Rules of engagement for Statistics Denmark - www.heart.dk

Security first

First of all the rules for statistics Denmark regarding data needs to be read, learned and understood. These rules are first of all designed to adhere to GDPR and avoid disclosure of microdata. The rules will not be repeated here, but from our experience particular care needs to be addressed in the following cases:

Descriptive tables can often have groups with few individuals. Such groups should first of all be avoided by proper selection of grouping definitions. If absolutely necessary numbers of 1,2 and 3 need to be replaced with ≤3, and be particularly aware that percentages in another column can also be used to calculate the real number. Such are not allowed.

Numbers at risk below Kaplan Meier graphs become small at the end of follow-up. Remove small numbers. It is rarely useful to report very small numbers, so please only provide decently large numbers and do not for example show “4” or “≤3” unless particularly necessary. In general limit the X-axis to only show the part of the graph with large numbers.

Graphs can be tricky, in particular scatter graphs. Depending on the ability to derive exact numbers on the axes such graphs may disclose microdata. The problem is particularly large with outliers that always need to be removed. You can report in the legend to figures that such outliers were removed.

Programs should always be generously commented on, but it is also necessary that programs can be exported from Statistics Denmark since these are the documentation of results if there should ever be critical questions to your research. Therefore, it is not allowed to disclose microdata in programs. If you for example need to remove/modify individuals, it is not allowed to use logic such as if-then structures based on then encrypted cpr-number (pnr) to remove such individuals. Instead, you need to take a longer course of creating variables that represent such individuals and use these variables to remove or modify. As an example an excessive number of children have a farther that represents the cpr-number “0”. Instead of identifying this number you can change these pnrs to missing based on the fathers not having a birth date or sex in other datasets.

Comments in programs are particularly dangerous. If-then structures can be searched, but identifying microdata in comments is not generally searchable.

Servers

Our network currently has 4 operating servers numbered 3-6. Servers 3 and 6 are back-up servers and available for use, but generally we advise not to use them. Servers 4 and 5 are large and program updates etc. are prioritized for these servers. Therefore only use servers 4 and 5.

Drives

There are a number of drives that are shared between our servers. Of these drives you should only use and know X, V and Z.

X is for raw data and user data cannot be placed here. As our projects become updated to DDV (Danmarks Data Vindue) all projects will have a similar basic data structure that is described in the document Basic_project_data_structure.pdf which you find on www.dst.heart.dk/girhub/programming guidance.

V and Z drives are for user data, and the following needs to be adhered to:

- Create a folder containing your full name and avoid blanks in the name using either underscore or camel case. The full name is necessary if we need to contact you, typically because of excessive disk use.
- Create a subfolder for each paper/report you work on
- We recommend that each folder is structured with the R-function heaven::createProject.
- Make sure that a very small and organized set of programs create the data for your publication. These programs should be commented on to an extent where others can understand the calculations. We encourage use of targets pipelines since it creates a useful and recognizable structure.
- Avoid excessive disk use. Habits of having consecutive series of large datasets with minor changes can easily use more space than we have. Check regularly your disk usage. For SAS users this implies avoiding permanent folders for intermediate data and only placing very selected datasets in such folders.
- The final dataset used for analyses should in general be kept in a permanent folder until the publication is final. This avoids the problem that your programs may provide new results if our data are updated.

Calculating

Currently SAS, R and STATA are available. Avoid STATA as it may soon disappear.
R users should constantly be aware of how much RAM they are using, in particular if programs are left open when you are not actively working. R users should run the carbage collector [gc()] regularly as R has an annoying habit of memory leakage meaning that memory you are not using any more is not released until you run the garbage collector.

Exporting

- On V and on Z there is a folder named “mail”. In one of these folders you need to create a subfolder with your email address as name.
- Place any files you want exported here. Not that csv-files are not allowed for some reason so change them to txt-files or save them as excel files. The reason we require this step is that it provides extra insurance that the files exported are also the files that are checked for microdata.
- Do not export long lists of raw output. Finish your data while on DST so that only final tables and graphs are exported. This rule is to make it feasible to check the data carefully.
- When exports are ready you can mail a person with export permission to check the data prior to export. If you have export permission you should not export yourself, but ask another, such that all exports are checked by two people.
- The exporter releases the files to DDV and you can then access them for download.
- Finally you should clear your mail-subfolder.

Data and programs for all

The V\data\alle folder is accessible from all projects and can only be modified by DST.
If you want to move programs from one project to another you can request our data manager to forward a request to move the program to V:\data\alle in a relevant subfolder.
A number of datasets without reference to individuals are available in this folder. Many will benefit from the presence of text codes for ICD8/10, details of medication based on vnr (varenummer).