OPEN Open Patient data Explorative Network

Guide til registerforskning

OPEN Open Patient data Explorative Network

Indhold

OPEN Open Patient data Explorative Network

Forord

Når mailen fra Danmarks Statistik (DST) eller Sundhedsdatastyrelsen (SDS) tikker ind i indbakken med meddelelsen om, at data er klar, kan det være med lidt blandede følelser, at man lukker op for Pandoras æske for at se, hvad den gemmer. Er det første gang, at man skal arbejde med registerdata, kan de forskellige mapper og filer m.m., der møder en, virke en smule uoverskuelige, og man kan blive usikker på, hvordan man begynder arbejdet med at klargøre og oprense sine registerdata til videre arbejde.

Formålet med denne "Guide til registerforskning" er at give dig en hjælpende hånd til at komme i gang med arbejdet med registerdata. Lige fra at få åbnet alle de mange mapper og filer, få renset data og klargøre det til analysen. Denne guide kan dog ikke anvendes som en "facitliste", men kan bruges som vejledning og inspiration til arbejdet. Fremgangsmåden varierer alt efter om du skal arbejde med data på forskermaskinen hos DST eller SDS, og derfor er en række afsnit i denne guide opdelt derefter. Denne guide er endvidere skrevet med udgangspunkt i, at statistikprogrammet Stata benyttes.

Det skal pointeres, at der er flere måder at tilgå og håndtere registerdata på, og nærværende guide bygger på OPENs forskeres erfaringer med at klargøre data på en struktureret og overskuelig måde.

Til denne guide er der knyttet en række hjælpe Stata-do-filer, som kan benyttes og som gennemgås i denne guide. Send en mail til OPEN.registry@rsyd.dk, hvis du ønsker, at vi skal sende dig nedenstående do-filer

"1_Konvertering af data fra SDS til STATA filer"

"2_Sammenlægning af filer med forskellige årstal"

"3_Tjek om det korrekte materiale er leveret"

"4_Crude to Clean"

"5a_Datakontrol_Overordnet datakontrol"

"5b_Datakontrol_Diverse tjek af tabeller"

"6_STATA kommandoer"

Attention: Hvis navngivningsprocesserne fra do-filerne følges, kan do-filerne køres uden de store ændringer – du skal selvfølgelig tjekke, at indholdet i do-filerne stemmer overens med det data, som du har fået udleveret!

Har du forslag til forbedringer af denne guide, tager vi med glæde imod forslag.
Kontakt os meget gerne på (OPEN.registry@rsyd.dk). OPEN fralægger sig ansvaret for eventuelle fej i do-filene.

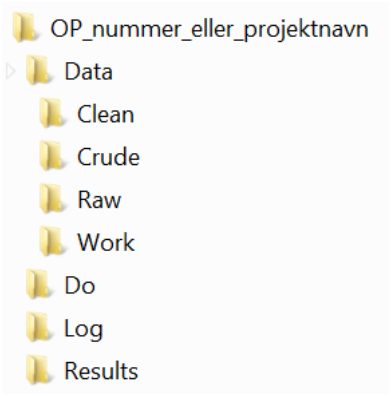
God arbejdslyst!

1 Strukturering af mapper

Arbejdet med klargørelse og oprensning af data kan nemt ende med rigtig mange datasæt, do-filer og log-filer, og det kan derfor blive rigtig svært at finde rundt i ens arbejde. OPEN anbefaler, at du strukturer dit arbejde med registerdata, som illustreret i nedenstående eksempel (figur 1). Formålet med denne struktur er at sikre en systematisk klargørelse og oparbejdelse af data, for på den måde at minimere risikoen for fejl. Ligeledes er de hjælpe-do-filer, som knytter sig til denne guide lavet, så de passer til denne struktur.

Overordnet oprettes en mappe, der kan navngives med f.eks. OP_nummeret eller projektnavnet.

Denne overordnede mappe underinddeles med følgende undermapper: ”Data”, ”Do”, ”Log”, ”Result” og ”Work”. Mappen ”Data” underinddeles yderligere i ”Raw”, ”Crude” og ”Clean”. Hvad de respektive mapper skal indeholde er beskrevet i figur 1.



