MRC Epidemiology Unit

Good Analytical Practice - Standard Operating Procedure

Version 2 , July 2018

1. Scope

This Standard Operating Procedure (SOP) applies to all members of the Unit* (including visitors working with Unit data and students) who perform any type of analysis on either Windows or Linux, where the results are then included in an output for which the Unit has primary responsibility. Examples of outputs include papers, reports, PhD theses, MPhil project reports, conference presentations.

It is the joint responsibility of each member of staff/student and their supervisor/line manager to ensure adherence to the requirements of this SOP.

In situations where analysis is performed for other collaborators (where the Unit is not primarily responsible for the output), this SOP should also be followed as closely as possible.

* In this document, "Unit" is used to refer to both the MRC Epidemiology Unit and CEDAR.

This SOP has been endorsed by both the Science Strategy Meeting and the Science Operations Meeting. Please feel free to contact Stephen Sharp (stephen.sharp@mrc-epid.cam.ac.uk) if you have any questions, comments or suggestions related to this SOP.

2. Rationale and benefits

The rationale of this SOP is to ensure that all analytical work is clearly justified, accurate, transparent and reproducible.

Benefits from compliance with the requirements and recommendations in this SOP include the following:

- all outputs for which the Unit has primary responsibility can be easily linked back to an analysis plan, data source(s) and analysis programs used to generate the results reported in the output.
- authors of an output should be able to locate analysis programs/datasets long time after the analyses were first performed, and address external queries or requests for data/results.
- individuals working on similar types of analysis should be able to share their analysis programs with each other.
- there will be increased transparency in how the results presented in an output have been obtained, thus increasing the opportunity for other co-authors on an output to identify and correct errors prior to an output being submitted or published.

3. Statistical Analysis Plan (SAP)

For all analyses, a Statistical Analysis Plan (SAP) should be written, agreed with co-authors/collaborators, and saved in an appropriate folder (see section 7) prior to analyses being performed.

The following is the type of information that should be included in the SAP:

- Date when SAP was finalised. Dates of and justifications for any subsequent revisions.

- Name of SAP author and SAP reviewers.
- Statistical software to be used.
- Objectives and hypotheses.
- Descriptive analyses.
- Outcomes.
- Exposures.
- Potential confounders and effect modifiers.
- Statistical methods.
- Modelling strategy.
- Sensitivity analyses (i.e. analyses that assess the sensitivity of the results to the model assumptions).

Some of the above sections may be less relevant if the SAP relates to genetic discovery analyses which are hypothesis-free.

The level of detail to be included in a SAP will vary between different projects. At a minimum, the SAP should contain the information that would be included in the "Statistical Analysis" section of a results paper.

For clinical trials, a more formal and detailed template has been developed based on guidelines published in Gamble et al, JAMA. 2017;318(23):2337-2343. See section 11 for the location of this template.

4. Analysis software

Where possible, analytical work should be performed using either Stata or R (or both). Python may also be considered, preferably only if at least some co-authors also have experience in using this package.

Other specialised software may be needed (e.g. for genetics analyses); this should be documented in the SAP.

If a software package is required that is not available on the network, please email helpdesk@mrc-epid.cam.ac.uk ; the cost of purchasing such software would usually need to be met by the grant or research programme that requires it. If someone (e.g. a visiting worker) does not have an MRC-EPID account and needs remote access to core Unit software, their line manager should discuss with IT.

5. Analysis programs

All analysis programs should be written and annotated in such a way that co-authors could use them to replicate the analysis. Analysis programs should be text format files which include the following information:

- name of study/project.
- brief description of the specific purpose of the analysis program.
- name of the person who has written the analysis program.
- version of the software being used.
- names of any add-on packages that need to be installed before the program can run, e.g. user-written Stata commands from the Statistical Software Components (SSC) archive, R libraries.

- command to change working directory to a relevant folder, so that any outputs from the program are saved in that folder rather than on the C: or U: drive or some other system folder.
- command to read in the relevant dataset.
- comments throughout the file to aid understanding of what each section of code is doing.

Line spacing/indentation can enhance clarity. A "master" program which runs all the analysis programs in sequence can be useful, rather than including all analyses in a single program, which can quickly become very large and hard to navigate.

