Att strukturera och organisera forskningsdata

Robert Lundqvist, Statistiker, Region Norrbotten

Anna Lindam, Statistiker, Region Jämtland Härjedalen

robert.lundqvist@norrbotten.se anna.lindam@regionjh.se

Kliniska Studier Sverige Forum Norr

Översikt

- Strukturera din forskning
- Organisera dina filer
- Ordning och reda bland dina variabler
- Tips för inmatning av variabler

Spårbarhet

Speciellt viktigt då forskning alltid omfattar:

- Data, ofta stora mängder data
- Instruktioner och SOP:ar
- Resultat från analyser
- Resultat från om-analyser
- Log-filer
- Grafer och bilder
- Referensartiklar
- Utkast

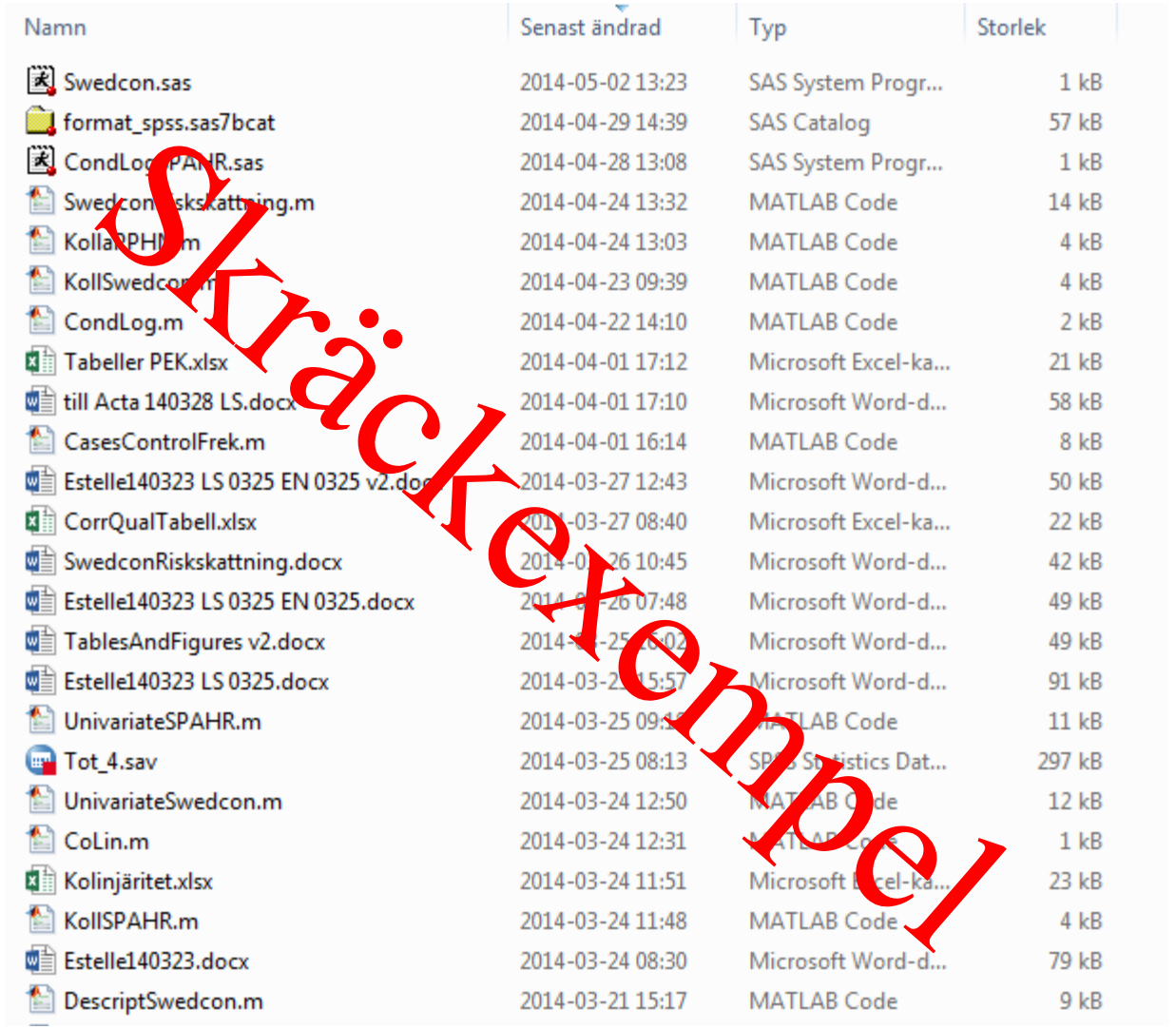
Struktur

Kliniska Studier Sverige Forum Norr

Tydlig arkitektur i mappstrukturen

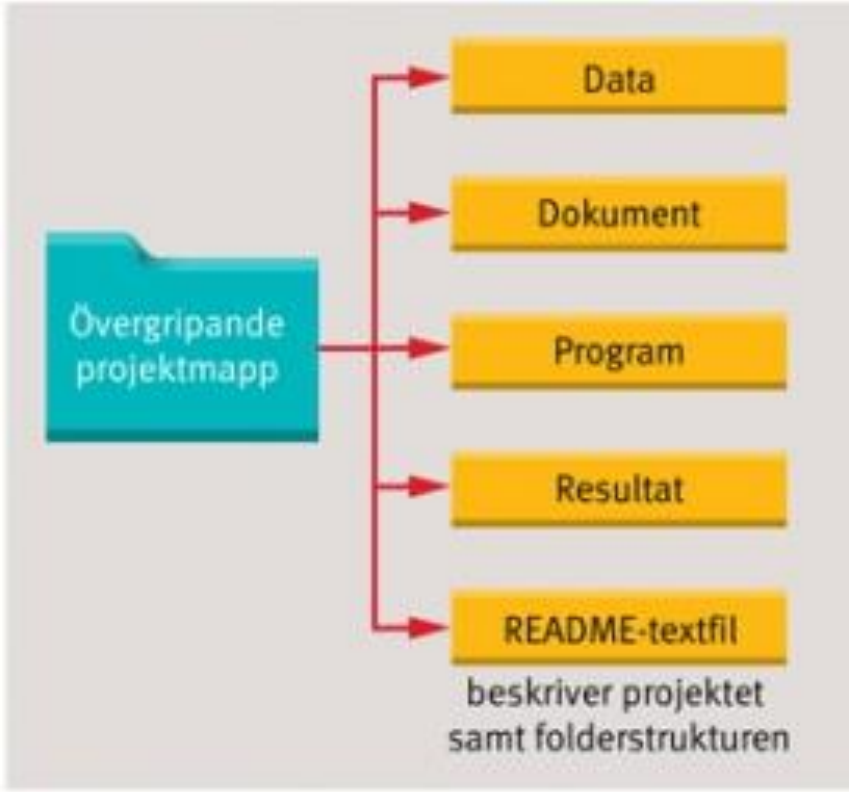
- Varför?
  - Lättare att hitta – både för dig och kollegor
  - Ordning och reda
  - Lättare hålla koll på dina data
  - Lättare att hålla koll på versioner om de ligger på samma ställen

Datastruktur



Namn	Senast ändrad	Typ	Storlek
Swedcon.sas	2014-05-02 13:23	SAS System Progr...	1 kB
format_spss.sas7bcat	2014-04-29 14:39	SAS Catalog	57 kB
CondLogSPAHR.sas	2014-04-28 13:08	SAS System Progr...	1 kB
Swedconriskskattning.m	2014-04-24 13:32	MATLAB Code	14 kB
