## Longitudinal Linkage of Cross-Sectional NCDS Data Files Using SPSS

## What are we doing when we merge data from two sweeps of the NCDS (i.e. data from different points in time)?

We are adding new information (i.e. new variables) for the same set of individuals. Each variable refers to an attribute at that specific point in time, not necessarily a lasting attribute. All variable names in NCDS are unique, so there is no danger of mixing up two variables from different NCDS surveys, even if they are asking the same question (e.g. what is your job now?).

The variable **serial** is used as the link variable to ensure the correct case is matched at one time-point to the same cohort member at a different point in their life.

🛅 nc ds4	4.sav - SPS	S Data Edito	r					🛅 nc ds 🗄	ōcmi.sav -	SPSS Data E	ditor				
File Edit	View Data	a Transform	Analyze Grap	ıhs Utilities	Window Help			File Edit	View Data	Transform	Analyze Grap	ohs Utilities V	Window Help		
2	<b>a</b> 🔍 -	0 GI 🔚	i? 🐴 📲		9 <b>15</b>   39 (0	<b>b</b>		<b>26 5 6 6 6 6 6 6 6 6 6 6</b>							
1 : serial		0.	10002M				1 : serial		01	0001K					
	serial	n622_4	n4region	resps4	n4113	n4118			serial	person	n622_5	n5region	n5gor	resps5 r	
1	010002M	2	2	1	729	1		1	010001K	1	1	8	2	1	
2	010003P	1	1	1	740	1		2	010003P	1	1	8	2	1	
3	010004R	1	2	1	717	1		3	010004R	1	1	9	10	1	
4	010006∨	1	2	1	608	1		4	010006∨	1	1	8	2	1	
5	010007X	2	10	1	209	1		5	010007X	1	2	10	11	1	
6	010008Z	1	2	1	32	1		6	010010L	1	1	8	2	1	
7	010009B	1	2	1	179	1		7	010011N	1	1	8	2	1	
8	010010L	1	2	1	652	1		8	010012Q	1	1	8	2	1	
9	010013S	2	2	1	9	1		9	010013S	1	2	8	2	1	
10	010014U	2	2	1	652	1		10	010014U	1	2	8	2	1	
11	010015W	1	2	1	229	1		11	010016Y	1	2	2	3	1	
12	010017A	1	8	1	559	1		12	010017A	1	1	2	3	1	
13	010019E	2	2	1	87	1		13	010019E	1	2	8	2	1	
14	010020P	1	2	1	166	1	_	14	010022T	1	2	4	6	1	
15	010021R	2	2	1	503	1		15	010023V	1	2	8	2	1	
16	010022T	2	2	1	648	1	_	16	010026B	1	2	8	2	1	
47	040000			4	500	4	_	47	0400070	4	4			4	

As an example, suppose we want to merge the data from NCDS4 with NCDS5:

When downloading from the UK Data Archive, data files are normally made available in SPSS portable format (e.g. ncds4.por), so they can be read by other software. But for the sake of the examples in this document, it's assumed that each file has been converted to an SPSS system file (e.g. ncds4.sav).

Each cohort member is referred to as a *case*, and the data for each case occupy one line of the displayed matrix. So we can see 17 cases in each of the diagrams above.

The variable names are here listed along the top line. For NCDS4 they are: serial, n622\_4, n4region, resps4, etc.

For NCDS5 (cohort member data) they are: serial, person, n622\_5, n5region, n5gor, etc.

Note that all the variable names are different, except for the link variable serial.

In merging the two files, we are concatenating the two sets of variables. This works perfectly for cases like 010003P, 010004R, 010006V, which are present at both time points.

But what about cases like 010002M, 010008Z and 010009B, which are present at NCDS4 but not NCDS5? Or cases like 010001K, 010011N and 010012Q which are present at NCDS5 but not at NCDS4?

The diagram below shows what happens in the merged file. We simply have missing values for all the variables corresponding to the sweep (NCDS4 or NCDS5) in which that cohort member wasn't present:

💼 nc ds4	4 and ncds	5cmi merge	d.sav - SPSS	Data Editor	r									JX
File Edit	View Data	Transform	Analyze Grap	ohs Utilities	Window Help									
28														
1 : serial 010001K														
-	serial	n622_4	n4region	resps4	n4113	n4118	person	n622_5	n5region	n5gor	resps5	resp5cmi	resp5wyt	n 🔺
1	010001K						1	1	8	2	1	1.00	1.00	
2	010002M	2	2	1	729	1								
3	010003P	1	1	1	740	1	1	1	8	2	1	1.00	1.00	
4	010004R	1	2	1	717	1	1	1	9	10	1	1.00	1.00	
5	010006V	1	2	1	608	1	1	1	8	2	1	1.00	1.00	
6	010007X	2	10	1	209	1	1	2	10	11	1	1.00	1.00	
7	010008Z	1	2	1	32	1								
8	010009B	1	2	1	179	1								
9	010010L	1	2	1	652	1	1	1	8	2	1	1.00	1.00	
10	010011N						1	1	8	2	1	1.00	1.00	
11	010012Q						1	1	8	2	1	1.00	1.00	
12	010013S	2	2	1	9	1	1	2	8	2	1	1.00	1.00	
13	010014U	2	2	1	652	1	1	2	8	2	1	1.00	1.00	
14	010015W	1	2	1	229	1								
15	010016Y						1	2	2	3	1	1.00	1.00	
16	010017A	1	8	1	559	1	1	1	2	3	1	1.00	1.00	
17	010019E	2	2	1	87	1	1	2	8	2	1	1.00	1.00	
18	010020P	1	2	1	166	1								

## How do we go about merging files, using analysis software like SPSS?

To merge two data files in SPSS (e.g. the combined childhood dataset ncds0123.sav and the 23-year-old dataset ncds4.sav), two approaches are possible. If you are familiar with SPSS syntax, this is the quickest and most elegant way. If not, you can use the drop-down windows menus. In section (a) below we give detailed instructions for using syntax, and in section (b) we explain the menus approach.