OP_nummer_eller_projektnavn

Data - I denne mappe arbejdes med klargøring af data

Clean - I denne mappe gemmes ”clean” (rensede) data

Crude - I denne mappe gemmes de sammenlagte filer midlertidigt

Raw - I denne mappe gemmes rå data

Work - I denne mappe gemmes det/de datasæt der arbejdes på

Do - I denne mappe gemmes de do-filer, der benyttes under arbejdet

Log- I denne mappe gemmes de log-filer, der dannes under arbejdet

Result - I denne mappe gemmes analyseresultaterne

Figur 1: Illustration af mappestruktur

2 Konvertering af data

Data fra Danmarks Statistik (DST) eller Sundhedsdatastyrelsen (SDS) vil som oftest skulle konverteres til Stata (eller et andet statistikprogram, som ønskes benyttet til projektet), før man kan gå i gang. I de følgende eksempler gennemgås, hvordan dette kan gøres afhængigt af om du har data hos DST eller SDS.

2.1 Data fra Danmarks Statistik

Når du har ansøgt om data via DST vil du skulle arbejde på deres Forskerservere via en fjernadgang. På nedenstående link kan du finde vejledninger, der beskriver, hvordan du logger på DST Forskerservere:
https://www.dst.dk/da/TilSalg/Forskningsservice/brug-af-forskermaskiner

På DSTs forskerserver vil data være placeret på E-drevet, der indeholder mapperne rawdata og workdata (og en række andre, men det er primært mapperne rawdata og workdata du skal bruge). Data er organiseret efter projektets sekscifrede projektnumer, der er på formen 70XXXX.

I mappen rawdata (E:\rawdata\<projektnummer>,) lægger DST Forskerservice det data, der er tilknyttet projektet. Data vil som oftest være placeret i følgende tre undermapper: Eksterne data, grunddata og population.

Data i mappen rawdata er skrivebeskyttet og kan kun ændres af DST Forskerservice.

Workdata er dit arbejdsområde på projektet E:\workdata\<projektnummer>. Her kan personer tilknyttet projektet arbejde med data, og det er også her både programmer, afledte data og resultater gemmes. Mappestrukturen i workdata anbefaler vi som illusteret i figur 1.

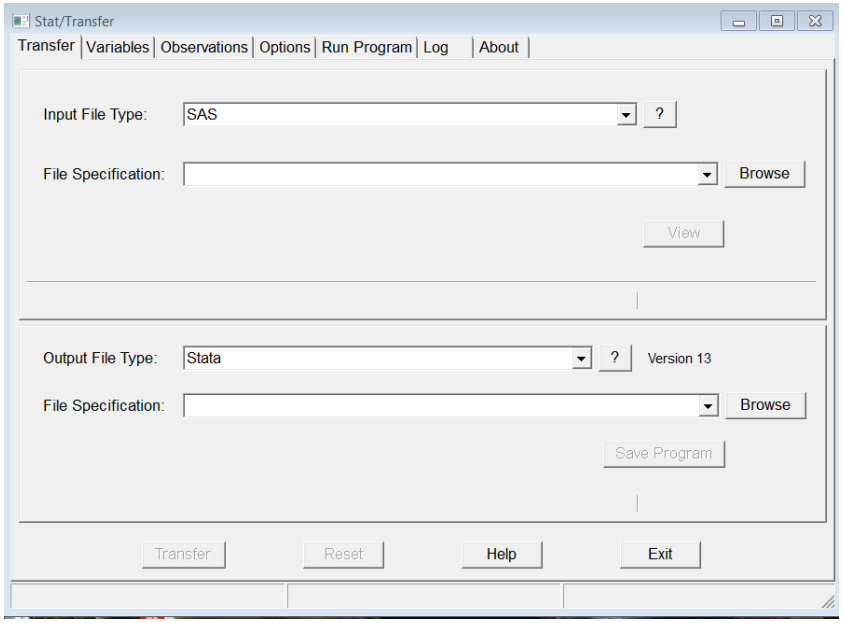
Data modtaget fra DST Forskerservice leveres som standard i SAS format og vil ligge i mappen ”Grunddata”, og hvis du ønsker at arbejde i Stata skal data først konverteres fra SAS til Stata.

Da data i ”Grunddata” er skrivebeskyttet, kan de ikke konverteres direkte i denne mappe. Derfor startes der med at tage en kopi af alle de SAS filer, der ønskes konverteret til Stata. Disse anbefales at gemme i mappen ”Raw” under ”Data” under Workdata (figur 1).

2.1.1 Konvertering af data via StatTranfer

Konverteringen af data kan gøres via ikonet StatTransfer, som du finder på skrivebordet ved DST

Når du klikker på ikonet ”StatTransfer” kommer dette billede frem (Billede 1)



Billede 1: StatTransfer

Forklaring - den øverste boks:
Input File Type: det format datafilen er i (som oftest SAS)
File Specification: hvor du henter datafilen

Forklaring nederst boks:
Output File Type: det format du ønsker datafilen konverteret til
File Specification: hvor du gemmer de konverteret data

Her kan du konvertere en fil ad gangen eller bruge "Run Program"-fanen, hvis der skal konverteres flere filer på en gang. Dette uddybes ikke yderligere i denne guide, men OPEN kan kontaktes ved behov.

