Taking into account missing data

Jouko Miettunen

Missing data

- People do not participate?
- People are lost to follow–up?
- Missing data on variables?
- Are these issues important?

Reporting attrition?

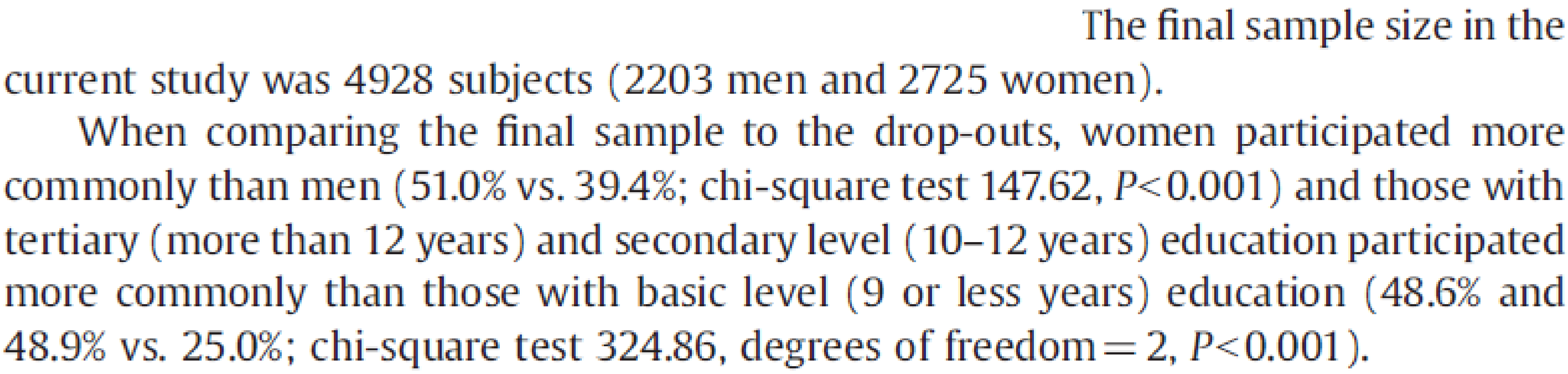
- Do you have any data from nonparticipants?
  - Baseline, register or clinical data?
- Have the previous studies on the sample presented results regarding attrition?
- You can report e.g. proportion of participants by some variables, e.g. gender, SES, education…
- Also continuous variables (e.g. hospital days) can be comparred among participants and nonparticipants

Examples on reporting

Miettunen J, Veijola J, Freimer N, Lichtermann D, Peltonen L, Paunio T, Isohanni M, Joukamaa M, Ekelund J.

Data on schizotypy and affective scales are gender and education dependent – study in the Northern Finland 1966 Birth Cohort.

Psychiatry Res 2010; 178:408–13.



The final sample size in the current study was 4928 subjects (2203 men and 2725 women).

