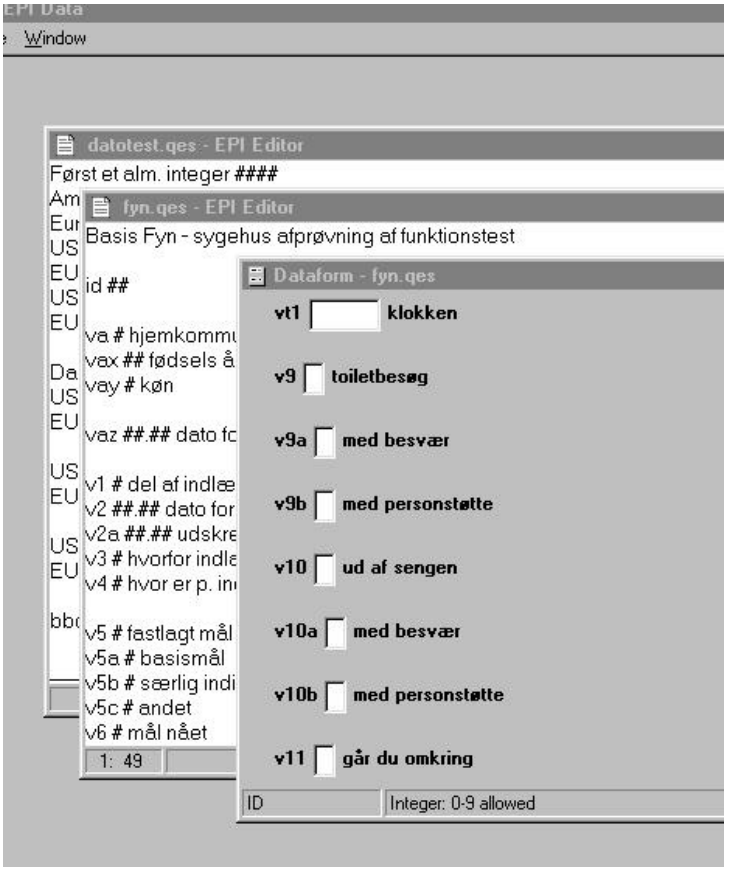
EpiData
Introduktion

Opbygning af skemaer, indtastning og fejlretning af data.

- Formulere spørgsmål i skemaer
- Omsætte spørgsmål til variable, der kan edb bearbejdes
- Indlæse data til edb bearbejdning
- Minimere fejl i data (kontrolleret indtastning)
- Rette data på en måde, der senere kan dokumenteres
- Udarbejde datadokumentation, der opfylder kravene fra udvalget vedr. videnskabelig uredelighed
- Foretage grundlæggende beskrivelser af data

- Introduktion til programmet EpiCalc 2000.
  - Confidensintervaller og statistisk testning ved direkte indtastning
  - Stikprøve størrelse beregning



Udgave 1.0. Sept. 2000

Om denne note:

1.udgave sept. 2000 (kladdeudgave pr forår 2000)

Noten er indlagt på www.epidata.dk til personligt brug. Noten må printes i sin helhed eller dele deraf til **personligt brug**. Hvis nogen ønsker at bruge hele noten eller dele af noten til undervisning, udgivelse eller distribution skal sådan brug anmeldes skriftlig til forfatteren og Copy-Dan.

Notens indhold svarer til et kursus på 4-8 timer afhængig af deltagernes forudsætning og hvor meget deltagerne har læst på forhånd. Ved 4 timer forudsættes at de første generelle dele er læst forinden.

ISBN: 87-987843-2-3 (trykt udgave)
ISBN: 87-987843-3-1 (elektronisk udgave)

EpiData - krav til computeren:
PC: Windows 95/98/2000/NT. Processor: 486, pentium eller hurtigere. Ram ingen særlige krav.
Machintosh: Mac G3 233 mghz, system 8.1 og 128 Mb RAM, samt emulator RealPC og windows95. (Muligvis også andre kombinationer).

Programmet er tilgængeligt fra internet og koster ikke noget. Se www.epidata.dk

Nogle nyttige internet sider om Stata, biostatistik eller hjælpeprogrammer:

Statistik rutiner og uddybende forklaringer af forskellige analysetyper i statistikprogrammet Stata: ("Resources to Help you Learn and Use Stata") http://www.oac.ucla.edu/training/stata/

Introduktionsnote inkl. øvelsesdatasæt til Stata på dansk: www.bola.suite.dk

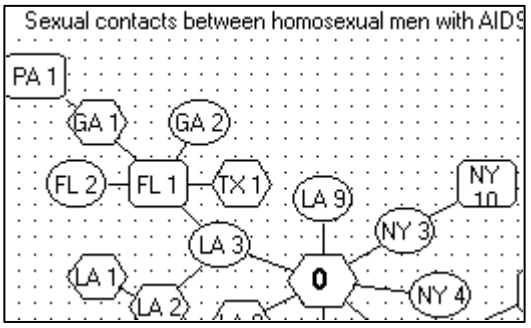
Instruktionsrutiner og forklaring af analyse af forskellige datatyper, inklusive øvelsesdata. Er baseret på EpiInfo, men principperne er generelt gyldige uanset statistikprogram:
http://www.sjsu.edu/faculty/gerstman/EpiInfo

Enkle manualer http://mkn.co.uk/help/extra/people/Brixton_Books

Indtastningsprogram EpiData (gratis) til win95/98/NT/2000: første udgave aug./sept. 2000.
Se www.epidata.dk

Hjælpeprogrammer (gratis). Blandt andet en nyttig "lommeregner" (Epicalc 2000) og et simpelt program til at tegne flowcharts og diagrammer (EpiGram). Figuren nedenfor er tegnet med EpiGram
http://www.myatt.demon.co.uk/

Uddrag af diagram tegnet med EpiGram:
(prikket baggrund findes ikke på udskrifter)



1 Forord

Der findes utallige programmer til analyse af data. Intentionen med EpiData er **ikke** at udvikle et nyt data analyse program, men at videreføre de principper for indtastning af data, som findes i programmet Epi-Info version 6[1]. Principperne er uovertruffet enkle og modsvarer alle krav til indtastning af langt de fleste data. Desværre fungerer EpiInfo ikke uden en del problemer i de nye styresystemer (Windows), specielt i forbindelse med nye netværk. Nutidens brugere har desuden vanskeligt ved at arbejde med et system baseret på "dos" principperne[2].