2.1.2 Konvertering af data via "StatTransfer Command Processor"

Skal der konverteres flere filer på en gang foreslår OPEN at bruge "StatTransfer Command Processor"

I det fremkommende vindue skriver du nedenstående kommado (kommandoen skrives på en linje)

copy E:/workdata/70XXXX/Projektnavn/Data/Raw/*.sas7bdat E:/workdata/70XXXX/Projektnavn/Data/Raw/*.dta

og trykker på Enter.

Forklaring
copy E:/workdata/70XXXX/Projektnavn/Data/Raw/*.sas7bdat = filstrukturen hvor SAS filerne ligger.

Hvis du følger mappe opbygningen som vist i figur 1, vil du i denne kommando kun skulle ændre følgende:

Projektnavn = Her skal du skrive navnet på den overordnet mappe du har lavet i "workdata"
70XXXX = her skal du skrive det nr., som dit projekt har fået ved DST, der starter med 70

E:/workdata/70XXXX/Projektnavn/Data/Raw/*.dta = filstrukturen hvor dine filer skal placeres efter, at de er konverteret fra SAS til Stata filer (i dette tilfælde i samme mappe, hvori SAS filerne allerede lå).
Her skal du også ændre til dit eget Projektnavn og DST nr. (som beskrevet ovenfor).

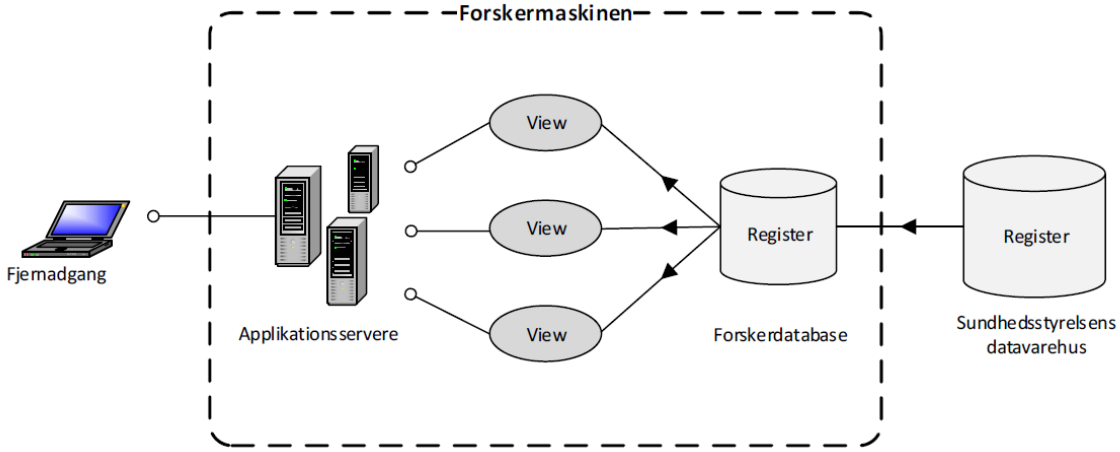
Når alle de ønskede filer er konverteret til Stata, anbefales det af pladsmæssige årsager at slette alle SAS filerne i din egen "Raw"-mappe, da de originale filer jo allerede ligger skrivebeskyttet i den mappe, som DST har lagt dem i ("rawdata→Grunddata").

2.2 Data fra Sundhedsdatastyrelsen (SDS)

2.2.1 Dataplacering og dataadgang på Forskermaskinen ved SDS

Dataplacering

Registrene på Forskermaskinen ved SDS er lagret i Forskerdatabasen. Herfra er det muligt at tilgå data med programmerne SAS, R og Stata fra applikationsserverne. På hvert projekt opretter Forskerservice dataadgang ved at danne projektspecifikke views i databasen udtrukket fra grunddata. Det betyder, at man derfor selv skal "downloade" sine data ud fra disse views (SDS har ikke på forhånd lavet dine filer og lagt dem i en mappe (som f.eks. ved DST)) (Figur 2).



Figur 2

SDS lægger én do-fil med alle de bestilte filer plus hjælpefiler, men de ligger i en stor pærevælling, så man skal selv sorterer. Forskellen fra DST er, at hos SDS er det en adgangssti til deres database man får udleveret, og derfor vil det altid være muligt, at hente den seneste version af data, så længe man har adgang til projektet. Nedenfor er det en vejledning til, hvordan det kan data kan hentes og sorteres.

