SUNDHEDSDATA-
STYRELSEN

27. januar 2020

FOR – formidling
Forskerservice

Vejledning til inklusion af egne data i et projekt på Forskermaskinen

På Forskermaskinen er det muligt at få inkluderet data fra andre kilder, end dem der er på Forskermaskinen. Det kan for eksempel være data, du selv har indsamlet, eller data du får videregivet fra andre. Uanset hvor disse data kommer fra, betegner Forskerservice dem som egne data. De egne data kan være personhenførbare[1] eller ikke personhenførbare. Denne vejledning beskriver, hvordan du skal gøre for at få egne data inkluderet i dit forskningsprojekt på Forskermaskinen.

Vejledningen er delt op i tre dele:

1. Ansvar for egne indsendte data
2. Sådan sender du dine data ind
3. Uddybning af regler vedrørende egne indsendte data
   a) Hjemmel - godkendelse og databehandleraftale
   b) Datasikkerhed - pseudonymitet og krav til formater
   c) Indsendelse af data - formater og restriktioner

Hvis du ønsker at indsende personhenførbare data, er alle tre underpunkter relevante. Hvis du ønsker at indsende ikke-personhenførbare data, er det alene punkt 2, der er relevant. Du skal derfor starte med at undersøge om data er personhenførbare. Hvis du er i tvivl, er dine data formentlig personhenførbare, og du bør behandle dem som sådan.

Ansvar for egne indsendte data

Selvom du sender dine egne data ind til dit projekt på Forskermaskinen, er dataansvaret fortsat dit. Det betyder, at det er dit ansvar, at data bliver behandlet i henhold til databeskyttelseslovgivningen og i henhold til reglerne på Forskermaskinen.

Data på Forskermaskinen må ikke indeholde direkte personhenførbare oplysninger som CPR-numre, løbenumre og andre identifikationsnøgler samt navne og adresser.

[1] Personhenførbare data, er data, der alene eller sammen med andre data, kan identificere et individ. Det kan for eksempel være gennem et personnummer, navn, adresse, løbenummer eller blot så mange oplysninger om en person, at personen bliver identificerbar.

Når du indsender data til Forskermaskinen, er det dit ansvar at sørge for:

- *Ikke* at indsende variable der indeholder navne og adresser
- At undersøge samtlige variable for indhold af CPR-numre, løbenumre og andre identifikationsnøgler samt navne og adresser
- At oplyse *alle* variable, der indeholder CPR-numre, løbenumre og andre identifikationsnøgler

Sanktion

Hvis Forskerservice ved indlæsning af data på Forskermaskinen eller efterfølgende bliver opmærksomme på, at der på et projekt ligger egne data, der indeholder navne, adresser eller ukrypterede CPR-numre, løbenumre og andre identifikationsnøgler, anses dette som et brud på retningslinjerne for arbejdet på Forskermaskinen. Brud på retningslinjerne vil blive sanktioneret på lige fod med hjemsendelse af mikrodata.

Sådan sender du dine data ind

1. Undersøg om dine data er omfattet af din dataansvarlige institutions godkendelse af dit projekt. Vi anbefaler desuden, at Forskerservice fremgår som databehandler/databehandlingssted på anmeldelsen til den dataansvarlige institution.

2. Undersøg om din dataansvarlige institution har indgået en databehandleraftale med Forskerservice.

3. Tjek at dine data opfylder følgende kriterier:
   a) Ingen navne og adresser
   b) Som udgangspunkt ingen lange karaktervariable. Det vil sige, at antallet af karakterer skal være < 40. Eventuelt længere prosa-variable, bør i stedet indsendes grupperet og standardiseret. Hvis du har behov for tekstvariable på > 40 karakterer, til opfyldelse af projektets formål, skal du fremsende en begrundelse for dette, samt en markering af de relevante variable.
   c) Som udgangspunkt ingen time-stamps. Disse skal som udgangspunkt reduceres til datovariable forud for indsendelse. Hvis du har behov for selve timestampet til opfyldelse af projektets formål, skal du fremsende en begrundelse for dette, samt en markering af de relevante timestamps.