EpiData er udviklet ud fra erfaringer med Epi-Info version 6 med tre begrundelser:

1. Der findes ikke noget let tilgængeligt dataindtastningsværktøj, som samtidig gør det let at "kvalitetssikre" indtastning.
2. At udvikle et let tilgængeligt program der udelukkende fokuserer på indtastning. Samt at sikre financiering af udviklingen for derefter at distribuere programmet gratis.
3. At indbygge muligheder for at dokumentere indtastningsprocessen og fokusere på datakvalitet. Bl.a. med tilføjelse af labels, logisk konsistenskontrol og udskrifter.

Med EpiData har brugeren et færdigt værktøj til de faser af dataindtastning, der starter med opbygning af skemaer og leder hen til arkivering af data eller distribution af data til samarbejdspartnere. Derefter kan data analyseres i en lang række statistikprogrammer.

Programmet udvikles i et samarbejde mellem Fyns Amt, Initiativ for Ulykkesforebyggelse ved Jens Lauritsen, programmør Michael Bruus og Mark Myatt fra forlaget Brixton Health UK ("http://mkn.co.uk/help/extra/people/Brixton_Books"). Dokumentation og manualer m.v. er skrevet udenfor Fyns Amt's regi.

Financieringen af udviklingen er sikret gennem bidrag fra:

- Fyns Amt, sundhedssekretariatet http://www.fyns-amt.dk.
- Brixton Health, UK http://www.brixtonbooks.demon.co.uk
- Danish Data Archives/ERAS, Denmark http://www.dda.dk
- University of Southern Denmark, Faculty of Health - Odense. http://www.sdu.dk/indexE.html
- Valid International. London UK http://www.validinternational.org/
- London School of Hygiene & Trophical Medicine, UK http://www.lshtm.ac.uk/
- International Centre for Eye Health, UK http://www.ucl.ac.uk/ioo/ircpb.htm

Bidrag til den videre udvikling modtages i alle størrelser med tak. Midlerne vil gå til aflønning af programmør, koordineringsomkostninger ved samarbejdet med Brixton Health, udvikling af dokumenta-

1 Se fx: "http://www.cdc.gov/epo/epi/epiinfo.htm" eller "http://www.gruk.no/epi-info"

2 CDC (Center for Disease Control) i USA begyndte at udarbejde en windows udgave af Epi-Info kaldet Epi-Info 2000 før 1998. Programmet er netop frigivet i første udgave, men udviklingsstrategien fastholder ikke **det simple**. Bl.a. fylder programmet knap 40 Mb og der er endnu ikke udviklet et modul til sammenligning af dobbeltindtastede data, ligesom flere moduler fortsat er dos programmer. Brugerne opfordres til at afprøve programmet og vurdere om princippet i EpiData eller Epi-Info 2000 er mest tiltalende. EpiData svarer til MakeView og EnterData modulerne i Epi-Info 2000. Data indtastet med EpiData kan senere bruges i Epi-Info 2000, hvis man skulle ændre mening.

tion, instruktionsmateriale og præsentation af EpiData ved videnskabelige møder mv. Se nærmere i hjælpefilen om dette. Anvendelse af midlerne revideres af Dansk Selskab for Samfundsmedicin. Forslag og ønsker til fremtidige versioner kan meddeleles info@epidata.dk

Denne note giver en introduktion til opbygning af skemaer, formulering af spørgsmål i skemaer og "oversættelse" af skemaer til variable og indtastningsbilleder.

Intentionen er: At læseren efter arbejde med **alle øvelser i kronologisk rækkefølge** er introduceret til principper for gode spørgeskemaer samt principper for datadokumentation. Desuden at brugeren kan omsætte et skema til et antal variable der indtastes på en computer og og efterfølgende forklares og dokumenteres. Som en del af indtastningen kan der tilføjes regler for konsistens af data og eller kontrol af gyldige værdier m.v. Dokumentation og datafil kan til sidst samles i en "pakke" der er velegnet til arkivering og videre forsendelse til samarbejdspartnere.

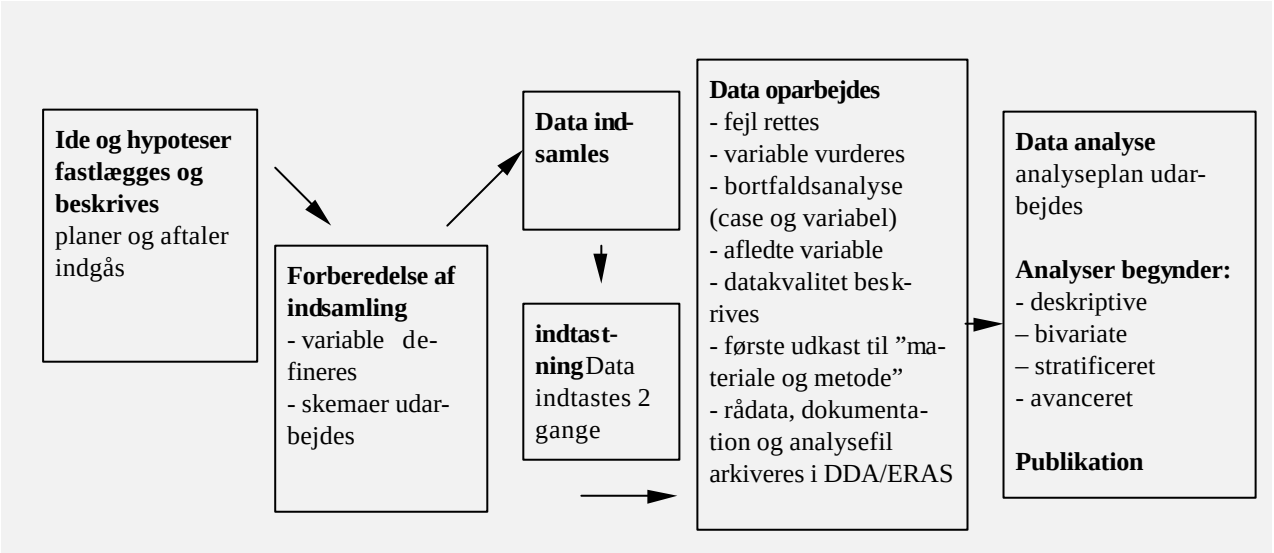
Forslag til rettelser af noten modtages gerne.

