AI tools were applied in this document to improve clarity and grammar. The content, structure, and analytical framework were developed independently by the author.

Data cleaning processes and procedures

The purpose of this document is to establish a standardized, stepwise protocol for the preparation of analytical datasets, following the Data Lifecycle framework. It guides the transformation of raw data—from initial extraction through cleaning, restructuring, and final integration—into clearly documented, analysis-ready files. This document ensures consistency, transparency, and compliance with FIPPA, FAIR data principles, and project-specific analytic plans.

This document is intended for study investigators, data managers, analysts, statisticians, research coordinators, and database developers responsible for research data preparation. It provides practical guidance on creating well-organized, metadata-rich datasets that are aligned with the study protocol, ethics submissions, and long-term data governance and sharing requirements.

Related documents

Data principles and governance

Designing a Research Database: Structure, Documentation, and Governance

Data dictionary and terminology

Raw dataset is a data table extracted directly from the data collection platform (such as REDCap), containing all variables specified in the study protocol or grant proposal, prior to any cleaning or transformation. It includes every variable defined in the data dictionary during the database design stage.

Raw cleaned dataset is a version of the raw dataset in which all variables are retained but cleaned for basic data quality issues. Variables are transformed to numeric formats where applicable (except for "other, specify" text fields), and the dataset is reviewed for outliers, invalid values, and missing data.

Transformed analytical dataset is a modified version of the raw cleaned dataset in which the data structure has been changed (e.g., reshaped from wide to long format) and reorganized into analysis-ready tables such as patient characteristics, event-level data, or study endpoints. This stage may involve recoding variables, creating or combining outcomes, and converting continuous variables to

categorical formats. All changes are documented, and the data dictionary is updated to reflect variable descriptions, formats, and structural modifications.

Final analytical dataset is a structured dataset created and tailored for a specific analysis or sub-project. It may be derived from raw cleaned data, transformed analytical datasets, or a combination of both. Multiple final datasets can be produced for different research questions. Each dataset is fully documented, including variable descriptions, coding, and formats. When multiple datasets are used, clear documentation of their linkage (e.g., via unique IDs or merge logic) is required to ensure traceability and reproducibility.

Algorithm for creating analysis ready set of data sets

I typically follow a structured algorithm (Figure 1) for creating analytical files that aligns with all project-specific documentation developed to date. This includes the study protocol, statistical analysis plan, data dictionary, and supporting materials (see Statistical Resources – Women's Health Research Institute for guidance on required documents). The workflow includes a privacy protection step to ensure compliance with FIPPA, PIPEDA, and other relevant regulations, consistent with the procedures outlined in the ethics submission, research proposal, and study protocols.

It illustrates the standard workflow procedures used to prepare analysis-ready datasets, beginning with raw data extraction and proceeding through cleaning, transformation, and final dataset creation. At the center of the process is a secure linkage step using random IDs and a master-person file, allowing integration across multiple datasets, including external data sources. This structure ensures consistent de-identification and linkage across all data tables. The blue arrows reflect iterative steps between dataset stages, supporting updates, restructuring, and refinement prior to final analysis. The entire process is embedded within a privacy-preserving framework that aligns with ethics and data governance protocols.

To support privacy protection, I begin by generating a file with a large set of randomly ordered IDs. In parallel, I extract the actual IDs from the source database and sort them in a random order. The two lists are then merged by position to create a master-person linkage file that maps real IDs to randomly assigned project IDs. This file serves as a centralized linkage reference across all datasets in the project—regardless of scope or complexity. The lead data analyst or statistician maintains the master-person file exclusively. All other team members are provided access only to project datasets containing the