🗰 ncds0123.sav - SPSS Data Editor													
File Edit	View Data	1 Transform	Analyze Grap	ohs Utilities W									
2	a 🔍	0 a 🔚	12 M 1										
1 : serial	1 : serial 010001K												
	serial	n622	nOregion	n1region									
1	010001K	1	2	2									
2	010002M	2	2	2									
3	010003P	1	2	1									
4	010004R	1	2	2									
5	010005T	2	2	2									
6	010006V	1	2	2									

🗰 nc ds4	1.sav - SPS	S Data Edito	r		
File Edit	View Data	a Transform	Analyze Grap	ohs Utilities '	W
28	<b>a</b> 🖳 🖢	୍ ରା 🔚	l? 🕺 州		
1 : serial		01	10002M		
	serial	n622_4	n4region	resps4	
1	010002M	2	2	1	
2	010003P	1	1	1	
3	010004R	1	2	1	
4	010006∨	1	2	1	
5	010007X	2	10	1	
6	010008Z	1	2	1	

## (a) Syntax approach

\*\* open each cross-sectional file, sort by serial and save. get file='c:\ncds4.sav'. sort cases by serial. save outfile='c:\ncds4.sav'.

get file='c:\ncds0123.sav'. sort cases by serial. save outfile='c:\ncds0123.sav'.

```
** append ncds4.sav to the 'active' file ncds0123.sav.
match files file=*
/file='c:\ncds4.sav'
/by serial.
save outfile='c:\ncds01234.sav'.
```

You can continue in this way to merge more files to the 'active file'. For instance, having first sorted ncds5cmi.sav by serial, you could proceed by executing the following syntax in place of the above *save outfile='c:\ncds01234.sav*' command:

```
match files file=*
/file='c:\ncds5cmi.sav'
/by serial.
save outfile='c:\ncds012345cmi.sav'.
```

## (b) Drop-down menus approach

First we need to sort the two data files. It makes sense for this to be done in reverse order, so that we end up with ncds0123.sav as the currently 'active' file at the end of the sorting process.

So read in ncds4.sav (File > Open > Data > Ncds4.sav).

🗰 Untitled - SPSS Da	ata Editor		
File Edit View Data	Transform	Analyze Graphs	Utilities Wind
New	•	I VA/idth	Decimals
Open	•	Data	
Open Database Read Text Data	•	Syntax Output	
Save Save As	Ctrl+S	Script Other	
Display Data Info Apply Data Dictionary. Cache Data			
Print Print Preview	Ctrl+P		
Switch Server Stop Processor Recently Used Data Recently Used Files	Ctrl+.		
Exit			
17 18 19			

🛅 Untitled - !	SPSS Data Editor					
File Edit View	Data Transform	Analyze Grap	hs Utilities	Window Help		
	💵 🗠 🗠	<b>[?] 孙  </b> ] 揮		1 5 00	1	
1:	1					
	var var	var	Var	Var	Var	var
1		T				-
2	Open File				? 🗙	
3		10.1 (0.)			ok 📼	1
4	Look in:   So Loc	al Disk (C:)		- E		
5	Name 🐨		Size	Туре	Date M 📥	
6	mcds0123.sav		34,892 KB	SPSS Data Docum	ent 16/11/2	
7	mcds6.sav		74,216 KB	SPSS Data Docum	ent 25/11/2	
8	mcds5p.sav		10,458 KB	SPSS Data Docum	ent 24/08/2	
9	mcds5mc.sav		7,696 KB	SPSS Data Docum	ent 24/08/2	
10	mcds5cmi.sav		46,038 KB	SPSS Data Docum	ent 16/11/2	
11	ncds4.sav		242 KB	SPSS Data Docum	ent 16/11/2	
12	exams.sav		4,685 KB	SPSS Data Docum	ent 08/11/2	
13	<				2	
14	File name: nco	ls4.sav			<u>O</u> pen	
15						
16	Files of type: SP	55 (^.sav)		<u> </u>	<u>aste</u>	
17					Cancel	
18						2

💼 ncds4.sa	av - SPSS	Data Edito	ог		
File Edit Vi	ew Data	Transform	Analyze	Graph	s Utilities W
2 🕞 🕞	De Co De	fine Variable I py Data Prop fine Dates	Properties. erties		¥ <b>=</b> 4
	ser Ins	ert Variable			resps4
1 01		ert Cases			1
2 01	000	to case			1
3 01	00C Soi	rt Cases			1
4 01	OOC Tra	anspose			1
5 01	000 <sup>Re</sup>	structure			1
6 01	OOC <sup>Me</sup>	rge Files		•	1
7 01		gregate Setifu Duslical			1
8 01	001 ou	hogonal Desi	ie Cases	· . ا	1
9 01	001	nogonal Desi	90		1
10 01	001 <sup>Spl</sup>	it File			1
11 01	001 Sel	ect Cases			1
12 01	001 ~~~ <sup>we</sup>	ight Cases			1
13 01	0019E	2		2	1

Then **Sort** the link variable (**Data > Sort Cases > by serial**)

🛄 ncds4	4.sav - SPSS Date	a Editor					
File Edit	View Data Tran	sform Anal	yze Graphs	Utilities Wind	dow Help		
2 I : serial	Sart Caree	01000	<b>₩ 1</b> [1] 2M	<u>= </u> 4	<b>\$</b> 0		n41
1 2 3 4 5	<ul> <li>♣ n622_4</li> <li>♣ n4region</li> <li>♣ resps4</li> <li>♠ n4113</li> </ul>		Sort I	oy: serial (A)		OK Paste <u>B</u> eset	
6 7 8 9	<ul> <li> <b>#</b>) n4118      </li> <li> <b>#</b>) n4120         </li> <li> <b>#</b>) n4121         </li> <li> <b>#</b>) n4122         </li> </ul>	>	Sol C	t Order Ascending Descending		Cancel Help	
10	0100140	2	2	1	652 000	1	
	01001344	1	2		229		

Now save the sorted file Ncds4.sav (**File > Save**)

Repeat the above procedure for the file Ncds0123.sav. At the end of this process, the sorted Ncds0123.sav will be the currently 'active' file.