Dataadgang

Har du ansøgt om data via SDS vil du skulle arbejde på deres Forskermaskine via en fjernadgang. På nedenstående link er der en vejledning til login på SDS Forskerservice
https://sundhedsdatastyrelsen.dk/da/forskerservice/forskermaskinen

Se vejledning til at logge på Forskermaskinen (PDF)

Når du har logget på SDS Forskermaskine skal du på F-drevet (Forskerdata(F:)), for at hente dine data.

Når du åbner mappen "Projekter" kommer der en lang liste med forskellige FSEID-projektnumre frem, og du skal finde mappen med projektets FSEID-nummer og åbne denne mappe. Projektmappen kan tilgås af alle de brugere, som er tilknyttet projektet. Endelige datasæt skal placeres heri.

For at hente data kan Stata bruges og det anbefales at gøre det via do-filer. Du kan anvende do-filen *"1_Konvertering af data fra SDS til STATA filer"* til at hente dine data i Forskerdatabasen og få dem gemt som Stata-filer. Denne do-fil kan bruges med få projektspecifikke rettelser, hvis mappestrukturen, som vist i

figur 1, er benyttet. Du skal dog være opmærksom på, om do-filen dækker alle de filer og årstal du har bestilt.

3 Dataoprensning

Når data er konverteret til Stata er næste opgave oprensning af data. Dette arbejde består af tre opgaver:
1) sammenlægning af datafilerne - de enkle årstal for hver tabel lægges sammen i én datafil for hver tabel,
2) gennemgang af det leverede data for at tjekke at de stemmer overens med de data, der er ansøgt om,
3) afslutningsvis oprensning af data.

3.1 Sammenlægning af årstal

Data fra henholdsvis DST og SDS vil som oftest blive leveret i rigtig mange datafiler- én datafil for hvert af de år, der er ansøgt om. Er der eksempelvis ansøgt om data fra Landspatientregisteret f.eks. fra tabellen t_adm i perioden 2001 til 2010 vil data ligge som ti t_adm-datafiler, men derudover vil der også være tabeller med uafsluttede patienter og tabeller med patienter fra privathospitalerne. Landspatientregisteret har lagt disse to i separate filer siden henholdsvis 2003 for uafsluttede og 2005 for privathospitaler.

Datafilerne for de enkle årstal sammenkøres i én datafil ved anvendelse af do-filen *"2_Sammenlægning af filer med forskellige årstal"*. I den pågældende do-fil er der vist to metoder til sammenlægningen.

Metode 1: Her sammenlægges årstallene trinvis – denne metode kræver lidt mere skrivearbejde i do-filen, men giver et godt overblik over, hvad det er, der bliver gjort.

Metode 2: Her sammenlægges de enkle årstal ved hjælp af en løkke – hvor alle årstal fra en tabel f.eks. t_adm sammenlægges på en gang i en kommando.

Hvilken metode der anvendes er underordnet. Det vigtigste er at få én datafil, der indeholder alle årstal for hver af de ansøgte tabeller jf. eksemplet med t_adm. Datafilerne med de sammenlagte årstal gemmes i mappen "Crude".

3.2 Tjek af leveret data

Næste trin i oprensningen af data er at sikre, at de modtagne data stemmer overens med det, der ifølge udtræksbeskrivelsen er ansøgt om. Dette gøres ved at tjekke, at de modtagne data indeholder de årstal, tabeller og variabler, der er beskrevet i udtræksbeskrivelsen. Do-filen *"3_Tjek om det korrekte materiale er leveret"* viser, hvordan dette kan gøres.

3.3 Dataklargøring

Som det sidste i processen skal data klargøres og renses. I denne proces ændres data fra "Crude" til "Clean". Dette medfører f.eks., at variabler omdøbes, der sættes label på nogle variabler, samt at f.eks. operations- og behandlingskoderne fra LPR samles, samt at de oprensede datafilerne flyttes fra "Crude" til "Clean"-mappen. Data i "Crude" og "Clean" mapperne indeholder stort set de samme data. Efter data er

klargjort og gemt i ”Clean”, kan man overveje at slette filerne i ”Crude” mappen, Formålet er at spare plads på serveren. Har man ikke problemer med pladsmangel, kan man vælge at beholde filerne i ”Crude” mappen.