Kontakt Jens M. Lauritsen pr e-mail: **JM.Lauritsen@dadlnet.dk** eller **info@epidata.dk**

1 Oversigt – indsamling, oparbejdning og analyse af data.

Når der indsamles kvantitative data (spørgeskema, kodede interviewspørgsmål …) er det meget væsentligt, at sikre god kvalitet af de registrerede data. Både teknisk (indtastning etc.) og indholdsmæssigt. Der findes en lang række af skala'er og afprøvede enkeltspørgsmål, som belyser en bestemt problemstilling. Find disse skala'er eller formuleringer og anvend dem[1]. Et af områderne er generelle helbredsstatusmål. Et af disse anvendes i øvelserne (EuroQol 5-d) og er gengivet bagest i noten. I Danmark vil et udmærket udgangspunkt være det standardspørgeskema, som Statens Institut for Folkesundhed (tidl. Dansk Institut for Klinisk Epidemiologi "www.dike.dk") har udarbejdet. Dette giver mulighed for at sammenligne med andre undersøgelser. Ved oversættelse fra andre sprog af standard skalaer er der en række forhold, som skal sikres[2]

Den samlede proces kan opdeles i disse trin:



Bemærk, at dataoparbejdning afsluttes med at første udkast til materiale og evt. også metode skrives. Megen tid og præcision af analysearbejdet vil gå tabt hvis der ikke er god forståelse og logisk konsistens af data **inden** analysen påbegyndes. En gennemgang af hele processen er beskrevet[2]. Oplysninger om arkivering af sundhedsvidenskabelige data kan rekvireres fra Dansk Data Arkiv/Eras (www.sa.dk/ERAS).

Supplerende litteratur:

1. McDowell I and C: Measuring Health: A guide to Rating Scales and Questionnaires. Oxford University Press. 2nd ed: 1995/96. Dirksen A, Christensen E, Jørgensen T, Kampmann JP, Kjær P. Klinisk forskningsmetode, en _____: Vejledning i god forskningspraksis (www.forskraad.dk.publ/vejl_vid_praksis/prorap.htm). Jørgensen PH, Kyvik KO. Registrering og arkivering af sundhedsvidenskabelige data. Ugeskr Læger 1997; 159: 963-4. (www.dda.dk/eras). Stewart AL and Ware JE (Eds): Measuring Functioning and Well-being. The MOS study approach. Duke University Press, 1992 (ISBN o-8223-1212-3). Modelspørgeskema til undersøgelse af befolkningens sundhed og helbred.

2. Kvamme,O.J.; Mainz,J.; Helin,A.; Ribacke,M.; Olesen,F.; Hjortdahl,P. Oversettelse av spørreskjema. Et oversett metodeproblem. Nord.Med. 1998; 113: 363-6.

3. Hansen JM, Lauritsen JM. Dataindsamling Og Analyse. Ugeskr Laeger 1999 (juli)

2 Fra ide til skema.

Før de konkrete skemaer udarbejdes skal der selvfølgelig findes en ide eller et tema der skal belyses, samt tilhørende formål og delmål. Når disse er beskrevet fastlægges den overordnede datastruktur i projektet. Herunder analyseenhed, antal skemaer og de konkrete spørgsmål.

Hvad er analyseenheden ?

Hvad observeres ? En antal patienter, skoleelever, skoleklasser, en lærer og de elever der undervises, en håndboldspiller eller et håndboldhold. Dette har betydning for valg af statististisk metode. For almindeligt anvendte statistiske metoder er det en forudsætning, at analyseenhederne er indbyrdes

Eksempel:

- I et studie indgår et håndbold hold med 17 spillere, der træner sammen. Er analyseenheden så et hold eller et antal spillere, hvor nogle egenskaber er knyttet til holdet. Hvis det sidste er tilfældet bør der i analyserne tages højde for en mulig gruppeeffekt (såkaldt klyngeeffekt eller på engelsk clustereffect). Hvis det primært er det sociale sammenhold på håndboldholdet der studeres ville

- Effekten af et kirurgisk indgreb - hvis kirurgens erfaring eller individuelle dygtighed har stor betydning kan det være nødvendigt at tage højde for det i analysen. Patienterne vil så i nogen grad være knyttet sammen i klynger (grupper) afhængig af kirurgen.

Heldigvis er der i en lang række situationer ingen stærk gruppetilknytning, men vær alligevel opmærksom på skjulte klynge effekter (fx behandler effekt)

Eksempel: I en blodbank ønskes det at belyse omfanget af komplikationer ved transfusioner. Hvilken af A..D er den rigtige analyseenhed ?

