

Pre-Submission QC — Is It Submittable?

Cannabis and Health Data Repository

Pre-Submission QC Checklist

A mechanical “is it submittable?” pass to run on your files before you transfer them. This is not full de-identification or harmonization — that is the Data Coordination Center’s job. This checklist covers the one de-identification step that must happen at your institution, plus the data-hygiene items that let us validate and harmonize your data without a round of back-and-forth.

Identity and de-identification (must happen at your institution)

- ☐ **Direct identifiers removed.** Names, addresses, phone numbers, email, MRNs, SSNs, full dates tied to an individual, and other direct identifiers are not present in the files you are transferring.
- ☐ **Opaque study ID assigned.** Each participant carries an opaque study identifier in place of any direct identifier.
- ☐ **Crosswalk kept at your institution.** The table mapping your opaque study ID back to real participants stays with you and is **never** included in the transfer. The Repository never receives it.

File structure

- ☐ **One record per case** where possible. For repeated events (visits, waves), the case is the row and repeated measures are columns or clearly-keyed related files — and the relationship is described in the documentation.
- ☐ **Consistent ID variable.** The participant-ID variable uses the **same name** across every file, in **character (string) format**.
- ☐ **Linkage described.** If you submit more than one file, the documentation explains how the files relate (keys, one-to-many, timepoints).

Variables and coding

- ☐ **Everything is labeled.** Every variable and every coded value has a clear label in the data file itself, not only in the codebook.
- ☐ **Missing data is coded explicitly.** Not-applicable, non-response, and refusal are distinguished with documented codes rather than blanks.
- ☐ **Technical variables included.** Weights, design variables, and disposition indicators needed for valid inference are present, with a note on how to use them.

- ☐ **Derived variables carry their code.** Any constructed or derived variable is accompanied by the program code that produced it.

Documentation completeness

- ☐ **Codebook covers every file** and matches the variables actually present (names, labels, value labels, missing codes).
- ☐ **Study documentation** (design documents, questionnaires/instruments, and any reports or publications associated with the data) is included.

Format and readability

- ☐ **Files open and are readable** in a preferred or accepted format (see the “What to Submit” checklist).
- ☐ **No stray identifiers in free text.** Open-text fields have been reviewed; if narrative fields carry disclosure risk, flag them — do not silently delete them. They can be routed to the restricted tier. (*See the De-identification one-pager.*)

You do not need to harmonize your data yourself. Mapping your variables to the common data dictionary is our role. Focus on delivering a well-documented, de-identified dataset: remove direct identifiers, assign the opaque ID, apply Safe Harbor de-identification as your IRB requires (we can assist), and document thoroughly. If de-identification would harm the analytic value of specific fields, contact us before stripping them.

Cannabis and Health Data Repository · NIH/NIDA Award 1UM1DA059000 · Johns Hopkins University and Realm of Caring.