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

Christian Torp-Pedersen

2023-11-04

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 are particularly tricky because they are generally long. In principle programs need to be kept securely since they are the documentation of our work if critical questions arise. We have provided some rules for "good programming practice" in a separate document. Such rules have in the past not been adhered to with a risk that microdata appears in programs. Therefore programs can neither be exported nor transferred to other projects. Any exception to this principle requires that at least 3 people have scrutinized the program and one of these should be the "older" associated with the project.

Penalties

If microdata are exported from Statistics Denmark and we notify Statistics Denmark of the violation - then the draconic penalties are not enforced. It is therefore important to immediately notify Statistics Denmark when violations are identified.

It has been a general observation that minor violations are very often caught when checking data prior to export. This creates an impression that the responsibility is that of the exporter rather than the requesting person. To create equal sense of responsibility it is therefore a rule that any minor violations causes that user to sustain a 30 day pause for any export.

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, data, programs

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/github/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*. * Adhere to the advice in the document "Good programming Practice". * Currently SAS, R and STATA are available. Avoid STATA as it may soon disappear.

Exporting results

- 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. Note 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. An office package is available on the serverse for this use.
- 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.
- When exporting data, do not assume that the person with the export license has the responsibility of ensuring against microdata. It is first of all the user that needs to ensure against microdata.

Exporting programs

Programs can be exported or moved to other project folders. These two procedures are both considered exporting by Statistics Denmark and both are subject to the draconic penalties used. Therefore we are amploying very strict rules resulting from the fact that the rules in "goodProgrammingPractice" cannot be assumed followed in the pase. * In general programs should neither be moved or exported. When absolutely necessary at least three people should check the proram and one of these should be from the network steering committee. This ensures that extreme care is taken to ensure against microdata. * We ensure that programs are made accessable even for old inactive projects

Search strings for R and SAS

When programs need to be exported (or moved to another project) it is important not to rely on just visual guidance to avoid microdata. The following search strings can be used. These are "regular expressions" and a good way to use them is from the Rstudio editor (also for SAS programs). When using the search facility you can tick the "Regex" option.

The following regex pattern looks for "pnr =", "pnr=", "pnr !" and "pnr!" (indicative of making logic depending on pnr numbers):

```
pnr\s*=\s*|pnr\s*!\s*
```

The following regex pattern looks for 10 consecutive numbers (indicative of a pnr number):

```
\d{10}
```

The following regex pattern looks for 10 consecutive letters or numbers (indicative of the special case of pnr numbers where there are missing values - note: this yields many false positives):

```
[A-Za-z0-9]{10}
```

The following regex pattern looks for places where you've written "1, 2, 3, one, two, three, en, to, tre" prefixed by a space and suffixed by a space, period or a comma (indicative of describing single patients in your script - note: may also yield false positives):

```
(?<=\s)1(?=[\s.,])|(?<=\s)2(?=[\s.,])|(?<=\s)3(?=[\s.,])|(?<=\s)one(?=[\s.,])|(?<=\s)two(?=[\s.,])|(?<=
```

Search strings for SAS

The following regex pattern looks for "pnr =", "pnr=", "pnr !" and "pnr!", "pnr eq", "pnr ne" (indicative of making logic depending on pnr numbers).

```
pnr\s*=\s*|pnr\s*!\s*|pnr\s*eq|pnr\s*ne
```

Data and programs for all

The *V\data\alle* folder is accessible from all projects and can only be modified by DST. 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).