The name of the analysis program should be a concise description of its main purpose.

In some situations it may be useful to retain previous versions of programs for a particular analysis. If this is the case, the version number of each program should be included at the start of the file; previous versions should be moved to a clearly labelled sub-folder, so they are not confused with the current/final versions.

6. Datasets

Original data sources (e.g. internal data releases or externally provided data) should be stored in an appropriate network folder. They should not be moved or copied to other folders.

Prior to undertaking any analysis, a program should be written, whose purpose is to create a dataset for analysis based on the original data source(s), and also keep a record of relevant information related to the data source(s).

If the data source(s) are from within the Unit, the release IDs of the original data source(s) and any updates, together with the title and date of the original data request, should be included at the start of the file. If the data source(s) are from outside the Unit, similar information should be recorded where possible.

Examples of tasks that could be performed by this program include:

- remove potentially personally identifiable data (these data should already have been removed from internal data releases).
- merge or append any other relevant data.
- generate any new variables required for the analysis (e.g. a variable representing BMI categories based on values of a continuous BMI variable).
- rename and/or label variables.
- label values of variables (e.g. if there is a variable coded 1/2 for men and women, these can be replaced with value labels M and F).
- reshape data to wide or long format.
- set up data for e.g. survival analysis.

This "analysis dataset" can then be saved under a different name from the original data source, and used for all subsequent analyses. Once an output is published, the analysis dataset, not the raw data source, will usually be the most appropriate "data" to be provided externally if required.

7. Location of analysis work related to an output

All analysis work relating to an output (SAP, analysis programs, analysis datasets, outputs) should be saved in appropriately located folders on a network (Windows or Linux) drive (not a personal drive or memory stick). Any analyses of personally identifiable data should be performed and saved in folders on the private network.

Multiple copies of the original data source(s) should not be made. If a dataset has been provided from a source external to the unit, then a single copy of this dataset should be saved. Datasets should not be stored as attachments to emails.

There should be sufficient information in the analysis programs to enable the original data sources to be easily identified (see section 6).

8. Internal peer review of analysis work

Peer review of analysis work (especially analysis programs) within the Unit is encouraged whenever feasible. This increases the chance of identifying any errors prior to submission, and facilitates sharing of good practice. This could be done by co-authors of a particular output or other members of the same research programme. Making analysis work easy to find and understand will help to facilitate this.

9. When an output is published

As soon as possible after an output is published, the analysis work relating to this output should be tidied up, removing any unnecessary or old versions of programs, temporary datasets etc. The location (i.e. folder path) of the analysis work for a particular output should be sent to DMMgrs@mrc-epid.cam.ac.uk, so that a central repository of this information can be maintained.

10. Leaving the Unit

Analysis work for Unit research projects should remain on the network after a member of the Unit has left. The line manager and/or corresponding author is responsible for ensuring that this is the case.

11. Resources

Template for analysis plan for a clinical trial

Unit V: drive: Functional_Groups\Statistics\Clinical trial SAP template

CEDAR V: drive: Group\Statistics\Clinical trial SAP template

Examples of SAPs

Unit V: drive: P3_NutrEpi\People\Jusheng\EPIC_N\EPIC N repeats fa\Proposal and analysis plan\EPIC_N_Analysis Plan_FA.docx

CEDAR V: drive: Studies\MOVED\ICAD\Analyses\EvanSluijs - Car ownership & PA\Car ownership and PA_analysis plan.docx

Examples of analysis programs

Unit V: drive: Studies\ADDITION\10 year analysis\Analysis

Unit V: drive: Programme1_DiabetesAetiology\People\Luca\Production\lipid_drugs_mr\dofiles

CEDAR V: drive: Group\Biobank\Analyses\PAPH\Jenna and Oli (AT and health)\Stephen Sharp

Linux: /genetics/data/good_analytical_practice/example_dofiles

In-house exemplar Stata do files for specific analytical tasks

Unit V: drive: Functional_Groups\Statistics\Stata resources\In-house do files

CEDAR V: drive: Group\Statistics\Stata resources\In-house do files

In-house Stata commands for genetics analyses on Linux (written by Jian'an Luan)

Within Stata (on Linux), type:

net from /genetics/data/GWA/jianan/ado/myprogs

and then click on the program you wish to install.