KollaSPHR.m	2014-04-24 13:03	MATLAB Code	4 kB
KollSwedcon.m	2014-04-23 09:39	MATLAB Code	4 kB
CondLog.m	2014-04-22 14:10	MATLAB Code	2 kB
Tabeller PEK.xlsx	2014-04-01 17:12	Microsoft Excel-ka...	21 kB
till Acta 140328 LS.docx	2014-04-01 17:10	Microsoft Word-d...	58 kB
CasesControlFrek.m	2014-04-01 16:14	MATLAB Code	8 kB
Estelle140323 LS 0325 EN 0325 v2.doc	2014-03-27 12:43	Microsoft Word-d...	50 kB
CorrQualTabell.xlsx	2014-03-27 08:40	Microsoft Excel-ka...	22 kB
SwedconRiskskattning.docx	2014-03-26 10:45	Microsoft Word-d...	42 kB
Estelle140323 LS 0325 EN 0325.docx	2014-03-26 07:48	Microsoft Word-d...	49 kB
TablesAndFigures v2.docx	2014-03-25 15:02	Microsoft Word-d...	49 kB
Estelle140323 LS 0325.docx	2014-03-25 15:57	Microsoft Word-d...	91 kB
UnivariateSPAHR.m	2014-03-25 09:1[illegible]	MATLAB Code	11 kB
Tot_4.sav	2014-03-25 08:13	SPSS Statistics Dat...	297 kB
UnivariateSwedcon.m	2014-03-24 12:50	MATLAB Code	12 kB
CoLin.m	2014-03-24 12:31	MATLAB Code	1 kB
Kolinjäritet.xlsx	2014-03-24 11:51	Microsoft Excel-ka...	23 kB
KollSPAHR.m	2014-03-24 11:48	MATLAB Code	4 kB
Estelle140323.docx	2014-03-24 08:30	Microsoft Word-d...	79 kB
DescriptSwedcon.m	2014-03-21 15:17	MATLAB Code	9 kB