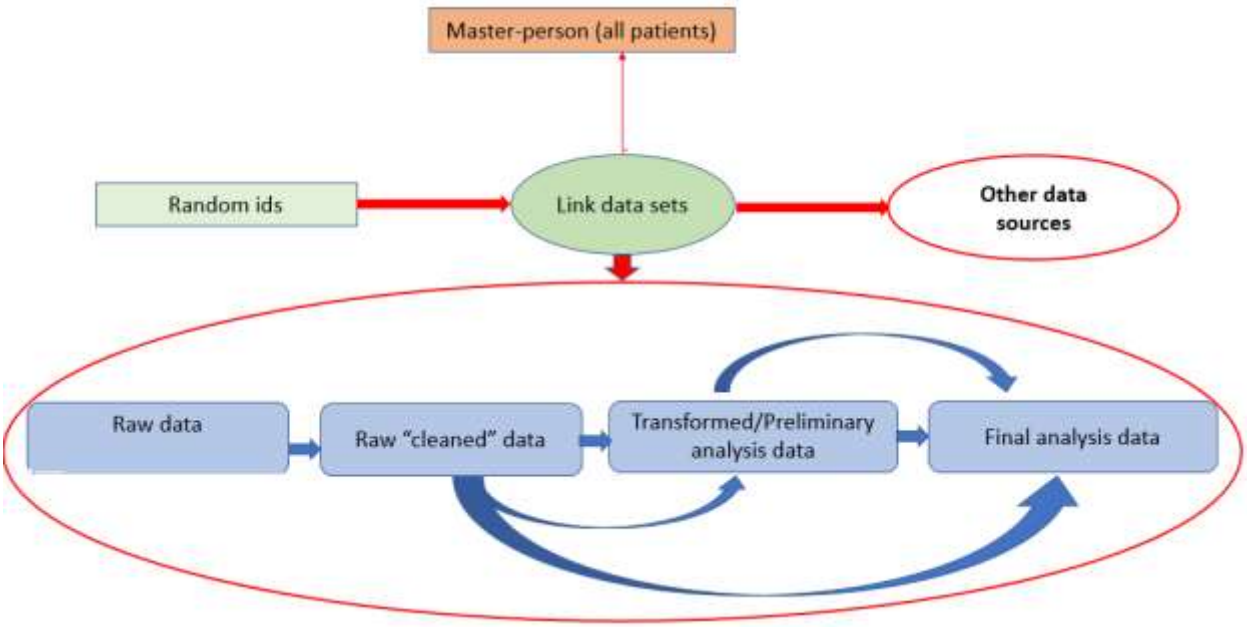
randomized IDs, with no access to the original identifiers. The below table shows the example on how dataset look like.

File example with random ids	
Random Id	Order
123345	1
123783	2
123944	3

File with original ids sorted randomly	
Actual id	Order
100	1
1	2
2022	3

Master-person file	
Actual id	Random Id
1	123345
100	123783
2022	123944

Figure 1:



Preparation of raw "cleaned" data set

- The raw dataset is extracted directly from your data collection system (such as REDCap) or from external sources (e.g., administrative data, clinical databases). Before it can be used for analysis, the raw dataset should be reviewed and modified to produce a raw cleaned dataset.
- Below are some practical tips I use when evaluating a raw dataset for the first time:
- Compare the variable list in the raw dataset against your original database documentation to ensure that all expected variables are present. This helps confirm that the extract matches what was planned in the protocol or data dictionary.

- If all variables are stored in one file (e.g., if you didn't follow my earlier suggestions about separating tables during database design, or if you received the data from an external source), make sure to document this step. Clearly record any decisions made about how the dataset was modified and specify the unique identifier(s) that will be used for future linkage.
- Use PROC CONTENTS in SAS (I usually sort by position order) or equivalent functions in other software—such as str() or glimpse() in R—to examine variable types and general structure in the raw dataset.
- Ensure all variables are coded as numeric, unless text is required (e.g., free-text responses like "other, specify"). This will simplify cleaning, analysis, and summary statistics.
- Standardize date/time formats across the dataset. Ensure that all date variables follow the same structure—ideally decided during the database design phase.
- Label and format all variables using your original data dictionary. This can be automated in SAS or R. (I'll be sharing examples of this process in a future WHRI GitHub repository—stay tuned!)
- Once the structure is validated, assess data quality. For survey data, begin by checking skip logic using the original instruments. Recode as needed to ensure consistency with documentation. Total counts across skip patterns should align—this will be important later when subsetting for analysis (e.g., restricting to a target population).
- Use PROC FREQ (for categorical variables) and PROC MEANS (for continuous variables) in SAS—or your preferred alternatives in other software—to summarize and review the data. Pay close attention to missing values, outliers, and inconsistent values. For clinical variables, confirm that values fall within expected ranges. Discuss any flagged issues with the research or clinical team and document all decisions.
- Recode values that need to be modified for analytical use, and make sure the rationale for recoding is clearly documented.
- Check for duplicate records. For datasets with one record per participant (e.g., baseline characteristics) or one record per test or follow-up period, make sure no unintentional duplicates exist. For example, lab datasets may contain multiple measures per day—if test time is missing, you may need to decide which record to retain based on study context.