💼 ncds0123.sav - SPSS Data Editor File Edit View Data Transform Analyze Graphs Utilities Window Help Define Variable Properties... 🖻 🔒 🎒 🖻 \* ₽**₽ ≣ % %** Copy Data Properties... 1 : serial Define Dates... Insert Variable n2region n1 region ser Insert Cases 1 01000 2 2 Go to Case... 2 01000 2 2 3 01000 Sort Cases... 1 1 Transpose... 2 2 4 01000 Restructure... n 5 01000 2 Merge Files Add Cases... Þ 6 01000 2 Add Variables.. Aggregate... 7 01000 Identify Duplicate Cases... 2 8 01000 2 Orthogonal Design ۲ 2 9 01000 2 Split File... 2 2 10 01001 Select Cases... 2 2 11 01001 Weight Cases... 1201001 2 2

Now we are ready to merge the files (**Data** > **Merge Files** > **Add Variables**)

You will see a dialogue box entitled 'Add variables: read file'. Enter the name of the other file (in this case, **Ncds4.sav**)

💼 ncds(	)123.s	iav - S	iPSS Dat	a Ed	itor							
File Edit	View	Data	Transfor	m A	inalyze Graj	ohs Utilities	Windo	w Help				
2	<b>e</b>	<b>u</b>		-	? 🐴 📲		ı.∣≣	0				
1 : serial				010	1001K							
	ser	ial	n622		nOregion	n1region	n2r	region	n3re	gion	resp	ps0
1	01000	IK I		1	2		2	2		2	_	1
2	0100	Add	Variable	s: R	ead File					?	X	1
3	0100							ī. e				1
4	0100	Look	tin:   🥯	Local	Disk (C:)		-	- E		•		1
5	0100	Na	me 🔻			Size	Туре			Date	A 1	1
6	0100	The second	cds0123.	sav		34,8	SPSS I	Data Docu	ment	16/11	1	1
7	0100	i i i i i i i i i i i i i i i i i i i	cds6.sav			74,2	SPSS I	Data Docu	ment	25/11		1
8	0100	i 🎬 r	cds5p.sa			10,4	SPSS I	Data Docu	ment	24/08	1	1
9	0100	j 🏢	ncds5mc.s	av		7,69	SPSS (	Data Docu	ment	24/08	1	1
10	0100		ocds5cmi.s	av		46,0	SPSS I	Data Docu	ment	16/11	1	1
11	0100		ods4.sav			24,1	SPSS [	Data Docu	ment	16/11	1	1
12	0100	• • • •	xams.sav			4,68	SPSSI	Data Docu	ment	08/11	/ <u>~</u>	1
13	0100	<								>		1
14	0100	File n	ame:	neds	4.sav			_		Open		1
15	0100										-	1
16	0100	Files	of type:	SPS	S (*.sav)			-		Cancel		1
17	0100											1
18	0100									_	1	1
10	01001			2	1		1	7		1		1

Another dialogue box appears, showing a list of all the variables in both files combined, with a separate list of 'excluded variables'. These are variables that appear in both files. Note that there is only one 'excluded' variable, since 'serial' is the only variable whose name is repeated across different datasets.

🛅 ncds(	0123.	.sav - S	SPSS Data	Editor				
File Edit	View	Data	Transform	Analyze	Graphs Utilities	: Window Help	ę.	
2	-		101 🟪	[? M	* 💼 📰	1. 	2	
1 : serial				010001K				
1	se	erial	n622	nOregio	n n1regior	n2region	n3region	resps0 res
1	010	Add V	ariables fr	om C:Anci	ded say	- î	1 1	
2	010	Auuv	arrables fi	om e. met	15-1.544			
	010	Exclud	led Variables	:		<u>N</u> ew Workin	g Data File:	ОК
5	010	serial	:_ <u>+</u>			seriak (*) n622 (*)	1	Paste
6	010				•	nOregion (*	2	Beset
7	010					n Tregion (" n2region (*	)	
8	010					n3region (*	)	Cancel
9	010	-	Pop			resps1 (")		Help
10	010		nenj	3008		(resps2 (*)		9
11	010	Ma	itch cases or Roth files pro	n key variable	es in sorted files	Key <u>V</u> ariable	es:	-
13	010	C	External file is	s keyed table	- D			-
14	010	C	 ∕∕_orking Dat	a File is keyer	d table	-		
15	010	E Inc	licate case si	ource as varia	able: source0	1	-	
16	010	(*) = W	orking Data	File (+	) = C:\ncds4.sav			
17	010		-	1	-	-		
18	0100	)18C		1	2	2 2	2	1

Check the box 'Match cases on key variables in sorted files', and underneath it make sure the (default) option 'both files provide cases' is checked. Then highlight the name of the link variable ('**serial**') in the 'excluded variables' box, and click the **right arrow** button to the left of the 'Key Variables' box at the bottom right. This will paste the variable name 'serial' into the 'Key Variables' box.

Then click **OK.** This will merge the two files ncds0123.sav and ncds4.sav.

🛅 ncds012	23.sav - SPSS Data Editor	
File Edit V	/iew Data Transform Analyze Graphs Utilities Window Help	
	) 🔍 🗠 🗠 🗛 📲 🖿 🏛 🌉 💊 🔕	
1 : serial	010001K	
	serial n622 n0region n1region n2region n3region resps0	res
2 01	Add Variables from C:\ncds4.sav	-
30 40 50 60 70 80 90 100 1100 120	Excluded Variables: ► New Working Data File: New Working Data File: New Working Data File: New Working Data File: Negon (1) n2region (1) n2region (1) n3region (1) n3region (1) resps1 (1) resps1 (1) resps2 (1) resps2 (1) resps2 (1) resps2 (1) resps2 (1) resps3 (1) Key Variables: C Both lise rouvide cases	
12 0 13 0 14 0 15 0 16 0 17 0	C External file is keyed table     Working Data File is keyed table     Indicate case source as variable: source01 (*) = Working Data File (+) = C:\ncds4.sav	

You can then save and name the combined dataset as, say NCDS01234.sav (**File > Save As**)