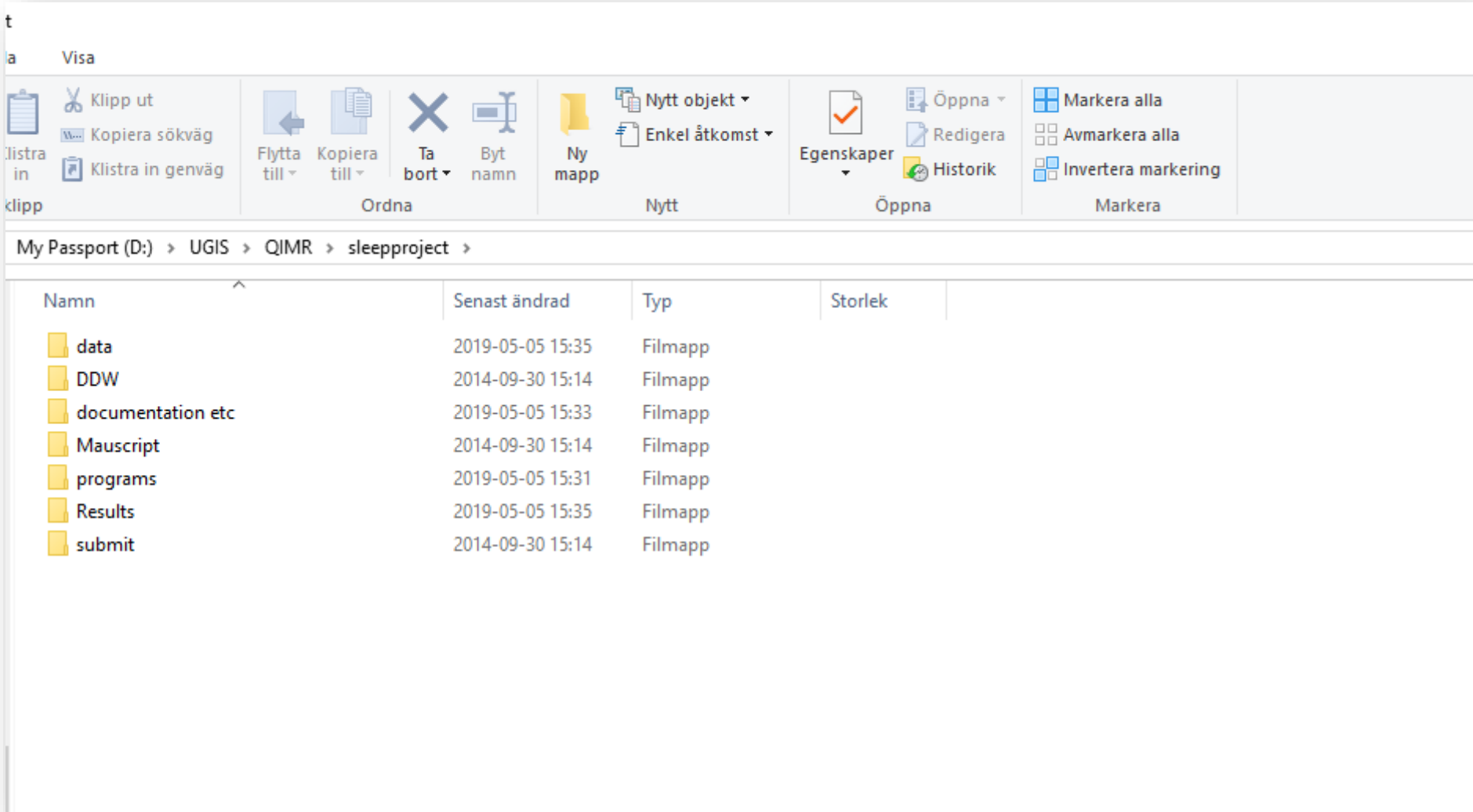
Mappstruktur



Figur 2. Rekommenderad mappstruktur.