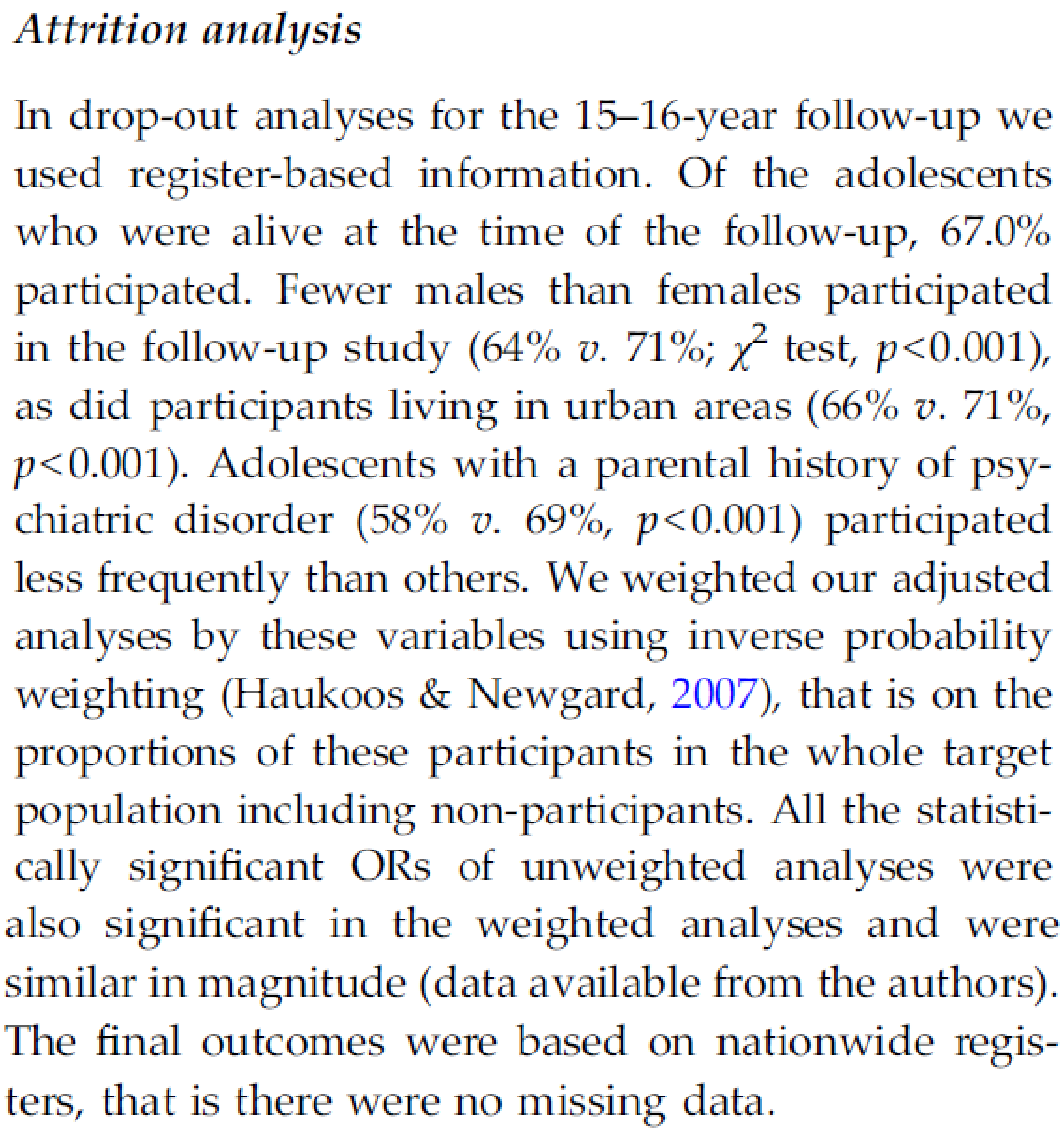
When comparing the final sample to the drop-outs, women participated more commonly than men (51.0% vs. 39.4%; chi-square test 147.62, $P<0.001$) and those with tertiary (more than 12 years) and secondary level (10–12 years) education participated more commonly than those with basic level (9 or less years) education (48.6% and 48.9% vs. 25.0%; chi-square test 324.86, degrees of freedom = 2, $P<0.001$).

Examples on reporting

Miettunen J, Murray GK, Jones PB, Mäki P, Ebeling H, Taanila A, Joukamaa M, Savolainen J, Törmänen S, Järvelin MR, Veijola J, Moilanen I.

Longitudinal associations between childhood and adulthood externalizing and internalizing psychopathology and adolescent substance use.