<b>1</b>	nc ds0	0123 -	SPSS	5 Data	Edito	r		
File	Edit	View	Data	Transf	orm	Analy	/ze	Grap
N	ew					►	88 I	>=
0	pen					•		
0	pen D	atabase	Э			•	ĸ	
R	ead Te	ext Data	a				regio	on
S	ave				Ctrl+	s		2
S	ave As	s						2
M	lark Fil	e Read	Only					2
D	isplay	Data Fi	le Info	rmation		•		2
C	ache [	Data						2
SI	top Pri	ocessor			Ctrl+			2
S	witch S	5erver.						2
P	rint Pri	eview						2
P	rint				Ctrl+	P		2
<u> </u>			Data					2
	ecenti ecenti	y Used v Used	Data Filec					2
	ecenti	y oseu	r iies					2
E	×it							2
	14	01001	4U		2			2
	15	01001	5W		1			2

You can continue to merge further datasets (e.g. NCDS5cmi.sav, NCDS6.sav) with this combined file in the same way as above.

Note that at each cross-sectional time point, there will be some NCDS members who were not present at that survey. This will mean all the variables corresponding to that survey will be system-missing for that person in the combined longitudinal dataset, but variables for other time points will have valid values (see first section above: "What are we doing...").

## Merging Child Data for NCDS5

The NCDS5 survey was unique in that, for a 1 in 3 sample, the mother was asked a detailed questionnaire about her children (the 'mother' interview), then asked a separate series of questions about each child in turn (the 'your child' interview), then tests were administered to each child (the 'child' interview).

The contents of each of these questionnaires can be summarised as follows:

#### **Mother Interview**

Family life Pregnancy and birth Health history Separations from mother Experience of being "in care" Pre-school experience Schooling history Experience of day care

#### **Your Child Interview**

(mother answering questions about each child in turn)

Motor and Social Development Behaviour Problems Index Temperament Home Environment

#### **Child Interview/Assessments**

(conducted directly with each child, tailored to specific age ranges)

Peabody Picture Vocabulary Test McCarthy Scale of Children's Abilities: Verbal Memory Subscale Peabody Individual Achievement Tests: Math Subscale Reading Recognition Subscale Reading Comprehension Subscale Weschler Intelligence Scale for Children: Digit Span Subscale Perceived Competence Scale Self-Perception Profile Plus, an interviewer evaluation of: Testing Conditions Child Temperament Home Environment Child Height and Weight Measurement The distribution of numbers of children in families interviewed is summarised below:

Number of Children in Family	Frequency	Percent
1	1383	53.4
2	821	31.7
3	301	11.6
4	69	2.7
5	12	.5
6	1	.0
8	1	.0
Total Mothers	2588	100.0

Number of children listed in mother interview, or undergoing assessments

There was also a fourth set of data where the interviewer filled in a series of observations about the home environment, and the mother's perceived relationship with the specific child: whether the mother talked to the child, responded to them, caressed them or slapped them, etc. (the 'home environment' variables).

The results of these four sets of data were stored in such a way that each child has its own separate record in the data file corresponding to that instrument. Unlike all the other NCDS data files, where the case is uniquely defined as the cohort member, these four sets of data have duplicate serial numbers. So for example, a cohort member with four children is liable to generate four records corresponding to the mother interview, four corresponding to the 'your child' interview, and four records for the 'child' assessments (it's slightly less clear-cut than this, as some children refused the interview or weren't at home, etc., despite being listed in the mother interview).

The particular child within the family is distinguishable by their 'person number' in the household grid (usually 3,4,5 etc., since the mother and father will normally be persons 1 and 2).

Users who have experience of NCDS data from the UK Data Archive in its older format may remember that these four datasets were previously available as four separate files (mother.sav, ychild.sav, child.sav and hoenv.sav). Great care was needed in merging these datafiles with each other and then with the rest of the NCDS longitudinal data. This was because one had to make sure each child of each cohort member was linked not only to their parent's record, but to their own record in the other child-based files, and not to one of their siblings.

We have now made this process much simpler by combining all the NCDS5 childbased files into one composite file, ncds5mc.sav. Like its constituent data files, it contains more than one record per cohort member, but the records for each child can be distinguished by the variable 'person'. Note that the unit of analysis in this file is the *child*, not the cohort member, so that a frequency count of any variable will yield a total of 4,287 (i.e. the number of children involved in the NCDS5 assessments) rather than 2,588, which is the total number of mothers. Note also that the mother is not necessarily the cohort member. In the case of male cohort members, it would normally be their partner who was interviewed.

There are records which have missing values for variables corresponding to some element of the child-based questionnaires. This is because not all children listed by the mother in the 'mother' or 'ychild' interview were present for the 'child' assessments, or agreed to do them.

To merge this combined dataset, ncds5mc.sav, with any other NCDS data file (e.g. ncds4.sav) it is necessary to use a slightly different procedure from the normal 'merge files' process outlined in the earlier part of this document. We have to tell SPSS to conceptualise the ncds4.sav file as a 'look-up table', whose values for all other variables are spread identically across all cases with that particular NCDS serial number.



In the above diagram, the variables 'status', n4113, n4118 and n4120 are from the NCDS4 data file, and apply to the cohort member whose serial number is 010030S.

The variables 'person', n516220, n516221 and n516223 are from the ncds5mc file, and apply to each of the four children of that cohort member, who are distinguished by their person numbers. When we merge the two files, we need to ensure the values of the ncds4 variables (which are not child-specific) are attached in exactly the same way to each of the four children, as in the diagram below:

	serial	status	n4113	n4118	n4120	person	n516220	n516221	n516223
6	010030S	8	137	1	2	3	1	18	4
7	010030S	8	137	1	2	4	2	9	3
3	010030S	8	137	1	2	5	2	14	6
3	010030S	8	137	1	2	6	2	29	5

#### result of merging, using ncds4 as look-up table

The values of the variables 'status' n4113, n4118 and n4120 are replicated four times in this case, but the values of 'person', n516330, n516221 and n516223 remain distinct for each child. In this merged file, the unit of analysis is the child, not the cohort member.

To do this merge in SPSS, as in the normal file merge procedure outlined earlier, one can use the 'syntax' approach or the 'drop-down menus' approach.