Eloranta S. et al, Att strukturera och dokumentera forskningsprojekt, Läkartidningen 2013;110;8

Kliniska Studier Sverige Forum Norr


t
a
Visa
Klipp ut
Kopiera sökväg
Klistra in genväg
listra in
klipp
Flytta till
Kopiera till
Ta bort
Byt namn
Ordna
Ny mapp
Nytt objekt
Enkel åtkomst
Nytt
Egenskaper
Öppna
Redigera
Historik
Öppna
Markera alla
Avmarkera alla
Invertera markering
Markera
My Passport (D:) › UGIS › QIMR › sleepproject ›
Namn
Senast ändrad
Typ
Storlek
data 2019-05-05 15:35 Filmapp
DDW 2014-09-30 15:14 Filmapp
documentation etc 2019-05-05 15:33 Filmapp
Mauscript 2014-09-30 15:14 Filmapp
programs 2019-05-05 15:31 Filmapp
Results 2019-05-05 15:35 Filmapp
submit 2014-09-30 15:14 Filmapp

Filer

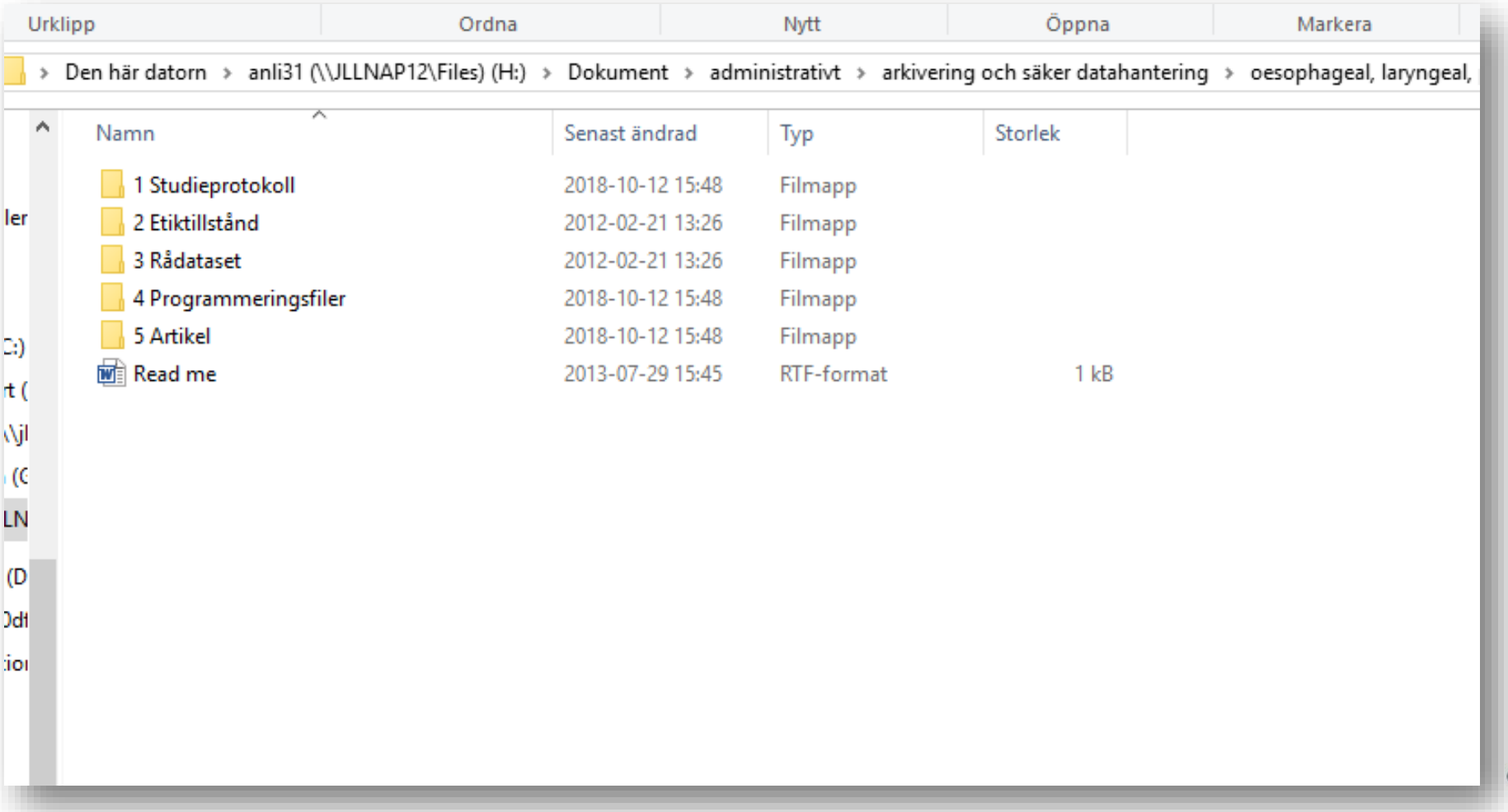
- Om du jobbar med uppdaterade datafiler, sätt namn på senaste med dagens datum
- Output-filer
    - I SPSS, innehåller text som kan kopieras över till syntaxfil och köras där, alltså ett slags arkiv
    - Spara även outputfiler med datum
    - Kom ihåg att starta varje SPSS-pass med att öppna senaste outputfilen, nya bearbetningar hamnar längst ner
    - Städa eventuellt bort helt överflödigt innehåll
- Om du använder syntax, sätt namn med dagens datum även på syntaxfil

