORD is endorsed by the International Neuroinformatics Coordinating Facility (INCF). OBI has organized this checklist by stage of study data collection. All OBI Integrated Discovery Programs (IDPs) ***must*** complete these checklists prior to each stage of their study and ***submit the updated checklist to OBI***. If the IDP makes use of Brain-CODE and the OBI neuroinformatics services, please consult with OBI if you require additional help with ensuring these quality steps are met.  The checklist is broken down by the two stages of the Data Quality Framework:   * Stage 1: Study Design, Training, and Testing * Stage 2: Study Run Time & Post-Collection   Please note that each item in this checklist refers to a DAQCORD item # (<https://www.daqcord.org/daqcord-questions>)  Once Stage 1 of the checklist is completed, please submit this to OBI.  Once Stage 2 of the checklist is completed, please submit the updated checklist to OBI. |

# **Stage 1: Study Design, Training, and Testing (33 items)**

Study Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Stage 1 Completion Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### **Correctness (Accuracy):**

* eCRFs are designed by a team with a range of expertise (item #1).
* Variables are named and encoded in a way that is easy to understand (item #5) and the same name should be used across the CRF and database design so that confusion is minimized.
* Free text is avoided unless clear scientific justification and analysis plan specified/feasible (item #12).
* Data collection fields include information on inclusion/exclusion criteria (item #16).
  + *Avoid enrollment errors by requiring completion of an eCRF with checklist confirming that eligibility criteria have been met for enrolled subjects.*
  + *For subjects that need to be enrolled within a specific window of time there are timestamp checks in the eCRF in place to verify the eligibility window.*
  + *Date of birth falls within the enrollment age criteria.*
* All personnel responsible for entering data receive training and testing on how to complete the eCRF (item #21).
* The eCRF is easy to use and includes a detailed description of the data collection guidelines and how to complete each field in the form. They are pilot-tested in a rigorous pre-specified and documented process until reliability and validity are demonstrated (item #22).
* Data collection that requires specific content expertise is carried out by trained/certified investigators (item #24).
* Assessors are blinded to treatment allocation or predictor variables where appropriate and blinding is explicitly recorded (item #25).
* There is a clear audit chain for any data processing that takes place after entry, and this should have a mechanism for version control (item #26).
* Imaging acquisition techniques are standardized (item #29).
* Biospecimen preparation techniques are standardized (item #30).
* Biospecimen assay accuracy, precision, repeatability, detection limits, quantitation limits, linearity and range are defined. Normal ranges are determined for each assay (item #31).
* There is automated entry of the results of biospecimen samples (item #32).

### **Completeness:**

* There is a robust process for choosing and designing the data set to be collected that involves appropriate stakeholders, including a data curation team with appropriate skill mix (item #2).
* Data that are mandatory for the study are enforced by rules at data entry, and user reasons for overriding the error checks (queries) are documented in the database (item #9).
* Missingness is defined and is distinguished from "not available", "not applicable", "not collected" or "unknown". For optional data, "not entered" is differentiated from "not clinically available" depending on research context (item #10).
* There is clear documentation for interdependence of eCRF fields, including data entry skip logic (item #15).
* A team of data curation experts are involved with pre-specified initial and ongoing testing for quality assurance (item #33).

### **Concordance (Reliability):**

* The data ontology is consistent with published standards (common data elements) to the greatest extent possible (item #3).
* Data types are specified for each variable (item #4).
* Database rule checks are in place to identify conflicts in data entries for related or dependent data collected in different eCRFs or sources (item #13).
* Data collection methods are documented in study manuals that are sufficiently detailed to ensure the same procedures are followed each time (item #20).
* Data collectors are tested and provided with feedback regarding the accuracy of their performance across all relevant study domains (item #23).
* For physiological data, the methods of measurement and units are defined for all sites (item #28).

### **Representation (Precision):**

* Relational databases have been appropriately normalized: steps have been taken to eliminate redundant data and remove potentially inconsistent or overly complex data dependencies (item #6).
* Each individual has a unique identifier (item #7).
* There is no duplication in the data set: data have not been entered twice for the same participant (item #8).
* There are mechanisms in place to enforce/ensure that time-sensitive data are entered within allotted time windows (item #14).
* The data entry tool does not perform rounding or truncation of entries that might result in precision loss (item #17).
* Internationalization is undertaken in a robust manner, and translation and cultural adoption of concepts (e.g., assessment tools) follow best practice (item #19).
* Data are provided in a form that is unambiguous to researchers (item #27).

### **Plausibility (Integrity):**

* Range and logic checks are in place for eCRF response fields that require free entry of numeric values. Permissible values and units of measurement are specified at data entry (item #11).
* Extract/transform/load software for batch upload of data from other sources such as assay results should flag impossible and implausible values (item # 18).

# **Stage 2: Study Run Time (10 items)**

Study Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Stage 2 Completion Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### **Correctness:**

* Source data validation procedures are in place to check for agreement between the original data and the information recorded in the database (item #39).
* Scoring of tests is checked. Scoring is performed automatically where possible (item #41).
* Data irregularities are reported back to data collectors in a systematic and timely process. There is a standard operating procedure for data irregularities to be reported back to the data collectors and for documentation of the resolution of the issue (item #42).

### **Completeness:**

* Proxy responses for factual questions (such as employment status) are allowed in order to maximize completeness (item #34).
* There is centralized monitoring of the completeness and consistency of information during data collection (item #36).

### **Representation:**

* Automated variable transformations are documented and tested before implementation and if modified (item #35).
* Known/emergent issues with the data dictionary are documented and reported in an accessible manner (item #43).

### **Plausibility:**

* Individual data elements should be checked for missingness. This should be done against pre-specified skip-logic/missingness masks. This should be performed throughout the study data acquisition period to give accurate "real time" feedback on completion status (item #37).
* Systematic and timely measures are in place to assure ongoing data accuracy (item #38).
* Reliability checks have been performed on variables that are critical to research hypotheses, to ensure that information from multiple sources is consistent (item #40).

# **Stage 2: Post-Collection (3 items)**

### **Correctness:**

* A plan for ongoing curation and version control is specified (item #45).

### **Representation:**

* The version lock-down of the database for data entry is clearly specified (item #44).
* A comprehensive data dictionary is available for end users (item #46).

1. Ercole A, Brinck V, George P, Hicks R, Huijben J, Jarrett M, Vassar M, Wilson L; DAQCORD collaborators. Guidelines for Data Acquisition, Quality and Curation for Observational Research Designs (DAQCORD). J Clin Transl Sci. 2020 Mar 13;4(4):354-359. doi: 10.1017/cts.2020.24. PMID: 33244417; PMCID: PMC7681114. <https://www.daqcord.org/> , <https://www.incf.org/sbp/daqcord> [↑](#footnote-ref-1)