## (a) Syntax approach

\*\* open the ncds4 data file and file ncds5mc.sav, sort by serial and save. get file='c:\ncds4.sav'. sort cases by serial. save outfile='c:\ncds4.sav'.

get file='c:\ncds5mc.sav'. sort cases by serial. save outfile='c:\ncds5mc.sav'.

\*\* spread the values of ncds4 variables identically over all the children in each family.

```
match files file=*
/table='c:\ncds4.sav'
```

/by serial.

```
save outfile='c:\ncds4_ncds5mc.sav'.
```

## (b) Drop-down menus approach

First read in the combined child-based dataset, ncds5mc.sav, as the active file (**File** > **Open** > **Data** > ncds5mc.sav).

١	Jntitl	ed - S	PSS D	ata Editor				
File	Edit	View	Data	Transform	Analyze	Graphs	Utilities	Wind
N	lew			•		/idth	Decima	ils
0	)pen			•	Data.			
	pen Da	atabase	•	•	Synta	ax		
R	lead Te	ext Data	3		Outpu	ut		
s	ave			Ctrl+S	Script			
s	iave As				Other	·	+	
D	)isplay I	Data In	fo					
A	pply D	ata Dict	ionary.					
	iache D	ata						
P	rint			Ctrl+P				
P	rint Pre	eview						
s	witch S	ierver						
s	itop Pro	cessor		Ctrl+.				
R	ecently	/ Used	Data	•				
R	tecently	/ Used	Files	•				
E	×it							
-	17				_			
	18							
	19							
<u> </u>	20							

	쾨 🛄 끄 ା 의 🖿	I? <u>例</u> 堆 F			
	Name	Туре	Width	Decimals	
1					
2	Carl Control				
3	Open File				
4	Look in: 🥪 Local [	Disk (C:)	•	+ 🗈 💣 🖩	
5	Name 🔻	1	Size	Tuna	-
7	monds0123.say		34.892 KB	SPSS Data Do	cument
8	mcds6.sav		74,216 KB	SPSS Data Do	cument
9	ncds5p.sav		10,458 KB	SPSS Data Do	cument
10	ncds5mc.sav		7,696 KB	SPSS Data Do	cument
11	ncds5cmi.sav		46,038 KB	SPSS Data Do	ument
10	ncds4.sav		24,134 KB	SPSS Data Do	cument
12	exams.sav		4,685 KB	SPSS Data Do	cument 📐
13	<				>
15	File name: ncds5	imc.sav			Open
16	Files of tupe: CDCC	[* eav]		-	Paste

Then **Sort** the link variable before merging files (**Data > Sort Cases > by serial**)

<b>i</b>	nc ds S	ōmc.sa	w - SP	PSS Data E	ditor				
File	Edit	View	Data	Transform	Analyze	Graph	is Ut	ilities	W
2			Def Cop	ine Variable by Data Prop	Properties. erties				
	1	serial	Ins	ert Variable			0	cima	iis.
	2	perso n5162	Go	ert Cases to Case			0 0		
	4	n5162	Sor	t Cases			0		
	5	n5162	Tra	nspose			0		
-	6 7	n5162 n5162	Me	rge Files		•	0 0		
	8	n5162	Ag( Ida	gregate otifu Duplical	to Coror		0		
	9	n5162	Ort	hogonal Desi	ian		0		
	10	n5162					0		
	11	n5162	Sel Sel	ect Cases			0		
	12	n5162	We	ight Cases			0		
	13	n5162					0		

🛄 ncds5	mc.sav - S	PSS Data Edito	r			
File Edit	View Data	Transform Ana	alyze Grapł	ns Utilities Wi	ndow Help	
🛎 🖬 🤞	s 🔍	) 🖂 🔚 🕼	两情		<b>5</b> 3	
	Name	Туре	Width	Decimals	Label	Values
1 2	Sort C	les ises	la.	lo.		
3 4 5 6	<ul> <li></li></ul>	20 21 23		ort by: A serial (A)	[	DK IIe} aste eset
7 8 9 10	<ul> <li></li></ul>	25 27 29 31 💌		Sort Order Ascending Descending		encel telp tural (
11	1010204	Numeric	0	U	WOTHER. TWJ	INUNE
12	n516236	Numeric	8	0	MOTHER:1 TI	None
13	n516238	Numeric	8	0	MOTHER:1 TI	None

## Click Data > Merge Files > Add Variables

Ĩ	ncds	õmc.sa	iv - SP	<mark>PSS Data</mark> E	ditor				
F	ile Edit	View	Data	Transform	Analyze	Graphs	Utilities	Windov	v Help
	2	<b>e</b>	Def Cop	ine Variable by Data Prop	Propertie: erties	5		ı I	<u>s</u> ¢
		Na	Def	ine Dates			Decima	ils	Labe
Γ	1	serial	Ins	ert Variable		0		SE	I JAIR
Γ	2	perso	Ins	ert Cases		0		Pe	rson N
Γ	3	n5162	GO	to Case		0		M	DTHER
Γ	4	n5162	Sor	t Cases		O O		M	DTHER
Γ	5	n5162	Tra	nspose		0		M	DTHER
Γ	6	n5162	Res	structure		مل		M	тнек
Γ	7	n5162	Mer	rge Files		•	Add Case	es	HER
Γ	8	n5162	Age	gregate			Add Varia	ables	HER
Γ	9	n5162		ntiry Duplica bogopal Dec	te Cases. ian	т <b>,</b> р		M	DTHER
Γ	10	n5162		nogonal Des	ign	o		M	DTHER
Γ	11	n5162	Spli	t File		0		M	DTHER
Γ	12	n5162	Sel	ect Cases		0		M	DTHER
Γ	13	n5162	we	ight Cases		0		M	DTHER
	14	n5160	13	Numeric	8	0		MI	P23:2
Γ	15	n5160	14	Numeric	8	0		MI	P23:2
	40	1 5400	40	KL 1					

As before, you will see a dialogue box entitled 'Add variables: read file'. Enter the name of the other file (here **NCDS4.sav**, which you should also have previously 'sorted' as above).