Arkivering

- All forskning ska sparas i 10 år eller mer, detaljer kan finnas i lokal dokumenthanteringsplan, kan skilja mellan olika slags studier
- Varför?
  - Studier ska kunna replikeras
  - Ny studie med samma struktur? Då behövs den tidigare studien.
  - Inte ovanligt att forskare vill göra uppföljande studier långt efter grundstudien
- Vissa uppgifter är offentliga handling
  - T.ex. ansökningar, beslut etikprövningsnämnden, avtal
- Kolla med arkivarie

Vad bör arkiveras i en studie?

- Dokument
  - Etiska ansökan + beslut
  - Studie/forskningsplan
  - Ev. loggbok
- Program (syntaxfil) och/eller loggbok
- Data
- Manuskript



Kliniska Studier Sverige Forum Norr

Read me

- The project "The risk of cancer of the esophagus and larynx after gastric resection for benign disease" resulted in 3 publications. As they share the same database, methods and parts of the programs, they are archived jointly.

Struktur datafil

2014-02-15 **8 patienter**

Patient	personnr	hö höft							vä höft						
17. NN1	YYMMDD-NNNN (18år)	alfa-vinkel	e	d	e/d	epi closure	pålagring		alfa-vinkel	e	d	e/d	epi closure	pålagring	
Bild 1 = kl 3		50,6	28,3	48,5	0,58	closed			48,2	31,1	47,8	0,65	closed		
Bild 2 = kl 2		46,1	28,4	47,8	0,59				47,1	29,4	47,8	0,62			
Bild 3 = kl 1		52,6	35,2	47,4	0,74			klart	47,8	31,4	47,2	0,67			
Bild 4 = kl 12		50,9	36,2	48,3	0,75			kollat	49,6	32,9	47,3	0,70			
Bild 5 = kl 11		60,6	34,4	48,1	0,72		finns	med F	61,4	37,1	47,6	0,78			finns
Bild 6 = kl 10		48,2	34,6	47,4	0,73				53,1	33	47,5	0,69			
Bild 1 = kl 9		50,3	32,8	48,5	0,68				49,7	28,6	47,8	0,60			
18. NN2	YYMMDD-NNNN (17år)	alfa-vinkel	e	d	e/d	epi closure	pålagring		alfa-vinkel	e	d	e/d	epi closure	pålagring	
Bild 1 = kl 3		46,9	26,1	49,9	0,52	closed			49,9	28	49	0,57	closed		
Bild 2 = kl 2		49,5	24,2	48,7	0,50			klart	48,5	23,4	49	0,48			
Bild 3 = kl 1		51,5	36,3	47,9	0,76			kollat	49,9	34,4	48,7	0,71			
Bild 4 = kl 12		49,7	30,6	49,1	0,62			med F	51,1	34	49,7	0,68			
Bild 5 = kl 11		58,4	36	49,4	0,73		finns		58,1	38,4	50,2	0,76			finns
Bild 6 = kl 10		54,8	28,8	49,9	0,58				54,1	29,9	50,1	0,60			
Bild 1 = kl 9		40,8	37,9	49,9	0,76				47,8	36	49	0,73			
19. NN3	YYMMDD-NNNN (18 år)	alfa-vinkel	e	d	e/d	epi closure	pålagring		alfa-vinkel	e	d	e/d	epi closure	pålagring	
Bild 1 = kl 3		49	27,6	50,3	0,55	closed			52,2	26,7	49,6	0,54	closed		
Bild 2 = kl 2		52,5	24,1	49,8	0,48				52,8	23,8	49,6	0,48			
Bild 3 = kl 1		51	31,1	49,7	0,63			klart	52,2	35,2	49,8	0,71			
Bild 4 = kl 12		51,9	33,4	49,5	0,67			kollat	53,1	30,8	48,6	0,63			
Bild 5 = kl 11		53,8	34,9	50,7	0,69			med F	52,5	34	50	0,68			
Bild 6 = kl 10		52,8	31,7	51,1	0,62				52,7	32,1	50	0,64			
Bild 1 = kl 9		45	34,5	50,3	0,69				50,8	35	49,6	0,71			

