How to structure your research project folders

Version number: 1.0
Date: 25May2022
Author: S Coffey

Overview
This document describes the shared folder structure that best works with an individual project, or group of related projects. These recommendations are based on personal experience, and should be adapted to individual projects.

Here we use the designation PROJ1 for the current project. The local path is written in full.

High level structure

/PROJ1
/PROJ1/PROJ1_Admin
/PROJ1/PROJ1_Articles
/PROJ1/PROJ1_Methods
/PROJ1/PROJ1_Dissemination
/PROJ1/PROJ1_Results

The Admin folder contains miscellaneous administrative documents relevant to the project that do not fit in to other areas. The Articles folder contains research articles directly relevant to the research project. Ideally a copy of all articles appearing in Dissemination documents would be kept here for ease of access.

PROJ1_Methods substructure
The methods folder contains all the information required for setting up and running a study.

/PROJ1/PROJ1_Methods/PROJ1_Amendments
/PROJ1/PROJ1_Methods/PROJ1_Charters
/PROJ1/PROJ1_Methods/PROJ1_DMC
/PROJ1/PROJ1_Methods/PROJ1_Ethics
/PROJ1/PROJ1_Methods/PROJ1_Funding
/PROJ1/PROJ1_Methods/PROJ1_Future_plans
/PROJ1/PROJ1_Methods/PROJ1_Locality
/PROJ1/PROJ1_Methods/PROJ1_Meetings
/PROJ1/PROJ1_Methods/PROJ1_Participant_info
/PROJ1/PROJ1_Methods/PROJ1_PISCF
/PROJ1/PROJ1_Methods/PROJ1_PROM
/PROJ1/PROJ1_Methods/PROJ1_Protocol
/PROJ1/PROJ1_Methods/PROJ1_Protocol_appendices
/PROJ1/PROJ1_Methods/PROJ1_SAP

/PROJ1/PROJ1_Methods/PROJ1_SOPs
/PROJ1/PROJ1_Methods/PROJ1_SIV
/PROJ1/PROJ1_Methods/PROJ1_Sponsor
/PROJ1/PROJ1_Methods/PROJ1_Trial_registration
/PROJ1/PROJ1_Methods/PROJ1_TSC

Notes:

Amendments contains a working document listing items required for inclusion in an amendment application, or, if minor, for notification at the annual report. Each amendment application or approval should have its own subfolder.
Charters for a data monitoring committee (DMC) and trial steering committee (TSC) may of course not be used if this is not a randomised trial.
The DMC and TSC folders contain the meeting notes, reports, and minutes of these meetings.
The Ethics folder should have a subfolder containing the current set of documents approved by the relevant Ethics committee (often collected from other subfolders such as PISCF). This can be superceded by future approvals if required.
The Future plans folder contains brief notes on potential future plans, substudies, follow-on studies or any research ideas that might pertain to the study, but are not part of the primary or prespecified analyses.
The Meetings folder contains any meeting minutes of local investigators or any non-TSC or DMC meetings that are minuted.
The PISCF and Protocol folders contains documents in progress for the patient information sheet and consent form and the protocol, respectively. The Protocol appendices folder can be used if there is a more complex appendix to be worked on, for example a Data Governance appendix.
The SAP folder contains the Statistical Analysis Plan.
The SOPs folder contains any Standard Operating Procedure documents.
The SIV folder contains details of any site initiation visits in the case of multicentre studies.
The Sponsor folder contains communication with Sponsors, insurance details etc.

PROJ1_Dissemination substructure
This should have a subfolder for every major research output (paper, presentation, thesis etc.) as they are worked on. Reports specifically for the DMC, TSC, Ethics committee, locality, or sponsor should be placed in the appropriate section of the methods. Each presentation folder should be detailed with the date of the presentation as in the filename recommendations below.

PROJ1_Results substructure
This folder contains all the study data, statistical analysis scripts, and outputs from this.

PROJ1_Results/PROJ1_Data_description
PROJ1_Results/PROJ1_Data_notes
PROJ1_Results/PROJ1_Data_received
PROJ1_Results/PROJ1_Output

PROJ1_Results/PROJ1_Scripts
PROJ1_Results/PROJ1_Working_data

Each subanalysis should have its own subfolder within the Output, Scripts, and Working_data folders.
The Data description folder contains any description of the data, such as variable lists or ideally a data dictionary. The Data notes folder contain any miscellaneous notes about decisions made during analysis that are not in the SAP.

The workflow for analysis is:

1. Raw results go directly from data capture (e.g. via Redcap) into Data_received. The data should not be altered manually.
2. Data cleaning scripts (held in Scripts) work on this data and save clean data for analysis in to Working_data. This data should be written both as a csv or Rdata file for most datasets.
3. Analysis scripts (also held in Scripts) work on cleaned data to produce tables, figures, and other results in Output. This can include Rmarkdown or Quarto documents generated in-line from the scripts.

Filename recommendations and version control
Filenames should be written as:
YYYYMMDD_PROJ1_nameoffile_vx.y.z.fileextension
YYYYMMDD is the year, month, date, and nameoffile contains some description of what the file is for, beyond its location in the file structure. Fileextension refers to the filesystem file extension, e.g. docx for Word documents.

In an ideal world, we would be using git for this, but in practice this will not be adopted widely in the near future in the medical sciences. Version numbers (outside of trial protocols) generally follow the convention that v1.0 is the first relatively clean public release. Thereafter version number x is moved up with major changes (e.g. after incorporating multiple changes from multiple authors), y is incremented for minor changes, and z is only used for very small changes, usually minor typographic errors (similar to patches for software). When getting comments on manuscripts a filename suffix "_AB" or "@AB" is used to denote comments from co-author AB. After incorporation of changes from one or more co-authors, the next version sent should have only changes from the last version left visible (i.e. older changes should have either been accepted or rejected).

Trial protocols have stricter version control, which can be seen in GCP guidance.