🛄 ncds5	ōmc.sav - SPSS Dat	ta Editor					
File Edit	View Data Transfo	orm Analyze Gra	iphs Utilii	ties Win	idow Help		
<b>2</b>		🏪 💽 🎮 🖻			<b>N</b>		
	Name	Туре		Width	Decimals		
1	serial	String	7	8 1	n	SERIA	
2	p Add Variables:	Read File				?	🔀 imbe
3	re Lastrin les los	al Diak (Cr)					o NC
4		al Disk (C.)					o NC
5	re Name 🔻			Size	Туре		🔺 o NC
6	re 🛅 ncds0123.sav		34	1,892 KB	SPSS Data Do	ocument	o NC
7	n mncds6.sav		74	1,216 KB	SPSS Data Do	ocument 📍	= 1i) C
8	n mncds5p.sav		10	),458 KB	SPSS Data Do	ocument	1iii)
9	n mncds5mc.sav		7	7,696 KB	SPSS Data Do	ocument	1iii)
10	n mncds5cmi.sav		46	5,038 KB	SPSS Data Do	ocument	1iii C
11	n ncds4.sav		24	1,134 KB	SPSS Data Do	ocument	1iv) (
12	n mexams.sav		4	1,685 KB	SPSS Data Do	cument	1iv) 1
13	n					>	1v) C
14	n File name: nc	ds4.sav				Open	1vi) I
15	n .					- I	1vi) I
16	n Files of type: SF	255 (f.sav)			<u> </u>	Lancel	1 TIN
17	n						1 TIN
18	noncono		-		÷		2 Q1
19	n516014	Numeric	8	9	Π	MOTH	FR-2 02

Another dialogue box appears as before, showing a list of all the variables in both files combined, with a separate list of excluded variables, which should just contain 'serial'.

	Name	Туре	Width	Decimals	Label	Valu	Jes	Missi
1	çoriol	Ctring	7	0		Mono		
2	Add Varia	ables from C:\	ncds4.sav					d ie
3	Excluded \	/ariables:		Nev	/Working Data File:		OK	
- 4	t Leavield (u)	anabioo.			rialz (8)			le
- 5				pe	riak ( ) rson (*)		Paste	le
6				ns ns	(*) 16220 (*)		Reset	le
- 7	r			nt	16223 (*)			le
8				nt	i16225 (*)		Cancel	le
9	1		1	ns n5	i16229 (*)		Help	
10		Rename		nt	i16231 (*)	~		le
11	🚹 🔲 Match (	cases on key varia	ibles in sorted	files Ke	y Variables:			le
12	r 💿 Both	files provide case	s					le
13	r C Exter	mal file is keyed ta	ble					le
14	r 🔿 Work	king Data File is ke	yed table					
15	I Indicate		ariable:	ource01				
16	r (°) = ) rtorkin	na Dista Filo	$(u) = C^{1} h h h h h$	laf anu				
17	r () - Wulkir	ny Data Alle	(+) = C. VICU	194.907				
18	n516020	Numeric	8	0	MIP23:2 Q3b	{O, Too	young}. 9	19
10		h I	0	0	MID22-2-02-	(п т		0

Check the box 'Match cases on key variables in sorted files', but this time make sure the checked option underneath it is 'external file is keyed table'. Then, as before, enter the name of the link variable ('**serial**') in the 'Key variables' box on the lower right by clicking the **right arrow** button to the left of that box. Then click **OK**.



Note that the resulting merged file will still only contain as many cases as there were in 'ncds5mc.sav'. All the additional cases in ncds4.sav will not have been added, because this latter dataset was merely used as a 'look-up' to spread data across cases with the relevant serial numbers in ncds5mc.sav. But all the ncds4 variables will have been added for those cases (assuming the corresponding cohort member was present at NCDS4 – if not, the ncds4.sav variables will all have missing values).

Remember that the values of the added NCDS4 variables will be identical for all children of the same individual mother, because they have the same serial number (although different 'person' numbers).

You can then save and name this dataset (e.g. ncds4\_ncds5mc.sav).

# Merging Child Data in SPSS, Keeping Cohort Member as the Unit of Analysis

The above method succeeds in merging NCDS5 child-centred data with the cohort member's data from other sweeps, but does have the disadvantage of leaving us with a dataset where the unit of analysis is the child.

If you need the unit of analysis to be the cohort member, it is possible in SPSS to produce various summary statistics derived from the variables in the ncds5mc data file, by using the AGGREGATE procedure.

This effectively 'distils down' information on multiple children's records into one variable, which can then be associated with the cohort member.



The summary statistics available are:

SUM	Sum	MEAN	Mean
SD	Standard Deviation	MAX	Maximum
MIN	Minimum	PGT	% cases less than specified value
PLT	% cases less than specified value	PIN	% cases between spec. values
POUT	% cases not in specified range	FGT	Fraction greater than spec. value
FLT	Fraction less than value	FIN	Fraction between values
FOUT	Fraction not in range	Ν	Weighted number of cases
NU	Unweighted number of cases	NMISS	Weighted number of missing cases
NUMISS	Unweighted number of missing cases	FIRST	First nonmissing value
LAST	Last nonmissing value	MEDIAN	Median

In the example above we have the following three child-based variables in file ncds5mc:

- n516220 (sex of child, 1=male, 2=female)
- n516227 (age of child in years)
- n521223 (whether completed basal score 5/5 in PIAT maths (1=yes, 2=no)

Each variable can give rise to a number of summary variables: here, the variable n516220 in the child-based file gives rise to the three variables n516220, n516220\_1 and n516220\_2 in the aggregated file:

- n516220 Total number of children
- n516220\_1 Percent male
- n516220\_2 Percent female

Similarly, variable n516227 gives rise to two variables in the aggregated file:

- n516227 Age of eldest child
- n516227\_1 Mean age of all children

To produce an aggregated file using SPSS, again we have two methods:

### (a) Syntax approach

\*\* open the ncds5mc data file. get file='c:\ncds4.sav'.