Kliniska Studier Sverige Forum Norr

Datafil

2014-02-15		8 patient													
Patient	**personnr**	**hö höft**							**vä höft**						
17. NN	YYMMDD-NNNN (18år)	alfa-vinkel	e	d	e/d	epi closure	pålagring		alfa-vinkel	e	d	e/d	epi closure	pålagring	
Bild 1 = kl 3		50,6	28,3	48,5	0,58	closed			48,2	31,1	47,8	0,65	closed		
Bild 2 = kl 2		46,1	28,4	47,8	0,59				47,1	29,4	47,8	0,62			
Bild 3 = kl 1		52,6	35,2	47,4	0,74			klart	47,8	31,4	47,2	0,67			
Bild 4 = kl 12		50,9	36,2	48,3	0,75			kollat	49,6	32,9	47,3	0,70			
Bild 5 = kl 11		60,6	34,4	48,1	0,72		finns	med F	61,4	37,1	47,6	0,78		finns	
Bild 6 = kl 10		48,2	34,6	47,4	0,73				53,1	33	47,5	0,69			
Bild 1 = kl 9		50,3	32,8	48,5	0,68				49,7	28,6	47,8	0,60			

Går inte att använda för beräkningar!

- Text och siffror blandade i samma kolumn
- Identifierare på egna rader
- Tom rad som avskiljare mellan patienter
- Samma variabelnamn höger och vänster höft

Kliniska Studier Sverige Forum Norr

Datafil

Visuell fil för förståelse och kommunikation

2014-02-15		8 patient														
Patient	**personnr**	**hö höft**							**vä höft**							
17. NN	YYMMDD-NNNN (18år)	alfa-vinkel	e	d	e/d	epi closure	pålagring		alfa-vinkel	e	d	e/d	epi closure	pålagring		
Bild 1 = kl 3		50,6	28,3	48,5	0,58	closed			48,2	31,1	47,8	0,65	closed			
Bild 2 = kl 2		46,1	28,4	47,8	0,59				47,1	29,4	47,8	0,62				
Bild 3 = kl 1		52,6	35,2	47,4	0,74			klart	47,8	31,4	47,2	0,67				
Bild 4 = kl 12		50,9	36,2	48,3	0,75			kollat	49,6	32,9	47,3	0,70				
Bild 5 = kl 11		60,6	34,4	48,1	0,72		finns	med F	61,4	37,1	47,6	0,78		finns		
Bild 6 = kl 10		48,2	34,6	47,4	0,73				53,1	33	47,5	0,69				
Bild 1 = kl 9		50,3	32,8	48,5	0,68				49,7	28,6	47,8	0,60				
		alfa-				epi			alfa-				epi			

Datateknisk fil för analys

Datum	Patnr	Alder	Bild	Klockan	ho_alfa	ho_e	ho_d	ho_ed	ho_epi	ho_palag	va_alfa	va_a	va-d	va_ed	va_epi	va_palag	comment
2014-02-15	17	18	1	3	50,6	28,3	48,5	0,58	1		48,2	31,1	47,8	0,65	1		
2014-02-15	17	18	2	2	46,1	28,4	47,8	0,59			47,1	29,4	47,8	0,62			
2014-02-15	17	18	3	1	52,6	35,2	47,4	0,74			47,8	31,4	47,2	0,67			klart
2014-02-15	17	18	4	12	50,9	36,2	48,3	0,75			49,6	32,9	47,3	0,7			kollat
2014-02-15	17	18	5	11	60,6	34,4	48,1	0,72		1	61,4	37,1	47,6	0,78		1	med F
2014-02-15	17	18	6	10	48,2	34,6	47,4	0,73			53,1	33	47,5	0,69			
2014-02-15	17	18	1	9	50,3	32,8	48,5	0,68			49,7	28,6	47,8	0,6			

Kliniska Studier Sverige Forum Norr

Datafil

Den generella principen är alla typer av datatekniska filer avsedda för beräkning är:

All information ska finnas på en rad

Varje kolumn (variabel) får endast innehålla en information

Datum	Patnr	Alder	Bild	Klockan	ho_alfa	ho_e	ho_d	ho_ed	ho_epi	ho_palag	va_alfa	va_a	va-d	va_ed	va_epi	va_palag	comment
2014-02-15	17	18	1	3	50,6	28,3	48,5	0,58	1		48,2	31,1	47,8	0,65	1		
2014-02-15	17	18	2	2	46,1	28,4	47,8	0,59			47,1	29,4	47,8	0,62			
2014-02-15	17	18	3	1	52,6	35,2	47,4	0,74			47,8	31,4	47,2	0,67			klart
2014-02-15	17	18	4	12	50,9	36,2	48,3	0,75			49,6	32,9	47,3	0,7			kollat
2014-02-15	17	18	5	11	60,6	34,4	48,1	0,72		1	61,4	37,1	47,6	0,78		1	med F
2014-02-15	17	18	6	10	48,2	34,6	47,4	0,73			53,1	33	47,5	0,69			
2014-02-15	17	18	1	9	50,3	32,8	48,5	0,68			49,7	28,6	47,8	0,6			

Variabelnamn

- Oavsett lösning, försök vara konsekvent
- Använd gärna enbart gemener för variabelnamn
- Undvik å, ä och ö
    - Flera program klarar att hantera detta, men inte alla…
- Korta men ändå någorlunda begripliga namn
- Använd inte mellanslag, det är bättre med ”_” eller ”-”

Namngivning av filer

- Var konsekvent, använd ett system
- Sätt gärna datum på uppdaterade filer
- Undvik mellanslag
- Undvik å, ä och ö
- Använd inte sådant som antyder att det är en färdig fil – ”_final”, ”_clean” eller liknande. Det är sällan sant.
- PHD Comics: notFinal.doc

Tänk ”tidy data”

- Varje variabel är en kolumn
- Varje observation är en rad

Undvik att:

- Använda färgade rader som ska betyda något särskilt, ex grupptillhörighet eller ”högt värde”
- Blanda text och siffror i samma kolumn, dela istället i två kolumner
- Lägga flera svar i samma cell, ex ”Vilket eller vilka av följande alternativ stämmer in på din situation?” med svar ”1,2”, ”2,3,4” eller liknande. Ta istället en kolumn per svarsalternativ.
- Blanda olika sätt att mata in datum

Missing

- Oftast bra att inte bara låta det vara tomt i cell
- Hur det ska kodas beror på programvara:
  - I SPSS läggs numeriska värden in som definieras som "missing", kan vara av olika typer
  - I R används NA
- Snyggast att använda särskilda koder för missing i SPSS, men det är sällan kritiskt, tomma celler fungerar också
- Vissa föreslår att "missing" i numeriska variabler inte ska kodas med siffror, dock ett måste i SPSS. Sätt in orimligt värde, ex –999.

Dokumentera och städa!

- Lägg till bra förklaringar till variablers innehåll
- Lägg till bra förklaringar till koder som används
- Om du arbetar med Excel för inmatning, lägg gärna in detta i egen flik
- I SPSS, sätt skalnivå (SCALE för numeriska/kvantitativa variabler,…)
- Sätt lämpligt antal decimaler
- Definiera upp eventuella ”missing”-värden
- Skapa kodbok!
- Se till att du på något sätt dokumenterar uppgifter om beräkningar som gjorts, definitioner på nya variabler,…
    - Syntax och/eller loggbok

Excel för manuell inmatning?

- Många gånger ett bra val, åtminstone lika bra som andra lösningar
- Excel har en massa "smarta" funktioner som kan ställa till det
- Om du jobbar själv är det lättare att vara konsekvent
- Är ni flera, se till att ni använder samma struktur: samma namn, samma flikar, samma koder, samma datumformat,…
- Använd gärna funktioner för "verifiering", dvs begränsning av vad som går att mata in i celler: heltal mellan 1 och 5, enbart datum,...
- Alternativ?
    - Databassystem (MS Access m fl), RedCap,…?

”Programmera” körningar - syntax

- Att använda menyerna går, men
  - det kan svårt att hålla reda på vad som görs och hur det ska göras
  - om data kommer att ändras – uppdateringar, fler poster – tar det tid att göra om alla bearbetningar
  - en hel del varianter saknas i menyerna
- Lösning: syntax/script/do-filer
- Kommandon i text, körs när du vill utföra en viss bearbetning

Några skäl att använda syntax

- Du får en direkt dokumentation av vad som gjorts
- Förklarande kommentarer är lätta att lägga in. De behövs!
- Om datamaterialet inte är stabilt utan kommer att ändras kan alla bearbetningar göras om väldigt lätt
- Inget motsättning mellan menyer och syntax, experimentera gärna i menyerna men kör sista variant med syntax

- Kan ersätta en massa filer: med grundfil för inläsning av data och bearbetning med syntaxfil behövs inga ”nya” datafiler eller outputfiler

Skäl att *inte* använda syntax?