Donorer: ①②③④⑤⑥⑦⑧⑨⑩ **Modtagere:** ❶❷❸❹❺❻❼❽❾❿ (Patienter)

Mulige analyseenheder:

A.En transfusion: ❷① ❷① ❼① ❼① (eksempler)

B.Et transfusionsdøgn: ❷①① ❼①① ❷⑧⑨⑩ (den samlede mængde blod,
❷②②② ❷③③③ ❷④③③ givet til hver patient i et døgn)

C.En patient og samtlige modtagne portioner blod:
❷①②③④④⑤⑥⑦⑧⑨⑩ (Patienten indeholder potentielt antigener fra alle 10 donorer)

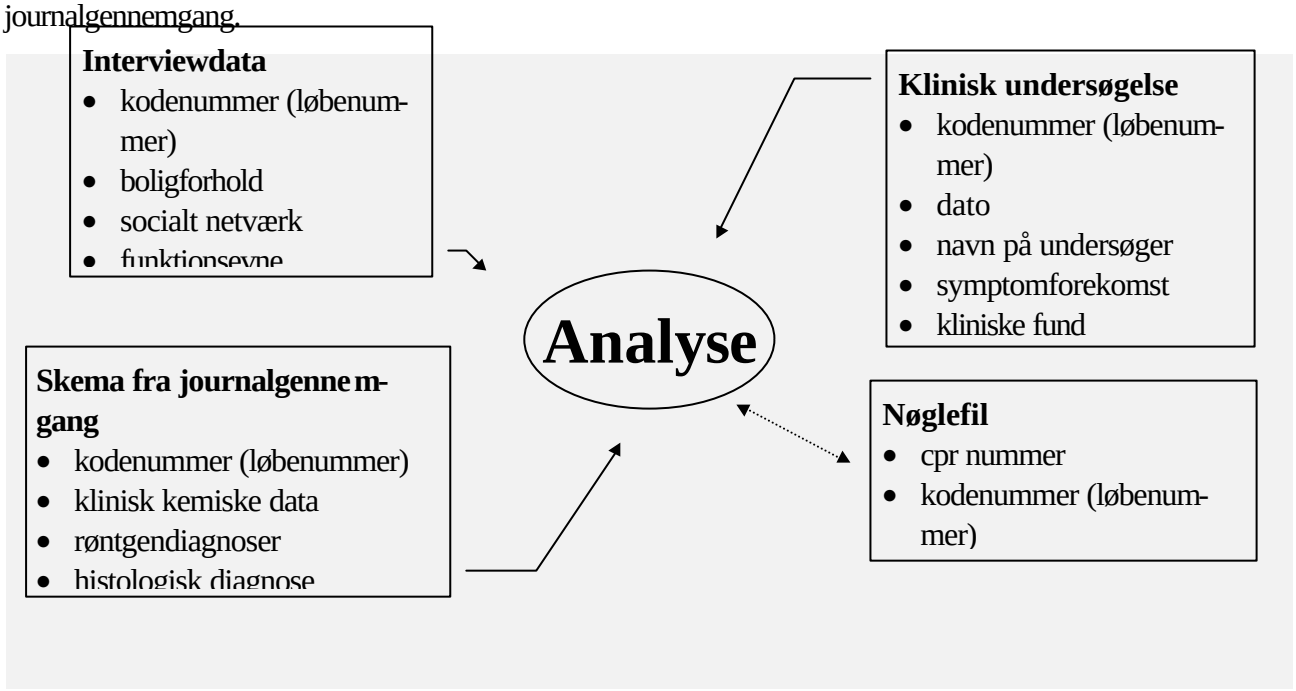
D.En donor og samtlige patienter, der har modtaget blod fra pågældende:
⑥❶❷❸❹❺❻❼❽❾❿ (Alle patienterne indeholder potentielt antigener fra donoren)

Svaret må selvfølgelig være: Det kommer an på arten af komplikationer der har interesse.

Hvor mange skemaer ?

Enkle undersøgelser kan gennemføres ud fra ét skema, fx et spørgeskema med 10 spørgsmål til alle respondenter (personer, som har svaret), mens andre har flere typer skemaer.

Eksempel:
En undersøgelse indeholder: interviewdata, klinisk undersøgelse og et skema til journalgennemgang.



Der opbygges et skema og en indtastningsrutine til hvert skema. Senere kobles alle skemaer sammen i data **Analysen** ud fra et fælles id nummer (identitets eller anonymiserings nummer).

En særlig fil med både identitetsnummer og cpr nummer gemmes særskilt og aflåst efter de regler, som datatilsynet har givet.

Indhold i det enkelte skema.

For hvert af de ønskede skemaer udarbejdes et første udkast med de forskellige emner, der skal indgå. Det er som regel en god ide, at starte med de spørgsmål, der specifikt vedrører projekts emne. Fødselsdato, køn etc kan indsættes senere i skemaet.

Opdel hvert skema i forskellige sektioner og tænk på om rækkefølgen er både logisk og acceptabel for de der skal svare. Det er nyttigt at anvende filterspørgsmål, især ved interview.

Eksempel: Filterspørgsmål

10	Har De været i arbejde siden 1. januar ?	☐[1] Ja ☐[2] Nej - fortsæt i spørgsmål 28

De konkrete formuleringer af enkelte spørgsmål tages fra standardiserede spørgsmål, en valideret skala eller en diagnostisk kode (fx ICD-10 kriterier), se side 5. Pas på med "private" oversættelser og typografiske fejl, som er meget lette at overse. Det er en fordel at skitsere de grafer og tabeller, der skal findes i de planlagte publikationer. Hvis du ikke kan anvise hvilke grafer mm. der skal bru-

ges til sidst, ved du heller ikke hvilke data der skal indsamles.

Alle formuleringer nedenfor er fra DIKE's sygdoms- sundhedsundersøgelse fra 1994. (sprøgsmålene indgik i en anden rækkefølge)

Øvelse 1 Præcision af formuleringer.
Sammenlign formuleringerne i spørgsmål 1 og 2. Forventer Du samme svarfordeling i de to formuleringer?