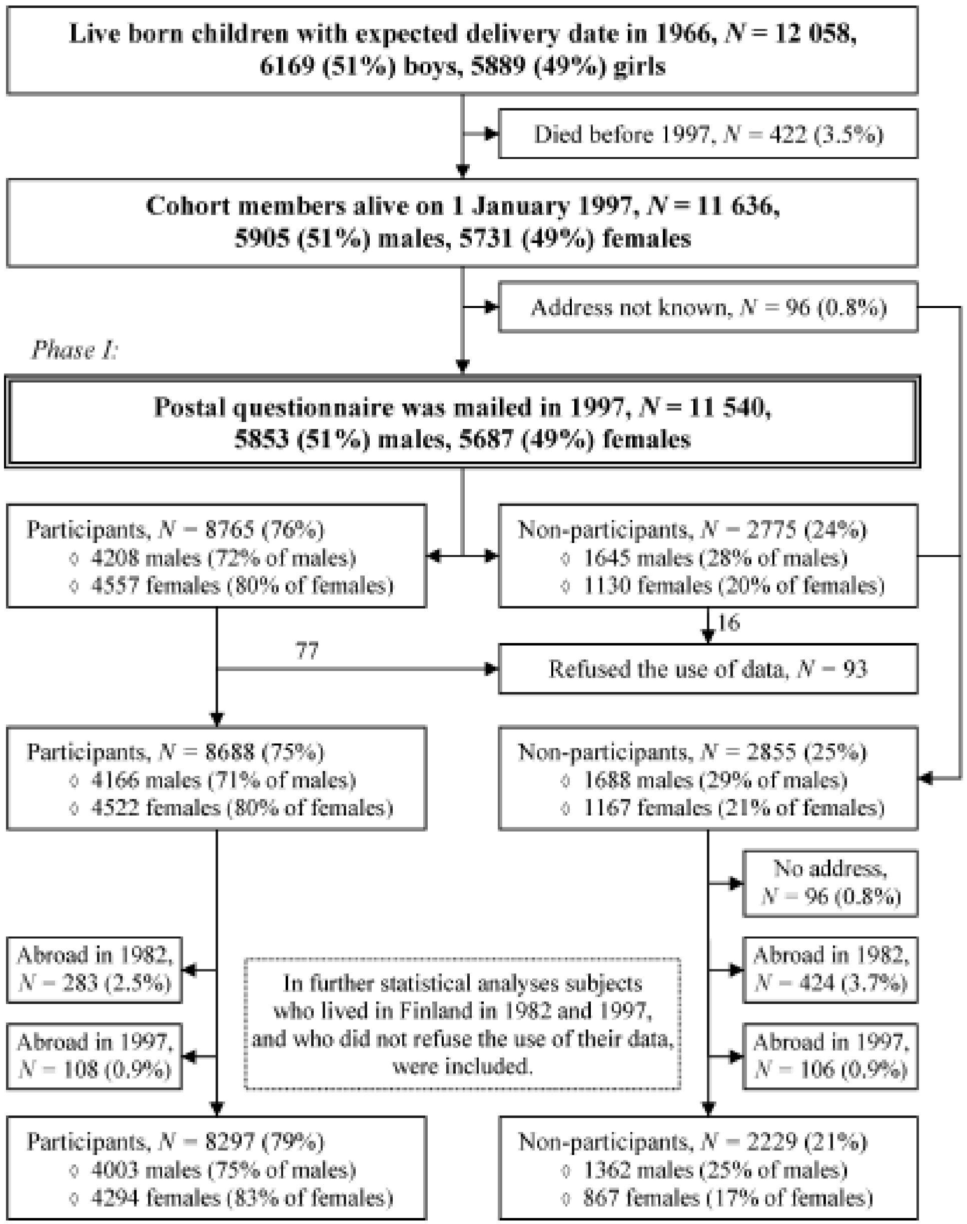
Psychol Med. 2014 Jun; 44(8):1727-38.



Attrition analysis

In drop-out analyses for the 15–16-year follow-up we used register-based information. Of the adolescents who were alive at the time of the follow-up, 67.0% participated. Fewer males than females participated in the follow-up study (64% *v.* 71%; χ^2 test, $p<0.001$), as did participants living in urban areas (66% *v.* 71%, $p<0.001$). Adolescents with a parental history of psychiatric disorder (58% *v.* 69%, $p<0.001$) participated less frequently than others. We weighted our adjusted analyses by these variables using inverse probability weighting (Haukoos & Newgard, 2007), that is on the proportions of these participants in the whole target population including non-participants. All the statistically significant ORs of unweighted analyses were also significant in the weighted analyses and were similar in magnitude (data available from the authors). The final outcomes were based on nationwide registers, that is there were no missing data.