\*\* aggregate records, using 'serial' as the break variable, and asking for summary statistics on three variables to be stored in the aggregated file 'c:\aggr.sav'.

```
aggregate
/outfile='c:\aggr.sav'
/break=serial
/n516220 = nu(n516220)
/n516220_1 = pin(n516220 1 1)
/n516220_2 = pin(n516220 2 2)
/n516227 = max(n516227)
/n516227_1 = mean(n516227)
/n521223 = pin(n521223 1 1).
```

### (b) Drop-down menus approach

First read in the combined child-based dataset, ncds5mc.sav, as the active file (**File** > **Open** > **Data** > ncds5mc.sav).

🧰 Untitle	ed - S	PSS D	ata Editor				
File Edit	View	Data	Transform	Analyze	Graphs	Utilities	Wind
New			•	, L NA	/idth	Decima	als
Open			•	Data.			
Open Da	tabase	•	•	Synta	a×		
Read Te:	xt Data	Э		Outpu	ut		
Save			Ctrl+S	Script		-	
Save As.				Other	r	<u> </u>	
Display D	ata In	fo					
Apply Da	ita Dict	ionary.					
Cache Da	ata						
Print			Ctrl+P				
Print Pre	view						
Switch Se	erver						
Stop Pro	cessor		Ctrl+.				
Recently	Used	Data	•	·			
Recently	Used	Files	•				
E×it							
47				-			
17							
18							
19							

	🛃 🖂 🗠 🔚	[? <u>两</u> 理自		5 30	
	Name	Туре	Width	Decimals	
1					
2	and and a				
3	Open File				_ <u>?</u> 🔼
4	Look in: 🗇 Local Di	sk (C:)	•	+ 🗈 💣 🖩	
5	Nome T	1	Sino	Turce	-
7	mane mode0123 cay		34 892 KB	SPSS Data Dor	ument
8	mcds6.sav		74.216 KB	SPSS Data Doc	
a	mncds5p.sav		10,458 KB	SPSS Data Doo	ument
10	ncds5mc.sav		7,696 KB	SPSS Data Doo	ument
11	mcds5cmi.sav		46,038 KB	SPSS Data Doo	ument
10	mcds4.sav		24,134 KB	SPSS Data Doo	ument
12	exams.sav		4,685 KB	SPSS Data Doo	:ument 🔽
13	<				>
15	File name: ncds5m	IC.SAV			Open
16	Files of type: SPSS (	*.sav)		•	Paste
17					

### Click Data > Aggregate

🗰 ncds5mc.sav - SPSS Data Editor									
File Edit	View	Data	Transform	Analyze	Graphs U				
1 : serial	ser 01000	Del Co Del Ins Ins Go	fine Variable py Data Prop fine Dates ert Variable ert Cases to Case	Properties erties					
2 3 4 5	01001 01001 01001 01002	Sor Tra Re: Me	rt Cases anspose structure rge Files						
7	01003	Ag Ide Ort	gregate entify Duplica thogonal Des	te Cases ign					
10 11 12	01004	Spl Sel We	it File ect Cases ight Cases ں		0 0 0				
13	01007	אכי	3		1.00				

A dialogue box appears, inviting you to nominate the break variable (serial), and the aggregate variables (i.e. which ones you'd like distilled into summary statistics in the aggregated file):





By default, SPSS assumes you want the MEAN of that variable, but if you want any of the other summary statistics, click the FUNCTION button. Another dialogue box appears, inviting you to choose other functions. In this case we choose 'Unweighted number of cases':

🛅 ncds5mc.sa	av - SPSS Data Edito				
File Edit View	Data Transform Ana	lyze Graphs Utiliti	es Window Help		
	n lala Isla				
Aggr	regate Data			×	<b> </b>
T. Selk		Break Variable(s)			
per:     per:     per:	son 🔷 👝	A serial		UK	<u>p16220 n5</u>
	John I. D. I. I.				2
res 🛞	Aggregate Data: Ag	gregate Function		et	2
res 🚸	Summary Statistics	Specific Values	Number of cases	Continue	1
	C Mean	C First	C Weighted	Cancel	2
(#) n⊃ (#) n5	C Median	C Last	C Weighted missing		1
n5	C Sum	C Minimum	Unweighted	Help	2
1 X	C Standard deviation	C Maximum	O Unweighted missing		2
	Percentages				2
	C Above				2
1 <u>S</u> av	C Below	er			2
	C Inside				2
C R <u>e</u> r	C Outside	· F	ligh:		2
15 0400	Freedow				1
15 0100	riacuons				2
17 0100	C Above	~			
18 0100	C Below	с. <b>.</b>			2
19/0100	C Inside		liah:		2
2010100	C Outside		1997 - I		2
21 0100	945 51	1.001	1.001 1.001	.00	1
22101005	NQ A	1 00	1.00 00	1.00	2

Continue to specify as many variables as you wish, and as many functions for the same variable as you wish (SPSS automatically attaches  $`_1'...'_2'$  to subsequent ones to distinguish them from the first statistic in the aggregated file):

🗰 nods5mo.sav - SPS	S Data Editor				
File Edit View Data 1	iransform Ana	yze Graphs Utili	ties Window He	elp	
Aggregate D           1: seit              • person             • resp5mth             • resp5mth             • resp5mth             • resp5chid             • resp5chid             • r516220             • n516221             • n516221             • n516223             • n516225             • 151             • 151		Break Variable(s Break Variable(s Aggregate Varia h516220 = ME h516220 = F h516220 = F	ble(s): NN(n516220 1 1) NN(n516220 2 2)		OK 51622 Paste Reset Cancel
□ Save number □ Save number □ C greate new du □ Rgplace work	of cases in breal ata file ing data file	<u>N</u> ame & Lal < group as variable: <u>File</u> C:/a	bel <u>F</u> N_BREAK ggr.sav	unction	
15 010074P	3	1.00	1.00	1.00	.00
161010076T	3	1.00	1.00	00	1.00

Note that, to request the percentage of cases having one specific value (e.g. '2' = girl for variable n516220), you need to request that it be in the range 'Low=2', 'High=2':



Finally, click the OK button in the 'Aggregate Data' dialogue box, and the aggregated file 'c:\aggr.sav' will be produced.