Helbred	
1. Hvorledes vil du vurdere din nuværende helbredstilstand i almindelighed?	1☐ Virkelig god 2☐ God 3☐ Nogenlunde 4☐ Dårlig 5☐ Meget dårlig
2. Hvordan synes du dit helbred er alt i alt ?	1☐ Fremragende 2☐ Vældig Godt 3☐ Godt 4☐ Mindre Godt 5☐ Dårligt

To andre generelle spørgsmål om helbred og dagligdag er:

3. Føler du dig frisk nok til at gennemføre det, som du har lyst til at gøre ?	1☐ Ja, for det meste 2☐ Ja, af og til 3☐ Nej, sjældent
4. Er du glad og tilfreds med din tilværelse ?	1☐ Ja, for det meste 2☐ Ja, af og til 3☐ Nej, sjældent

Svarmuligheder

Det er vigtigt at opbygge hvert spørgsmål så alle personer kan finde ét svar (såkaldt exhaustivt eller udtømmende krav), og samtidig at du har besluttet om der kun må være et svar i hvert spørgsmål (kategorier er gensidigt udelukkende eller exclusive). Endelig om svarmulighederne skal være kategoriale (mand, kvinde), ordinale eller rangordnede (for det meste, af og til, sjældent) eller kontinuerte (hvor mange meter ?). Datoer er med **EpiData** lette at indtaste og regne på (fx alder ved et bestemt ambulatorie besøg). Tekster og forklaringer kan give gode uddybende forklaringer, som ikke nødvendigvis skal tastes ind. I indtastningen kan man nøjes med at angive om der findes en tekst eller ej.

Svarmulighederne omsættes senere i variable med de samme typer. Nogle spørgsmål kodes i én variabel (fx køn) andre gange må der flere til (hvilke af følgende aviser læser du ?).

Indsamling af data.

Det er nødvendigt at etablere en rutine til at sikre, at alle skemaer indsamles og en opsamlings/rykkerprocedure hvis der mangler information. Dette kan være en omfattende arbejdsopgave,

hvor grundighed betaler sig. Det skal være let at overskue (fx ved at udskrive lister) hvem der har svaret eller hvilke journaler der er gennemgået på et givet tidspunkt.

2 Fra skema til indtastning.

For hver oplysning der indsamles besluttes, hvordan oplysningen kodes som variable i dataanalysen. En variabel er den mindste informationsenhed i et skema. Variable kaldes også parametre, kovariater, items mm afhængig af faglig baggrund og tradition hos den der udfører undersøgelsen.

Antal og typer af variable ?.

En oplysning kan fx være et spørgsmål i et spørgeskema, en blodprøveværdi eller en histologisk diagnose for en vævsprøve. Nogle gange kodes et spørgsmål i én variabel (fx køn) andre gange må der flere til (hvilke aviser læser du ?).

Variable er enten grupperede (d.v.s. et afgrænset antal svarmuligheder ja/nej/ved ikke), rangordnede (grupperet, men med en bestemt rækkefølge stor/større/størst), kontinuerte (numeriske), fritekst strenge ("Han tog hans hat og gik hans vej"), logiske (Sand/Falsk - som regel kodet som J/N, Y/N eller 1/0) eller datoer (datoen indtastes).

Variable er af forskellig type. De er alle vigtige for et samlet reproducerbart resultat. Udover de emnemæssigt mest relevante (de der afspejler formålet) er det som hovedregel nødvendigt med nogle baggrundsvariable (alder, køn ....), nogle kontrolvariable (til såkaldt confounder kontrol), styringsvariable (dataindsamlings processen) og kvalitetsvariable. Se nærmere nedenfor.

Gruppering af data - NEJ !!

Det er **meget vigtigt** at **INGEN** information grupperes/summeres på tidspunktet for dataopsamling eller dataindtastning. Det gøres meget lettere og bedre under analysearbejdet.
Indtast fødselsdato - ikke alder, vægt i kg ikke grupperet vægt, sygedage i antal dage - ikke over/under 7 dage. Indtast hele datoen for besøg i et ambulatorie, ikke kun årstal. Hvis du bruger en tidskode (1=start 2= 3mdr 3=6 mdr ...), så tast også datoen for en sådan oplysning ind. Det kan være der er årstidsvariation eller at det senere viser sig, at der var problemer med kalibrering i en bestemt kalender periode.

Ingen datareduktion under indsamling eller indtastning af data !!!!

Styringsvariable.

Et antal styringsvariable er som regel nødvendige: dato for interview, hvornår er skema retuneret, hvem har kodet tekstinformation, hvem har interviewet .....

Kvalitet af en oplysning - "kontrolvariable".

Ofte varierer oplysningers kvalitet med andre kendetegn. Fx må det forventes, at en røntgen diagnose stillet af en speciallæge er mere præcis end hvis diagnosen stilles af yngste reservelæge. Et andet eksempel kunne være dato for en analyse, der er kendt for at have årstidsafhængig variation (Fx holdbarhed i forhold til transport temperatur).

For at bruge præcisionen af en oplysning i analysen kan der inkluderes en eller flere variabler, som angiver usikkerheden på hovedvariablen (diagnostikers erfaring eller prøvedato). Senere i dataanalysen kan det undersøges om der er systematiske afvigelser ud fra undergrupperne i disse "kontrol-

variable".

Navngivning af variable.

De forskellige variable gives navne. Det er nødvendigt for at sikre, at det er de rigtige tabeller der tolkes.

Det er en smagssag om variable skal have numre (v1,v2....v129) eller betegnelser, som antyder indholdet (alder, gender, ACTH, ...). Men i analyser mv er det hurtigere at skrive fx "v1" end "hoejde78". Derfor foreslår jeg at man bruger nummerlignende betegnelser. Navnet skal være entydigt og kan med fordel referere til de spørgsmål, der er i registrerings- /interview- /spørgeskemaer (Fx s1,s2 ... for spørgsmål og L1,L2 ... for laboratoriedata). Anvend højst 8 bogstaver/tal i variabelnavne, **aldrig æøå**. Første tegn i et navn må ikke være et tal **1navn** er ulovligt, mens **navn1** er ok.

Forskel mellem uoplyst og irrelevant.