Flowchart



Live born children with expected delivery date in 1966, *N* = 12 058, 6169 (51%) boys, 5889 (49%) girls

Died before 1997, *N* = 422 (3.5%)

Cohort members alive on 1 January 1997, *N* = 11 636, 5905 (51%) males, 5731 (49%) females

Address not known, *N* = 96 (0.8%)

Phase I:

Postal questionnaire was mailed in 1997, *N* = 11 540, 5853 (51%) males, 5687 (49%) females

Participants, *N* = 8765 (76%)
◊ 4208 males (72% of males)
◊ 4557 females (80% of females)

Non-participants, *N* = 2775 (24%)
◊ 1645 males (28% of males)
◊ 1130 females (20% of females)

77

16

Refused the use of data, *N* = 93

Participants, *N* = 8688 (75%)
◊ 4166 males (71% of males)
◊ 4522 females (80% of females)

Non-participants, *N* = 2855 (25%)
◊ 1688 males (29% of males)
◊ 1167 females (21% of females)

No address, *N* = 96 (0.8%)

Abroad in 1982, *N* = 283 (2.5%)

Abroad in 1997, *N* = 108 (0.9%)

In further statistical analyses subjects who lived in Finland in 1982 and 1997, and who did not refuse the use of their data, were included.

Abroad in 1982, *N* = 424 (3.7%)

Abroad in 1997, *N* = 106 (0.9%)

Participants, *N* = 8297 (79%)
◊ 4003 males (75% of males)
◊ 4294 females (83% of females)

Non-participants, *N* = 2229 (21%)
◊ 1362 males (25% of males)
◊ 867 females (17% of females)

Figure 1. Data collection of phase I (postal questionnaire) in a health survey conducted in 1997 in the Northern Finland 1966 Birth Cohort.