Processen fra ”Crude to Clean” er vist i do-filen *”4_Crude to Clean”.* I denne fil har vi lavet et eksempel på forskellige tabeller med indhold, og hvordan et oprensningsforløb kan se ud. Du skal dog være opmærksom på, at dine data meget vel kan afvige fra viste eksempel. Du skal også være opmærksom på, at har du fået data fra SDS kan navngivningen i do-filen, bruges uden de store ændringer.

3.4 Optimering af kørselstid i Stata

Det kan anbefales at udarbejde do-filen på en mindre del af datasættet i de tilfælde, hvor datasættet er meget stort. Det gør arbejdet hurtigere. Når alle do-filer er klar, kan man så køre hele sit datasæt igennem do-filerne. OPEN har lavet en selvstændig guide til, hvordan man kan gøre for at optimere kørselstiden i Stata, hvis dette er et problem for en. Denne guide hedder ”reduktion af filstørrelse…” og findes ligeledes på vores hjemmeside. Til denne guide har vi ligeledes udarbejdet en do-fil med relevante kommandoer til dette.

4 Datakontrol

Data er nu oprenset, men inden analysearbejde kan påbegyndes, er det vigtigt, at der udføres datakontrol. Dette er *IKKE* at forveksle med det tidligere udførte tjek af de leverede data. Datakontrollen har til formål at sikre datafilerne indeholder de korrekte data og er sammensat korrekt. Dette kan udføres som en overordnet datakontrol efterfulgt af tjek af relationerne mellem tabeller.

4.1 Overordnet datakontrol

Den overordnede datakontrol kan f.eks. bestå i at tjekke at ingen cpr-numre er repræsenteret mere end én gang, at alle individer har et køn, og at recnum er unikke (recnum er et id, som bruges til at identificerer den enkelte kontakt på et hospital).

Do-filen *”5a_Datakontrol_Overordent datakontrol”* indeholder forskellige Stata-kommandoer til den overordnede datakontrol. I forbindelse med denne datakontrol er det vigtigt at være opmærksom på, at dit data kan afvige, derfor er do-filen *”5a_Datakontrol_Overordnet datakontrol”* kun tiltænkt som en inspiration og *IKKE* er en facitliste. I do-filen er dog beskrevet, hvilke opmærksomhedspunkter der findes for de enkelte variabler. Har du flere i dit eget datasæt, så find inspiration i do-filen til at lave den kontrol du har brug for.
Derudover er det vigtigt, at Stata-kommandoerne i do-filen køres trinvist, så resultaterne for hver af de kørte kommandoer gennemgås/tjekkes løbende.
Hvis navngivningsforslagene fra de tidligere do-filer er brugt, vil der være mindre chance for at do-filen fejler under kørsel!

4.2 Tjek af relation mellem tabeller

Efter den overordnede datakontrol er det vigtig at tjekke relationer mellem de modtagne tabeller. Det kan f.eks. være tjek af om der er sammenhæng mellem datoer, f.eks. at dødsdatoen i CPR-registeret den sammen som i dødsårsagsregisteret. Hvis der er uoverensstemmelse mellem dødsdatoer i cpr og dødsårsagsregisteret, så anbefaler OPEN, at cpr registeret bruges.

I do-filen *"5b_Datakontrol_Diverse tjek af tabeller"* er der en række eksempler på diverse tjek af tabeller og relationer. På samme måde som med den overordnede datakontrol er det vigtigt, at Stata-kommandoerne i do-filen køres enkeltvis og resultaterne for hver kommando gennemgås løbende samt at do-filen IKKE bruges som en facitliste, men fungerer som inspiration.

5 Stata-kommandoer

I do-filen *"6_STATA kommandoer"* er der listet nogle af de mest anvendte Stata-kommandoer. I Stata er der også en indbygget hjælpefunktion, hvor der er masser hjælp at finde. Denne hjælpefunktion fremkommer ved enten at trykke på "help" i menulinjen, og derefter "Search", hvorefter man indtaster sit søgeord: f.eks: date. Derudover er det også muligt at taste "help" og det ønskede søgeord (bemærk det skal skrive med småt og uden "") i kommandofeltet. Eksempel: help date

Det er også muligt at google sig frem til masser af hjælp, og der findes timevis af undervisningsvideoer på YouTube; det er ikke alt der lige anvendeligt, men der er masser af virkelige godt materiale at finde derude.

Derudover tilbyder OPEN en række workshops, seminarer og kurser omkring emnet, blandt andet et introduktionskursus i Stata samt forskellige registerkurser. Se nærmere via dette link:

https://www.sdu.dk/da/om_sdu/institutter_centre/klinisk_institut/forskning/forskningsenheder/open/kurser_seminarer

God arbejdslyst!