Det er vanskeligt at opstille en fast regel for hvordan dette skal håndteres under indtastningen og i skemaer.

- **Uoplyst:** Uoplyst betyder at oplysningen **ikke** er tilvejebragt. Om det skyldes svar personens manglende lyst til at give svaret eller om det ikke kan fremskaffes er ligegyldigt. Svaret findes ikke tilgængeligt for analysen.
- **Irrelevant:** Irrelevant er et svar, som ikke kan og ikke skal besvares. Fx er antal gennemførte graviditeter uden betydning for mænd (i biologisk somatisk forstand).

Under indtastningen kan det være en fordel at alle uoplyste værdier blot overspringes, dvs tastes med en blank (ingen) værdi. Ved et filterspørgsmål springes de irrelevante over, og der indsættes en irrelevant værdi (fx 8) i de spørgsmål der springes over. Den irrelevante værdi kan enten indsættes under indtastningen (automatisk) eller det kan ske i oparbejdningen..

Hvordan skal der kodes hvis et filterspørgsmål er uoplyst ?

Eksempel: Filterspørgsmål

10	Har De været i arbejde siden 1. januar ?	☐[1] Ja ☐[2] Nej - fortsæt i spørgsmål 28

Der kunne kodes uoplyst for spørgsmål 10-27 for alle, som ikke har besvaret ovennævnte spørgsmål, men der kunne også kodes irrelevant. Hvis der kodes irrelevant svarer det til at uoplyst tolkes som et nej. Beslutningen må dokumenteres i undersøgelsen.

En kodebog

En kodebog viser hvilke variable der findes, hvordan de enkelte variable kodes og hvilken kontrol der udføres under indtastningen. (se næste figur) I kodebogen er hver oplysning omsat til en eller flere variable, der hver er defineret nøje. Hvilken datatype (kontinuert, grupperet, åben tekst ...), hvilke svarkategorier findes, hvilken talkode skal anvendes, hvis variablen er uoplyst etc. Bredden er det antal cifre eller antal bogstaver der skal bruges, fx 3 ved 999 og 2 ved 99. For kontinuerte variable skal minimum og maksimum anføres. Desuden skal nøglevariable besluttes, d.v.s. variable som skal belyses for hver person.

Der kan som nævnt ovenfor være brug for to typer "uoplyst". Den ene type gælder når en ønsket oplysning ikke findes (*personen svarede ikke, blodprøvesvaret er endnu ikke ankommet etc.*), mens den anden type er "irrelevant spørgsmål", der opstår ved et "filterspørgsmål".

Alt dette vil fremgå af en kodebog, som **EpiData** kan udskrive når indtastningen er defineret.

Endelig udarbejdelse af skemaer

Når analysevariabler er defineret kan skemaer finpudses og dataindsamlingen begynde. "Pilot-afprøvning" foretages på en gruppe der svarer til den endelige modtagergruppe. Efter pilotfasen gennemgås hvert spørgsmål - spørges der kun om én dimension (éet spørgsmål), er alle mulige svar

Simpel kodebog for et skema

Variabel navn	Variabel betydning	Værdier	Bredde	Kommentar
Idnr	løbenummer (kodenummer) for personen	001-999	3	Case: 0-500, refområde: 501-900
S11n1	Kilometer på cykel/uge	000-9000 9999* .*	4	
S11n2	Kilometer i bil/uge	000-9000 9999* .*	4	
S11n3	Kilometer i bus/uge	000-9000 9999* .*	4	
S12	Boligtype	1 Lejlighed 2 Parcelhus 3 Rækkehus 4 Værelse 5 Andet 9 Uoplyst*	1	Defineret ud fra SFI/DIKE undersøgelsen fra 1990 (ref 2)
s13txt	Beskrivelse af trafikopfattelse	Indskrives som tekst (max 60 tegn)	60	

2 Dokumentation.

Dokumentation består af de oplysninger, som i princippet er nødvendige for at gentage en given undersøgelse en anden gang. Det vil sige at det både indeholder projektplaner, beskrivelse af beslutninger i processen, godkendelser fra datatilsyn og samarbejdspartnere, eventuel godkendelse fra videnskabsetisk komite og en række tekniske dokumenter om selve de indtastede data. Der er ingen formelle lovbestemte krav til dokumentation, men i praksis viser erfaringen at det er meget nyttigt at dokumentere processen. Ingen kan huske hvad der er besluttet i en konkret fase af et projekt nogle år senere. Med "datadokumentation" menes den del af dokumentationen, som er nødvendig for at kunne anvende et givet indsamlet datasæt til analyse og beskrivelse.

I forbindelse med sundhedsvidenskabelig forskning er der udarbejdet vejledende retningslinier fra "udvalget vedr. videnskabelig uredelighed" (se side 5). Reglerne er retningslinier og - desværre - udformede i generelle henstillinger. Fx skrives - kvalitetskontrol - uden at specificere dette nærmere.

1. Mulighed for at finde tilbage til originalmateriale.

Det skal være muligt for enhver oplysning (fx et punkt i en figur) at finde tilbage til originalmaterialet. Det vil sige, at der skal være knyttet et id nummer til alle observationer. Id nummeret følger med i alle udgaver af data og er **entydigt** forbundet med de originale observationer. Originalmaterialet skal opbevares i 10 år. Originalmaterialet omfatter notater (også håndskrevne retteblade mm), spørgeskemaer, analyseskemaer, ……

2. Kvalitetskontrol, dokumentation og arkivering.

Samtidig med ovenstående proces dokumenteres antal fundne fejl under indtastning mv. og de konsekvenser det har haft. Data arkiveres sammen med projektplaner, kopi af anvendte skemaer og den udarbejdede dokumentation. For større projekter med fordel i Dansk Data Arkiv/ERAS. Ved arkivering i ERAS bevarer den person der har afleveret data den fulde kontrol over hvem der må få data udleveret igen. Kontakt ERAS for nærmere oplysning: "http://www.dda.dk/eras" eller

Udarbejdelse af datadokumentation.