Northern Finland Birth Cohort 1966 – The 31y study attrition

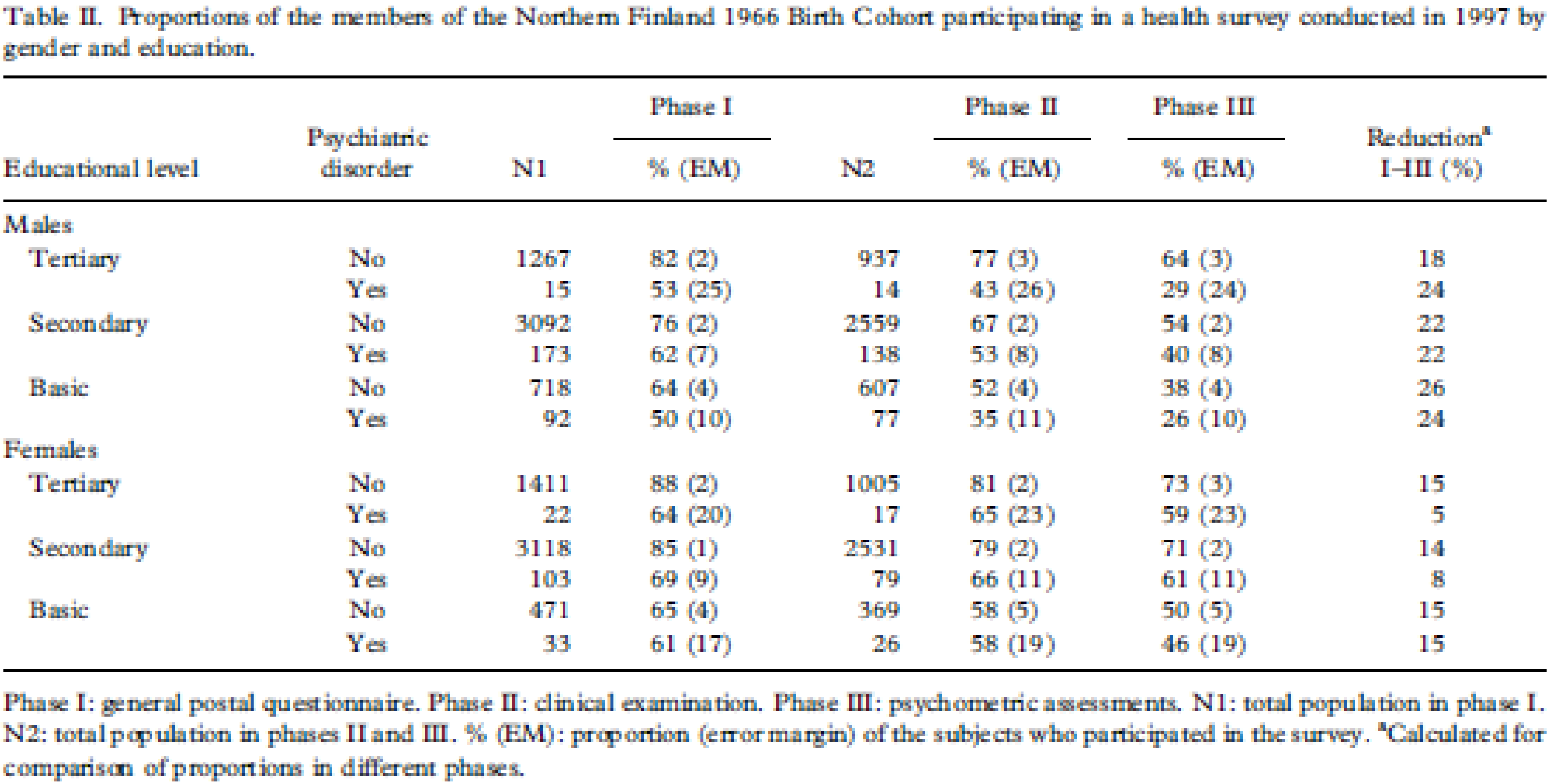


Table II. Proportions of the members of the Northern Finland 1966 Birth Cohort participating in a health survey conducted in 1997 by gender and education.

Educational level	Psychiatric disorder	N1	Phase I	N2	Phase II	Phase III	Reduction[a] I–III (%)
			% (EM)		% (EM)	% (EM)	
Males							
Tertiary	No	1267	82 (2)	937	77 (3)	64 (3)	18
	Yes	15	53 (25)	14	43 (26)	29 (24)	24
Secondary	No	3092	76 (2)	2559	67 (2)	54 (2)	22
	Yes	173	62 (7)	138	53 (8)	40 (8)	22
Basic	No	718	64 (4)	607	52 (4)	38 (4)	26
	Yes	92	50 (10)	77	35 (11)	26 (10)	24
Females							
Tertiary	No	1411	88 (2)	1005	81 (2)	73 (3)	15
	Yes	22	64 (20)	17	65 (23)	59 (23)	5
Secondary	No	3118	85 (1)	2531	79 (2)	71 (2)	14
	Yes	103	69 (9)	79	66 (11)	61 (11)	8
Basic	No	471	65 (4)	369	58 (5)	50 (5)	15
	Yes	33	61 (17)	26	58 (19)	46 (19)	15

Phase I: general postal questionnaire. Phase II: clinical examination. Phase III: psychometric assessments. N1: total population in phase I. N2: total population in phases II and III. % (EM): proportion (error margin) of the subjects who participated in the survey. [a]Calculated for comparison of proportions in different phases.

Haapea M, Miettunen J, Läärä E, Joukamaa MI, Järvelin M-R, Isohanni MK, Veijola JM. Non-participation in a field survey with respect to psychiatric disorders. Scand J Public Health 2008; 36: 728-36.

Haapea M, Veijola J, Tanskanen P, Jääskeläinen E, Isohanni M, Miettunen J. Use of inverse probability weighting to adjust for non-participation in estimating brain volumes in schizophrenia patients. Psychiatry Res Neuroim 2011; 194:326-32.

Use of inverse probability weighting to adjust for non-participation in estimating brain volumes in schizophrenia patients

Marianne Haapea [a,b,c,*], Juha Veijola [a,d], Päivikki Tanskanen [b], Erika Jääskeläinen [a,d], Matti Isohanni [a,d], Jouko Miettunen [a,d,e]

[a] *Department of Psychiatry, Institute of Clinical Medicine, University of Oulu, Oulu, Finland*
[b] *Department of Diagnostic Radiology, Oulu University Hospital, Oulu, Finland*
[c] *Department of Physiology, Institute of Biomedicine, and Biocenter Oulu, University of Oulu, Oulu, Finland*
[d] *Department of Psychiatry, Oulu University Hospital, Oulu, Finland*
[e] *Institute of Health Sciences, University of Oulu, Oulu, Finland*

ARTICLE INFO

Article history:
Received 20 October 2010
Received in revised form 13 May 2011
Accepted 3 June 2011