4. Indsend anmodning om at få inkluderet dine data på dit projekt på Forskermaskinen.
   - *Hvis det er et nyt projekt indsendes anmodningen sammen med selve ansøgningen om at få adgang til data på Forskermaskinen*
   - *Hvis der er tale om et projekt, hvor I allerede har adgang, skal anmodningen sendes til forskerservice@sundhedsdata.dk*

Anmodningen skal indeholde:

a) FSEID (kun hvis dit projekt er igangværende, og dermed har fået tildelt et FSEID)
b) Angivelse af kilde til hvert datasæt
c) Angivelse af omtrentlig størrelse på datasættet i GB (max. 100 GB per fil)
d) Angivelse af filtype
    - Bemærk, at dine datasæt skal indsendes som enten sas7bdat-fil eller flad fil (.csv eller .txt).
e) Angivelse af den anvendte delimiter ved flad fil
    - Indsender du en flad fil, skal du oplyse, hvilken delimiter du anvender. Bemærk, at den anvendte delimiter ikke må indgå som værdi i filen. Forskerservice anbefaler brug af ¤ som delimiter, da denne sjældent indgår som værdi i datasæt.
f) Angivelse af tabelnavn(e)
g) For hver tabel angives de variable, der indeholder direkte personhenførbar information, og dermed skal krypteres ved indlæsning. Eksempler på personhenførbare værdier er:
    - CPR-nummer
    - Projektdeltagers løbenummer
    - Patient ID eller unikke prøve ID
h) For hver tabel angives eventuelle variable, der indeholder timestamps, samt begrundelse for, hvorfor der er behov for timestamp
i) For hver tabel angives eventuelle tekstvariable > 40 karakterer, samt begrundelse for, hvorfor der er behov for variablen.

Du skal således ikke indsende en komplet variabelliste.

5. Afvent svar fra Forskerservice.

6. Du modtager link til at kunne uploade data. Uploader du mere end én datafil, må du meget gerne pakke filerne i én zip-komprimeret mappe.

7. Forskerservice giver besked, når du har adgang til dine data på Forskermaskinen.

Uddybning af regler om egne indsendte data

Hjemmel - godkendelse og databehandleraftale

Dit forskningsprojekt skal være opført på den dataansvarlige institutions fortegnelse. Det er det, når du har fået en godkendelse til at gennemføre projektet fra din dataansvarlige institution. Når du inkluderer egne data i projektet, skal du sikre dig, at de egne data også er omfattet af fortegnelsen. Hvis de egne data ikke er omfattet af fortegnelsen, skal der foretages en ændring på fortegnelsen, og projektet skal have godkendt ændringen af den dataansvarlige institution,

inden data kan inkluderes i dit forskningsprojekt på Forskermaskinen. En ændring af fortegnelsen og fornyet godkendelse af ændringen kan for eksempel være relevant, hvis inklusionen af egne data medfører ændringer i populationsstørrelsen, formål eller andre typer af persondata.

Datasikkerhed - pseudonymitet og krav til formater

Data, der skal på Forskermaskinen, skal være pseudonyme. Pseudonymitet betyder, at der ikke må være data, der indeholder CPR-numre, navne, adresser og andre informationer, der gør data direkte personhenførbare. Disse regler gælder også for dig, når du skal inkludere egne data.

For dig betyder det konkret, at:

- Data må ikke indeholde navne og adresser
- Data må som udgangspunkt ikke indeholde karaktervariable på over 40 karakterer. Længere prosa-variable skal som udgangspunkt standardiseres og grupperes inden indsendelse
- Data må som udgangspunkt ikke indeholde time-stamps. Timestamps betyder angivelse af klokkeslæt
- Du skal sikre dig, at der ikke er variable, der 'gemmer' på værdier som navne og adresser eller CPR-numre, løbenumre og andre identifikationsnøgler.