Datadokumenation er en løbende proces. Det består helt enkelt i en samling af noter (fejlrutiner, rykkerprocedurer, håndskrevne fejllister m.v.) og kopi af de værktøjer (indtastningsfiler m.v.) der er anvendt under indtastning og fejlfinding. Endelig indgår kopi af alle filer der er anvendt undervejs og dokumentationen afsluttes med en liste over filer der er tilgængelige i dataanalysen.

Det vigtigste formål med datadokumentationen er at redegøre for indtastningsprincipper, ændringer foretaget som konsekvens af kvalitetskontrol og at begrunde hvor mange personer, der indgår i den konkrete undersøgelse. Herunder vurdere omfanget af uoplyst og beskrive særlige forhold som der skal tages hensyn til under analyse af data.

Omfanget af dokumentation skal svare til ambitionsniveauet med studiet. For flerårige studier bør der findes omfattende dokumentation. For mindre evalueringer eller kortlægninger fylder det måske kun

Som en del af dokumentationen er det en god ide at arbejde med et dokument, hvor beslutninger løbende indskrives.

Eksempel: Beslutninger under indtastning og oparbejdning af data fra et landbrugsprojekt:

For personen 82809 er det en fejl, at v7a2x er kodet 1, dette er rettet til 0.

Den ene bedrift og tre personer uden information udelukkes:

```
select if (bnr <> 841 and v1 <> 82309 and v1 <> 8299 and v1 <> 2599).
```

Grunden til at de var med i filen var, at de som besøgende har haft en materielskade ulykke.

Disse personer er kodet som bystander og givet 1 time for alle risikotider:

V1	V3	TIMER2	RISKÅR1	RISKÅR2	RISKÅR3	RISKÅR4	RISKTOT	V7A2	V7A2X
1102	1	1	1	1	1	1	1	2	2
2299	1	1	1	1	1	1	1	1	1
3999	1	1	1	1	1	1	1	1	1
5099	1	1	1	1	1	1	1	1	1
7203	1	1	1	1	1	1	1	1	1
7304	1	1	1	1	1	1	1	1	0

Det samlede antal ulykker er sat til ovenstående i variablen v7a2 og v7a2x, men er sat til 0 i variablerne v7a2b og v7a2bx.

Brugstype

Denne var uoplyst for brug med nr: 16, 41,252,260,293

Disse kodes indtil videre som ”andet” i brugstype og uoplyst i rentabilitet.

På denne måde er der ikke tvivl om hvad der besluttet. Selve teksten er selvfølgelig noget kryptisk for udenfor stående, men er med til at opfylde kravet "data skal kunne følges fra oprindeligt skema til endelig analysefil". Noterne skal opfattes som et arbejdsdokument.

Nedenfor vises endnu et eksempel med en samlet variabelliste. Eksemplet er tilpasset ud fra en konkret arbejdsmiljøundersøgelse af psykisk helbred og stress indenfor politiet (ref: Bjarne Ibsen. Politiets Psykiske Arbejdsmiljø. Arbejdsmiljøfondet). Undersøgelsen blev gennemført som en kombination af en landsdækkende tværsnitsundersøgelse og en opfølgende undersøgelse i Fyns Amt, hvor der skulle udfyldes skemaer 5 gange på et år. Formålet var at nedbringe de psykiske konsekvenser af politiarbejde i forbindelse med voldsomme hændelser (fysisk vold, trusler mod politiet, oplyse forældre om børn's død, skudepisoder mv.).

Ind imellem teksten kommer nogle øvelser, hvor du skal tage beslutninger.

Eksempel: Datadokumentation for et konkret projekt.

Personer:
For at vide hvor mange personer der har været ansat er der udarbejdet en slags råfil med personoplysninger. Grundlaget for denne var den registrering, som har været anvendt til at udsende skemaer efter. Fra filen er anvendt idnr, start og slutdato, samt beregnet varighed af ansættelse. Filen havde desuden en variabel charge (er personen overordnet ?), aktiv (personen skulle modtage skema nr 5) og en variabel der angav køn på basis af navnene (en række personer var benævnt med initialer fx P.K.Hansen og har fået værdien 2 for køn, de fleste af disse vil være mænd). Kodningen er sket ud fra filen: 'Find datoer.do' (vedlagt). ***OBS***. *Af anonymitetshensyn har navne og id nr altid været adskilt på tidspunkter, hvor skemaet er blevet håndteret.*

Skemaer er modtaget fra Esbjerg Centralsygehus udelukkende med id nr på. Indtastningen er sket fortløbende med kontrol af lovlige værdier m.v. ved hjælp af indtastningssystemet EpiData. Al indtastning er foretaget af NN.

Alle skemaer er indtastet en gang. De 10 vigtigste variable er kontrolleret mod skemaernes indhold efter indtastning.

Besvarelse af de udsendte skemaer:
Det skal nu besluttes hvilken del af skemaerne der skal anvendes ved analyserne. Det er vanskeligt at afgøre hvilke skemaer, der skal indgå i analyserne. Valget indebærer en afvejning af komplethed af skemaer i forhold til besvarelsesprocent. For at gøre analyserne gennemskuelige vil jeg foreslå at basere analyserne på alle der har været ansat i 10 mdr. Det er denne gruppe, som har været længst i kontakt med projektet som var et interventionsprojekt. Desuden må det også gælde, at mindst første og sidste skema er afleveret.