Keywords:
Missing data
Non-response bias
Unit non-response
Weighting methods

ABSTRACT

Low participation is a potential source of bias in population-based studies. This article presents use of inverse probability weighting (IPW) in adjusting for non-participation in estimation of brain volumes among subjects with schizophrenia. Altogether 101 schizophrenia subjects and 187 non-psychotic comparison subjects belonging to the Northern Finland 1966 Birth Cohort were invited to participate in a field study during 1999–2001. Volumes of grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) were compared between the 54 participating schizophrenia subjects and 100 comparison subjects. IPW by illness-related auxiliary variables did not affect the estimated GM and WM mean volumes, but increased the estimated CSF mean volume in schizophrenia subjects. When adjusted for intracranial volume and family history of psychosis, IPW led to smaller estimated GM and WM mean volumes. Especially IPW by a disability pension and a higher amount of hospitalisation due to psychosis had effect on estimated mean brain volumes. The IPW method can be used to improve estimates affected by non-participation by reflecting the true differences in the target population.

Weight cases (SPSS: Data → Weight cases)

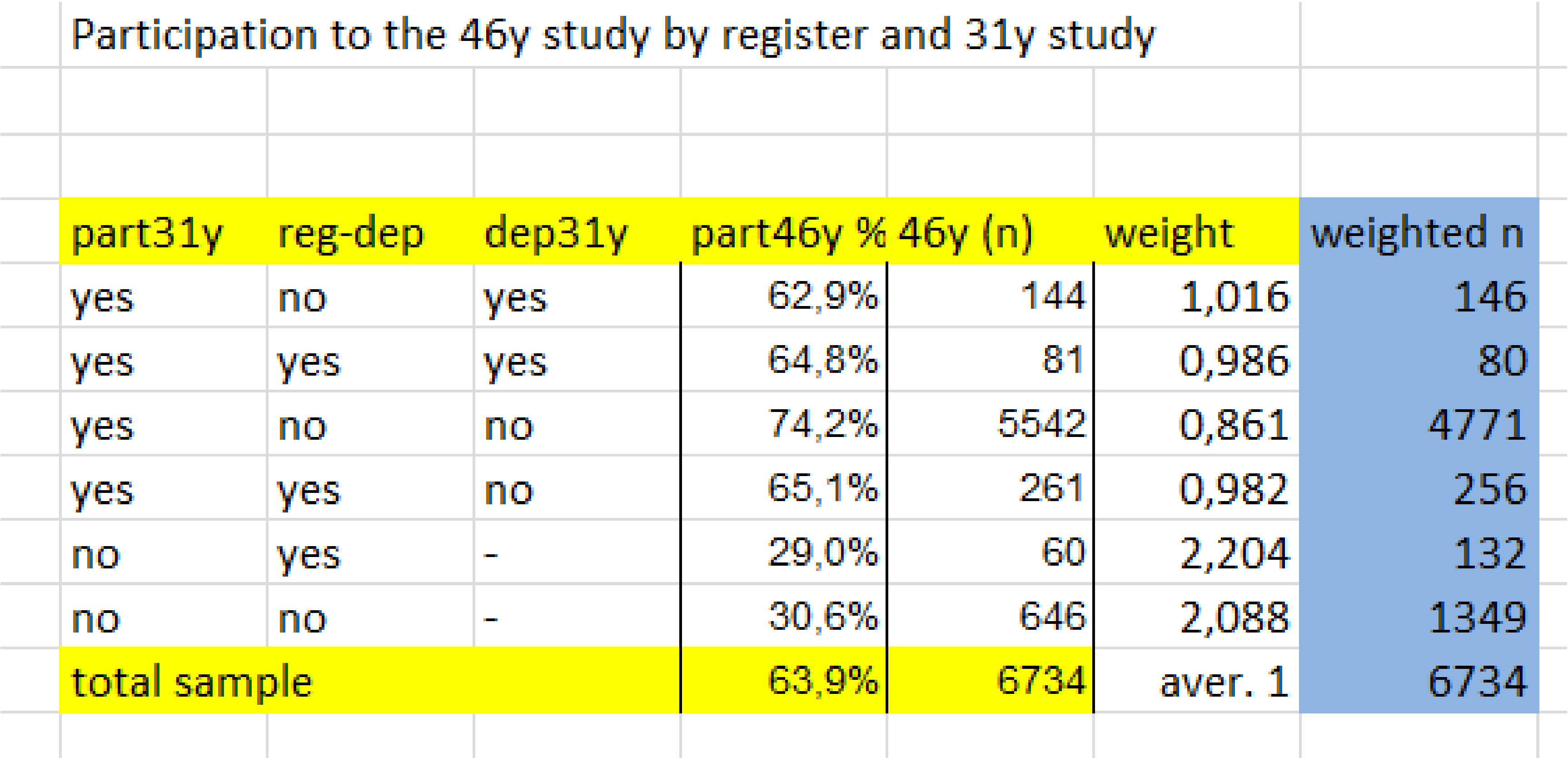
Weight Cases gives cases different weights (by simulated replication) for statistical analysis.