- Verify that clinical data units are consistent. Descriptive summaries will often reveal mismatches. If multiple units exist, convert them to a standard unit and document the change. For example, in one of my projects involving CMV viral load, the assay changed mid-study and values were reported differently. We contacted the lab to understand how to harmonize values across assays and described the conversion clearly in the paper for transparency.
- Update the data dictionary based on the new "cleaned" raw data created.

Preparation of transformed analytical files

- Begin by restructuring your raw cleaned dataset as needed for analysis. This often means reshaping the data from wide to long format (or vice versa) depending on the analysis requirements.
- In my lecture about data (see my resource page) I stated to split data in this stage. However, after working with some different types of data sets here at WHRI I moved it to the previous data stage. You can decide when to do it based on the nature of the data.
- Create or re-create outcome variables based on the analytic plan. This might involve combining multiple variables, categorizing continuous values (e.g., symptom score → mild/moderate/severe), or defining time-to-event outcomes.
- Convert variable types when appropriate. For example, convert a continuous variable into a categorical one based on clinical thresholds. If you are creating new formats at this stage, please make sure to include new information in tour data dictionary.
- If you have longitudinal data that you are planning to share with interdisciplinary teams, my suggestion would be to lock the data and create master file with all the key and standardized variables. Keep a list of the variables in the master files (commonly used) and update it any time new key variable is created.
- Clearly document all transformations. Update your data dictionary to reflect any changes in variable names, types, coding schemes, or structure. Include notes on derived variables and how they were calculated.
- Keep intermediate versions of each dataset if you're doing multiple transformation steps. This is especially helpful when working in teams or when you need to backtrack.

Preparation of final analytical files

There is not much advice here, as the preparation of final analytical files would depend on the specific analysis and will be tailored to it. But here are some valuable tips

- Create tailored datasets specifically for the analysis or sub-project you're working on. These may be derived directly from the raw cleaned data, the transformed dataset, or a combination of both—depending on what's needed for your model or hypothesis.
- You can prepare multiple final datasets for different types of analyses (e.g., cross-sectional vs. longitudinal, baseline-only vs. full follow-up). Each one should be streamlined to include only the variables relevant to that analysis.
- Ensure that all datasets are fully documented, including updated variable labels, formats, coding schemes, and derived fields. The data dictionary should reflect the final structure used in modeling or reporting.
- If your project uses multiple final datasets, make sure to clearly document how they are linked—using consistent IDs, merge keys, or time indicators. This is especially important for hierarchical models, time-to-event analyses, or when linking administrative data with survey responses.
- Before sharing or archiving, verify that the dataset contains no personal identifiers (unless explicitly approved) and that all formatting is consistent and ready for export, publication, or replication.
- When merging two or more data sets, make sure to check the unmerged records. For example, if you have two files A and B, you need to ensure that all the intendent records are included. For example, you expect to have one to one merge for the A and B files. If you have records that are in A and not in B (or opposite) you need to check every record that was unmatched. There can be a good explanation, why your merge did not work - > then it would be a good idea to document any records and explanation for mismatch. However, in most of the cases, in my experience, the mismatch is due to the errors. For example, you merge your data by postal code, but postal codes have different structure in two data sets (V7A3E7 V7A 3E7).