- I SPSS: det blir några extra steg som ska utföras, först *Paste*, sedan gå till syntax-fil och till sist köra kommandon
- En till filtyp att hålla reda på

- Tunna argument…

Varningsflagg!

- Märker du att du (eller handledare) börjar mixtra med klipp-och-klistra, till och från Excel,… ? Aldrig ett bra tecken!
    - Ibland ohjälpligt, oftast mindre lyckade format i grunddata
    - En åtgärd - om det inte gjorts för mycket - är att börja om med import av grundfil, och då styra inläsningen hårdare
- Räkna förfluten tid mellan två datum för hand? Sällan bra.
- Rensa i fil utan att dokumentera vad du gjort, ex exkludera patienter. Du måste åtminstone spara filer både före och efter rensning.

Hur spara data?

- Under tiden du gör dina bearbetningar, använd det program du råkar köra
- När du är klar, spara ner data i det format som använts men också som "rena" text-filer: csv, tab,…, dvs s k icke-proprietära format
- Glöm inte att också spara ner dokumentation av variabler och koder, gärna även det i csv eller tab-filer

Sammanfattningsvis

- Struktur är viktigt!
- Det mesta går att lösa, knasiga strukturer går ofta att stuva om i, men det kan kräva väldigt mycket tid
- Syftet med fil måste styra:
  - Läsa visuellt? Kolla enstaka detaljer?
  - Bearbeta, analysera?
- Bearbetningsbara material bygger på en rätt enkel struktur
- Bli inte förvånad om det tar oväntat mycket tid att skapa en bra struktur och att dokumentera innehållet

Referenser

- Eloranta, S. m fl, Att strukturera och dokumentera forskningsprojekt, Läkartidningen, nr 8, 2013
- Broman, K.W. & Woo, K.H., Data Organization in Spreadsheets, The American Statistician, 72(1), 2018
- Collier, J., SPSS syntax - a beginner's guide, SAGE, 2010
- Svensk nationell datatjänst, Hantera data, del av SND:s webbplats
- Riksarkivet, Värdera och gallra

Kika gärna in på vår CANVAS-sida

https://www.canvas.umu.se/courses/2600

Nästa föredrag:

1/6 kl. 12:00 - 13:00 | Introduction to Survival Analysis (In English)

Survival analysis, also known as time-to-event analysis, is widely used in medical research to investigate the time it takes from a specific event of interest to occur, such as death or disease progression. This seminar will provide an introduction to survival analysis, such as drawing survival curves, comparing survival curves, and modelling the effect of variables on survival.

Wendy Yi-Ying Wu, Region Västerbotten / Institutionen för Strålningsvetenskaper, Onkologi, Umeå Universitet

Kliniska Studier
Sverige
Forum Norr

Kliniska Studier Sverige Forum Norr

- så förbättrar vi möjligheterna att bedriva kliniska studier i Norrland

REGION NORRBOTTEN

region västerbotten

UMEÅ UNIVERSITET

REGION JÄMTLAND HÄRJEDALEN

Region Västernorrland

Kliniska Studier Sverige – Forum Norr

- Plattform för forskningsstödjande infrastruktur och en bro mellan hälso- och sjukvård, näringsliv och akademi i Norra sjukvårdsregionen
- Samarbete mellan:
  - Region Jämtland Härjedalen
  - Region Västernorrland
  - Region Västerbotten
  - Region Norrbotten
  - Umeå universitet
- Regional nod för Norra sjukvårdsregionen

Kliniska Studier
Sverige
Forum Norr

Kliniska forskningscentrum

Forum Norr vill sänka trösklar för klinisk forskning i Norra sjukvårdsregionen

Lokaler i **Sunderbyn**, **Sundsvall**, **Umeå** och **Östersund**

Samverkan som möjliggör inkludering av studiedeltagare från en större del av norra sjukvårdsregionen

Foto: Lisette Marjavaara

Foto: Sara Rönnberg

Foto: Beatrice Backman Lönn

Foto: Erik Holmstedt

Kliniska Studier Sverige Forum Norr

Tack för oss!

Finns det frågor?