Forskerservice krypterer CPR-numre, løbenumre og andre identifikationsnøgler. Den anvendte krypteringsnøgle sikrer, at det er muligt at koble dine egne data med de øvrige data i forskningsprojektet.

Inden indsendelse af egne data til Forskerservice skal du indsende:

a) Angivelse af kilde til hvert datasæt
b) Angivelse af omtrentlig størrelse på datasættet i GB (max. 100 GB per fil)
c) Angivelse af filtype
    - Bemærk, at dine datasæt skal indsendes som enten sas7bdat-fil eller flad fil (.csv eller .txt).
d) Angivelse af den anvendte delimiter ved flad fil
    - Indsender du en flad fil, skal du oplyse, hvilken delimiter du anvender. Bemærk, at den anvendte delimiter ikke må indgå som værdi i filen. Forskerservice anbefaler brug af ¤ som delimiter, da denne sjældent indgår som værdi i datasæt.
e) Angivelse af tabelnavn(e)
f) For hver tabel angives de variable, der indeholder direkte personhenførbar information, og dermed skal krypteres ved indlæsning. Eksempler på personhenførbare værdier er:
    - CPR-nummer
    - Projektdeltagers løbenummer
    - Patient ID eller unikke prøve ID
g) For hver tabel angives eventuelle variable, der indeholder timestamps, samt begrundelse for, hvorfor der er behov for timestamp

h) For hver tabel angives eventuelle tekstvariable > 40 karakterer, samt begrundelse for, hvorfor der er behov for variablen.

Du skal således ikke indsende en komplet variabelliste.

Indsendelse af data - formater og restriktioner

Når Forskerservice har modtaget og godkendt din anmodning om indsendelse af dine data, har Forskerservice lov til at modtage dine data. Forskerservice orienterer dig, når vi er klar til at modtage dine data. Data skal uploades til Forskerservice via vores upload-løsning.

For upload af data benytter Forskerservice ShareFile® til deling af data, hvilket gør det nemmere og mere sikkert at sende dine data til os.

Når du er klar til at uploade dine data, skal du give Forskerservice besked om dette, hvorefter du vil modtage en mail med et personligt upload-link. Dette link kan kun anvendes i 7 dage. Linket kan også videresendes til en anden, hvis det ikke er dig selv, der skal uploade data.

Ved at følge linket vil du komme til Sundhedsdatastyrelsens upload-side, hvor du kan uploade din datafil til Forskerservice. Når du uploader datafilen, bliver den automatisk krypteret. Det er derfor ikke nødvendigt, at du krypterer datafilen.

De filer, du uploader skal overholde følgende:

- Dit datasæt skal indsendes som enten en sas7bdat-fil eller en flad fil (.csv eller .txt).
  - Indsender du en flad fil, skal du oplyse, hvilken delimiter du anvender. Bemærk, at den anvendte delimiter ikke må indgå som værdi i filen. Forskerservice anbefaler brug af ¤ som delimiter, da denne sjældent indgår som værdi i datasæt.
- Variabelnavnene skal være SAS-kompatible. Det vil sige at variabelnavnene skal overholde følgende:
  - Vælg variabelnavne, der er max 32 tegn lange
  - Anvend ikke mellemrum i variabelnavnet (anvend evt. _ til angivelse af mellemrum)
  - Anvend ikke specialtegn i variabelnavnet (f.eks. ÆØÅ,.;:`*#%&/)()
- Labels kan ikke umiddelbart indlæses på Forskermaskinen, navngiv derfor dine variable med omhu.

Uploader du mere end én datafil, må du meget gerne pakke filerne i én zip-komprimeret mappe.

Det er vigtigt, at du alene indsender, data som er omfattet af den aftale, du har indgået med Forskerservice. Det er også vigtigt, at du indsender data i et format, der lever op til ovenstående krav om formater. Hvis du indsender data, der ikke lever op til aftalen eller formatet, vil Forskerservice returnere dine data og bede dig om at fremsende nye data.