To conclude

This guide provides a structured and practical approach to preparing high-quality, analysis-ready datasets, grounded in the principles of data governance, documentation, and ethical research practice.

From initial extraction of raw data to the creation of final analytical files, each step in the process is designed to ensure transparency, consistency, and alignment with the study protocol, statistical analysis plan, and privacy regulations such as FIPPA and PIPEDA.

By following this staged framework—moving from raw data, to raw cleaned, to transformed, and finally to tailored analytical files—research teams can build reproducible, well-documented datasets that support valid, efficient, and responsible analysis. Practical tips throughout the document highlight common challenges, offer examples from real-world projects, and emphasize the importance of ongoing documentation through the data dictionary and team communications.

Ultimately, strong data preparation process is not just about technical accuracy—it is about trust, auditability, and long-term usability. Whether you're working with clinical trials, administrative linkages, or survey-based studies, investing time and care at this stage of the research process improves the quality of your findings and the integrity of your study.

As always, adapt these principles to the specific context of your project, and don't hesitate to document your rationale when deviations or decisions are made. And remember, clear structure, consistent naming, and thoughtful documentation will always save time later.

References and related materials:

Arts, D. G. T., De Keizer, N. F., & Scheffer, G. J. (2002). Defining and improving data quality in medical registries: A literature review, case study, and generic framework. Journal of the American Medical Informatics Association, 9(6), 600–611. https://doi.org/10.1197/jamia.M1087

DAMA International. (2017). The DAMA guide to the data management body of knowledge (DAMA-DMBOK®) (2nd ed.). Technics Publications. https://www.dama.org/cpages/body-of-knowledge

European Open Science Cloud. (n.d.). Data life cycle. RDMkit. https://rdmkit.elixir-europe.org/data_life_cycle

Government of British Columbia. (1996). Freedom of information and protection of privacy act. https://www.bclaws.gov.bc.ca/civix/document/id/complete/statreg/96165_00

Government of British Columbia. (2003). Personal information protection act. https://www.bclaws.gov.bc.ca/civix/document/id/complete/statreg/03063_01

Government of Canada. (2021). Tri-agency research data management policy. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, & Social Sciences and Humanities Research Council of Canada. https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management/tri-agency-research-data-management-policy

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. Epidemiology, 10(1), 37–48. https://doi.org/10.1097/00001648-199901000-00008

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. Journal of Biomedical Informatics, 95, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics, 42(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Kahn, M. G., Brown, J. S., Chun, A. T., Davidson, B. N., Meeker, D., Ryan, P. B., Schilling, L. M., Weiskopf, N. G., Williams, A. E., & Zozus, M. N. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMs (Generating Evidence & Methods to Improve Patient Outcomes), 4(1), 1244. https://doi.org/10.13063/2327-9214.1244

National Institutes of Health. (2023). NIH data management and sharing policy. https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/dms

Partridge, E. F., & Bardyn, T. P. (2018). Research electronic data capture (REDCap). Journal of the Medical Library Association, 106(1), 142–144. https://doi.org/10.5195/jmla.2018.319

Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226–1227. https://doi.org/10.1126/science.1213847 https://www.science.org/doi/10.1126/science.1213847

Provincial Health Services Authority. (n.d.). Data access & privacy. https://www.phsa.ca/researcher/data-access-privacy

Provincial Health Services Authority. (n.d.). Research privacy. https://www.phsa.ca/researcher/data-access-privacy/research-privacy

University of British Columbia. (n.d.). Data governance. UBC Office of the Chief Information Officer. https://cio.ubc.ca/data-governance

University of British Columbia. (n.d.). Research data management. UBC Library. https://researchdata.library.ubc.ca/

University of British Columbia. (n.d.). Data management plans. UBC Library Research Data Management. https://researchdata.library.ubc.ca/plan/

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. PLOS Medicine, 2(10), e267. https://doi.org/10.1371/journal.pmed.0020267
Walther, B., Hossin, S., Townend, J., Abernethy, N., Parker, D., & Jeffries, D. (2011). Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. PLOS ONE, 6(9), e25348. https://doi.org/10.1371/journal.pone.0025348

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18