General data organisation

Christian Torp-Pedersen / Mikkel Porsborg Andersen

2023-07-21

Data to projects are provided from many sources, the most important being Statistics Denmark (DST) and the Health data organisation (Sundhedsdatastyrelsen, SDS). Data from these sources are by DST conceived as "base data" (Grunddata) and are systematically placed in a folder named **Grunddata**.

The current listing of dataset is very general and the project you are working on may contain fewer datasets or additional datasets.

Data supplied from other sources are placed in other folders according to permissions for that project.

Therefore the general structure of all projects will be a folder names "population" which has the populationk of the study when relevant, a folder names "external data" with data develivered from other sources than DST and SDS.

This document provides an overview of the structure of **Grunddata**. An important principle for the structure of data has been to change as little as possible to the orginal data supplied. The reason for this is that it allows all users to interrogate DST and SDS for description of registers and definitions of variables. These institutions have on their web-pages descriptions of datasets and well as variable explanations.

A full list of all external data including variable explanations is available on DDV (Danmarks Data Vindue).

Data management with the new structure will result in changes of many programs. To ease this process we have written a SAS and R program to read data in a structured way. These programs are found in V:/Alle/skeleton and are also visible on www.heart.dk/github/programming guidance/skeleton.

One very important change for many users is that we provide multiple files for each register according to the DST habit of having one file per year. This change is made to ease data organisation as many users only need to interrogate data from selected years. It further can help to increase speed of data processing by parallising calculations - the document **Parallel processing with SAS and R** describes this. Also, the programs for SAS/R enabling data management are available in the folder V:/data/alle/skeleton

The following headlines correspond to the subfolders in **Grunddata** on projects.

Cancer

This includes the cancer registry from SDS (t_tumor)

Death

T_dodsaarsag_1 is the national cause of death register until 2001 when new death certificates were introduced

T_dodsaarsag_2 is the new cause of death register. Note that attempts have been made to provide simple translations of new to olds cause of death definition, but this should be discouraged. It is important to note the update of causes of death is very much delayed.

Simple registration of death is in the **dod** table. This table combines information from DST and SDS and is updated as much as other data in the project

Laboratory

This folder first of all holds **lab_forsker**. NPU codes corresponding to tests are found in V:/data/alle/blodprøver

Historically we have collected laboratory results from willing sources and these are placed in separate subdirectories corresponding to the old municipal of Copenhagen (Kbh amt), Copenhagen general practitioners laboratory (KPLL), North Region of Denmark and Roskilde. The advantage of these datasets is that they reach further back in time than lab_forsker.

When relevant the **pato** register of microscope examinations is also supplied in this directory.

LPR

LPR1 uses ICD8-codes and was 1994 replaced with LPR2 and ICD_10 codes. This continued until about march 2019 where LPR3 started. LPR3 is structured very differently from LPR1/2. It has therefore been decided to maintain the partially digested LPR1/2 data in the following deliveries: **diag_indl** contains all diagnoses and selected administrative data. **opr** includes all procedure codes and all examination codes (sksube) - again with added administrative data.

Some projects will also include **opr_old** which are ICD8 procedures prior to approximately 1994. Other projects also have the file **lpr_bes** which are outpatient contacts from lpr2.

Psychiatric admissions were provided in independent tables. Note than occasionally psychiatric diagnoses appear in the somatic data, but not consistently.

LPR3 datasets are supplied without changes for the populations relevant for the project.

Medication

This directory includes all **lmdb** files available and also **laegelmiddeloplysninger** which holds further details of medications.

Mfr

mfr is the national birth register and this dataset has data from 1997-2019.

We have older data also **mfr_lfoed** - born alive prior to 1997 **mfr_dfoed** - stillborn prior to 1997 **nydfoed_2010** and **nylfoed_2010** are also delivered and may not provide new data.

With the introduction of LPR3 all births from 2019 are found in the LPR3-table **nyfoedte**.

Nursinghome

The dataset **plhjem** is the data we have from nursinghomes generated by DST by request from us and using the number of old people at an address to qualify an address to be examined as to whether is was a nursinghome.

From 2016 the more official data **aepi** has nursing home data.

The registers **aefv**, **aelh**, **aeph** and **aetr** are various types of personal assistance

Population

The main population register is the **bef** tables for each year since 1985 and from 2008 each quater of a year. This is the main official register to define danish residents at any time.

Prior to 1985 the **fain** tables provide some of the data found in **bef**.

In addition the dataset **sexBirth** provides all sex definitions (0=female,1=male) available from the cpr-register (t_person) Projects may or may not directly have the **t_person** register

Country of origin is found in the **iepe** dataset and all immigrations into and away from Denmark are found in the **vnds** table.

Social

The tables **indXXXX** hvor Xs represent years include income and fortune of invididuals. This table also includes individual income adjusted for family composition. The table does not adjust for inflation.

The tables **uddaXXXX** has maximal education for all individuals. There are SAS formats to translate to understandable values and also a R-function in heaven to do this.

The table **uddfXXXX** contains also maximal education. Only the last year is necessdary to include as it is uptated with new data annually. It appears to include maximal education for more people than udda-files, perhaps because imported educations are also registered here.

The table **dreamXXXXXX** is an updated version of dream that for each week has codes indicating public support and for each month indication of association to particular working areas.

Som projects include **kotre** and **koto** which are student registers.

SSS

SYSI and **SSSY** are data from practitioners to many types representing different time periods.

Other data

For all projects there is a folder named **External Data**. The data in this folder can be found in the description files for the project - named variables and datasets. Further explanations can be found in the approvals from DST and SDS or on DDV.

V:/data/Alle

On the V-path subfolder “alle” is a number of moderately organised folders for a variety of anonymous and useful data. Not in particular details of medication in LMDBdata, text versions of diagnoses in ICD8 and ICD10

Note also here the “skeleton” folder that includes sample programs to import data from the structure required by Statistics Denmark