We demonstrated this process on the file ncds5mc.sav, but note that this whole process could equally have been done on a file which had already been longitudinally linked (e.g. to ncds4.sav) using the method in the previous section.

## Working with smaller subsets of data

Linking datasets together longitudinally can produce an enormous file, which may be unwieldy to work with. If you know you are only going to be interested in a certain number of variables for your research, you can produce a subset very simply by running the following syntax:

save outfile='c:\ncdssubset.sav'/keep=serial var1 var2 var3 var4 var5 var6'.

If any of the variables you want to save to the subset happen to lie next to each other in the large dataset (i.e. in the order you normally see them listed on your screen), you can shorten this syntax by referring to the first and last ones in the list:

save outfile='c:\ncdssubset.sav'/keep=serial var1 to var6'.

It is possible to do this using drop-down menus, by clicking File > Save As

۳, the second	ıc ds 5	cmi.s	av - 9	SPSS Da	ta E	dito	ľ		
File	Edit	View	Data	Transfo	rm	Analy	/ze	Grap	ohs
N	ew					•	åå.	>	i e
0	pen					•			
0	pen Da	atabase	e			•	ĸ		
R	ead Te	ext Data	э				22	5	
s	ave			0	Ctrl+	s		1	
S	ave As							1	
M	ark File	e Read	Only					1	
D	isplay I	Data Fi	le Info	rmation		•		1	
c	ache D	ata						2	
S	top Pro	cessor		0	Ctrl+			1	
S	witch S	ierver.						1	
P	rint Pre	eview						1	
P	rint			c	Ctrl+	P		2	
	ocopth	Ucod	Data					2	
	ecentry acaptly	/ Used	Dala Filoc			1		2	
	ecentry	, osea	i iies					1	
E	xit							2	
	14	01002	2T		1			2	
	15	01002	237		1			2	

Then click the Variables button in the dialogue box you see:

🛅 ncds5cmi.	.sav - SPSS Data Edit	tor				
File Edit View	) Data Transform Ar	alyze Graphs Util	ities Window Help			
693	🔍 🖂 🛌 📭	<b>M</b> <u>-</u> [ [ ] ]	34 <b>15</b> 30			
1 : serial	0100	001K				
SE	erial person	n622_5   n5reg	jion n5gor	resps5	resp5cmi	resp5wyt
1 9					1.00	1.00
20	Save Data As				1.00	1.00
30 9	ave in: 🥪 Local Disk (	C:)		## <b>.</b> •	1.00	1.00
4 0 -		1 22		1.100	1.00	1.00
50	Name 🔻	Size	Туре	Date I 🛆	1.00	1.00
60	mncds0123.sav	34,892 KB	SPSS Data Document	16/11/	1.00	1.00
70	ncds6.sav	74,216 KB	SPSS Data Document	25/11/	1.00	1.00
80	ncasop.sav	10,450 KB	SPSS Data Document	24/00/	1.00	1.00
90	ncds5rmi.sav	46.038 KB	SPSS Data Document	16/11/	1.00	1.00
10 0	ncds4.sav	24.134 KB	SPSS Data Document	16/11/	1.00	1.00
11 0	:	101		>	1.00	1.00
12 0					1.00	1.00
13 0	Keeping 31	90 of 3190 variables.	Ω	ariables	1.00	1.00
14 C F	ile name:			Save	1.00	1.00
15 C	1			0010	1.00	1.00
16 C S	ave as type: SPSS (*.sa	(V)	-	Paste	1.00	1.00
17 C				Cancel	1.00	1.00
18 0	🔽 💆 rite va	ariable names to sprea	adsheet	Cancel	1.00	1.00
19 C	☐ S <u>a</u> ve va	alue labels where defi	ned instead of data value		1.00	1.00
20 0	C Save va	alue labels into a sas			1.00	1.00
21 0 10	JZ W 1	10000000	01 21		1.00	1.00

🗰 ncds5cmi.sav - SPS	S Data Edito	1					
File Edit View Data Tr	ansform Anal	yze Graphs l	Jtilities W	indow	Help		
	× 🔚 🕼	<b>M</b>		<b>I</b>	0		
1 : serial	010001	IK					
serial p	erson n	622 5   n5r	egion	n5go	r	resps5 resp	5cmi
1 0'	1967				-		1.0
2 C : Save Data	As					2 🛛	1.0
			_			_	1.0
Save Data As: N	ariables						1.0
							1.0
Only variables mar	ked with an X w	vill be saved to th	e specified	data file	э.		1.0
Keep Name	Label			Order	~	Keep All	1.0
resp5cmi	Responde	d to Cohort Mem	ber Int	7			1.0
resp5wyt	Responde	d to What Do Yo	ou Thin	8		Drop All	1.0
resp5cl	Responde	d to Your Life Si	nce 19	9			1.0
🖾 n500518	Time of In	erview (mins)		10			1.0
M n500520	CMI:2,A1a	i) Current main ei	conomi	11		Continue	1.0
M n500522	CMI:2,A15	) WORKING-No	. jobs la	12		Continue	1.0
M n500524	CMI:2,A2a	JNUT WURKIN	G Any J	13		Cancel	1.0
	CAROCO:	INUT WURKIN	U-NO.J NT /I.A	14			1.0
	CASOC2.2	A4a/b) CURRE	NTZLA	10	~	Help	1.0
11241 11348011	CA3002.2			10			1.0
Selected: 3190 of	3190 variables.						1.0
<u>jaln</u>	1 Save valu	e labels where d	enned inste	ad or da	ita vai	Jes	1.0
20 0	🔽 Save valu	e labels into a .s	as file				1.0
21 01003277	1		0	_	4		1.0
22 010034A	1	2	8		2	1	1.0
2310100350	1	2	8		2	1	1 🛙

You can then check the boxes on the left next to each variable you want to keep (or 'uncheck' the ones you want to drop). Then click **Continue** and enter the name of the new data file: just the specified variables will be saved to this file.

This works where only a small number of variables need to be checked or unchecked, but can be tiresome for saving a large number which all appear consecutively. In this case, it's much better to use the syntax method explained above.