Analyserne kunne baseres på følgende subpopulationer:

Skemaer:	Besvarelsesprocent:		ansat mindst 3 mdr		ansat mindst 10 mdr	
Første	554/674 =	82 %.	534/641 =	83 %	423/498 =	85 %
Sidste	487/674 =	72 %	468/641 =	73 %	381/498 =	77 %
Første og sidste	448/674 =	67 %	435/641 =	68 %	359/498 =	72 %
Første og sidste + ét mere	446/674 =	66 %	433/641 =	68 %	357/498 =	72 %
Første og sidste + to mere	424/674 =	63 %	413/641 =	64 %	343/498 =	69 %
Alle fem	355/674 =	53 %	346/641 =	54 %	290/498 =	54 %
Procentandel af de ansatte:		100 %		95 %		74 %

Øvelse 2 Beslut hvilken delmængde af skemaerne du vil bruge til analysen.

tabellen og tag en beslutning. Uanset hvilken beslutning du tager bliver der et bortfald. Enten et skemabortfald eller et personbortfald. Udform derefter et afsnit til metodeafsnit, der beskriver dit materiale og bortfaldets størrelse.

Efter en yderligere vurdering af omfanget af uoplyste udformes en tabel med oversigt over bortfald og variable i analysen. Afhængig af de valg der blev truffet ovenfor kunne tabellen se sådan ud:

Materialet for data fra Fyn 1995 bliver derfor:

	Personer I alt	Antal skemaer afleveret	Antal hændelser rapporteret i alt i foregående 3 mdr.	Uoplyst om hændelser
Samtlige ansatte personer	674			
Ingen skemaer afleveret	120			
Samtlige indtastede skemaer	554	1577	240	8
Mere end 19 uoplyste spørgsmål		42	21	6
Herefter tilgængelig	554	1535	219	2
Mindre end 3 mdr's ansættelse i 1995	9	19	2	
Indgår i analysen	545	1516	217	2

Udover de indsamlede data fra Fyn beskrevet ovenfor foreligger et landsmateriale fra 1992. Der er herefter tre datafiler til rådighed for alle de afsluttende analyser:

aler.dta'	Anvendes til dannelse af skalaer alene	1516 records	Variable se bilag
'Fyn 1995.dta'	Hver person én record, dvs analyser på tværs af året.	545 records	Variable se skema nedenfor
'Hele landet.dta'	Analyser for hele landet, bl.a. afsluttende ph.d. analyser	9035 records	Variable se skema nedenfor

En record eller observation svarer til en person (et skema).

Analyseprincip for sammenligninger i løbet af året. Herunder vurdering af interventionens effekt.

Da projektets dataindsamling og analyse blev planlagt var det hensigten, at der skulle indsamles de samme data (forløbs data) ved starten af projektet og ved afslutning af hvert af de 4 kvartaler i løbet af 1995. Denne indsamlingsmetode måtte opgives, fordi der var stor modstand mod at besvare skemaet fra nogle afdelinger. Følgende variable er tilgængelige. Andre problemstillinger kan analyseres fra de oprindelige data efter yderligere oparbejdning.

Indhold	**Lands-projektet**	**Fyn** startskema	1. kvartal	2.kvartal	3.kvartal	4.kvartal
Arbejde på fyn	sted					
Alder (agegrp er grupperet)	age agegrp			age agegrp		
Køn	gender			gender		
Antal skemaer afleveret i alt				antal		
Skema afleveret i dette kvartal ?		in_0	in_1	in_2	in_3	in_4
Ordens /kriminalpoliti	besk	besk_0				besk_4
Leder (charge)	charge	charge				
Anciennitet i politiet	ancp					
Anciennitet i nuværende position	anc	anc				
Skiftet afdeling siden 1.1.95						anc_4
Skiftende arbejdstider	skift	skift_0				skift_4
Bureaukratisk sagsgang	d3a					
For stor arbejdsmængde	arbmgd	arbmgd				
Støtte og opmuntring fra leder	e10	e10				
Støtte og opmuntring fra kolleger	e11	e11				
Det sociale miljø på arbejdspladsen	i1c	i1c				
Mulighed for råd og hjælp fra leder	e8	e8				
Overvejet at finde andet arbejde	c1a	c1a				
Social kontakt med andre	social	social				
Deltagelse i aktiviteter	aktiv	aktiv				
Generelt selvvurderet helbred	j13a	j13a_0	j13a_1	j13a_2	j13a_3	j13a_4
Glad og tilfreds med tilværelsen	j14a	j14a_0	j14a_1	j14a_2	j14a_3	j14a_4
Tilfreds med arbejdsforhold - alt i alt	L7	L7_0	L7_1	L7_2	L7_3	L7_4
Mindst én uges sygefravær sidste år	j5a					
Mindst 3 dages sygefravær sidste 2 mdr	j4a1					
Mindst 3 dages sygefravær sidste 3 mdr		j4a_0	j4a_1	j4a_2	j4a_3	j4a_4
Stadig påvirket af hændelse fra tidligere	b5y	b5y				
Angivet voldsom hændelse for foregående 12 måneder	b5t	b5t_0				b5t_4
Udsat for voldsom hændelse foregående	b5	b5_0	b5_1	b5_2	b5_3	b5_4
Hændelserne (-en) påvirkede privatlivet	b5f	b5f_0	b5f_1	b5f_2	b5f_3	b5f_4
Er du stresset ?	j8k	j8k_0	j8k_1	j8k_2	j8k_3	j8k_4
Manglende lyst til at gå på arbejde pga	c5d8	c5d8_0	c5d8_1	c5d8_2	c5d8_3	c5d8_4
Skala: Graden af gener fra hændelse	gener	gener_0	gener_1	gener_2	gener_3	gener_4
Skala: **Fysisk** velbefindende (4 item)	velbf4	velbf4_0	velbf4_1	velbf4_2	velbf4_3	velbf4_4
Skala: **Psykisk** velbefindende (4 item)	velpf4	velpf4_0	velpf4_1	velpf4_2	velpf4_3	velpf4_4
Medvirket i intervention						delt

Variable markeret med grå er kun oplyst for personer, der har haft hændelser i de foregående tre

Variablene (**gener** og **c5d8)** samt (**gener_0** og **c5d8_0)** refererer til besvarelsen af spørgsmålet om hændelser i de foregående 12 mdr. I de resterende kvartaler i Fyns undersøgelsen refererer det til hændelser fra de foregående 3 måneder i hvert af kvartalerne.