The values of the weighting variable should indicate the number of observations represented by single cases in your data file.

Some procedures ignore the weighting variable completely, and this limitation is noted in the procedure-specific documentation.

Once you apply a weight variable, it remains in effect until you select another weight variable or turn off weighting. If you save a weighted data file, weighting information is saved with the data file. You can turn off weighting at any time, even after the file has been saved in weighted form.

SPSS help

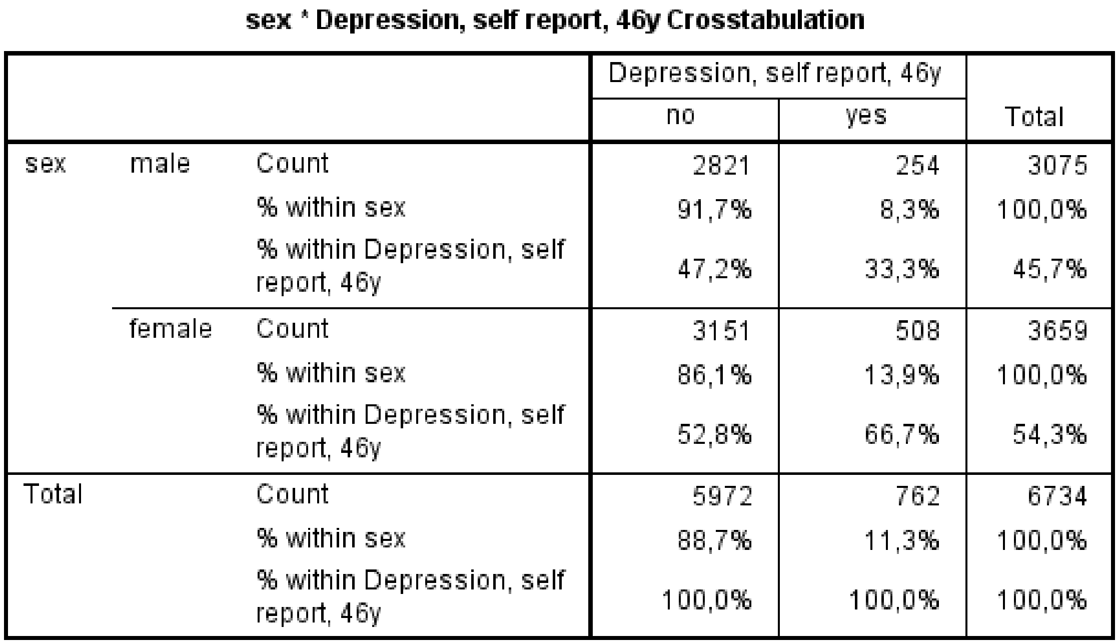


Participation to the 46y study by register and 31y study

part31y	reg-dep	dep31y	part46y %	46y (n)	weight	weighted n
yes	no	yes	62,9%	144	1,016	146
yes	yes	yes	64,8%	81	0,986	80
yes	no	no	74,2%	5542	0,861	4771
yes	yes	no	65,1%	261	0,982	256
no	yes	-	29,0%	60	2,204	132
no	no	-	30,6%	646	2,088	1349
total sample			63,9%	6734	aver. 1	6734

- Weights are calculated as ratios of total participation rate and participation rates in each group.
- There can be several variables included here, e.g. education.

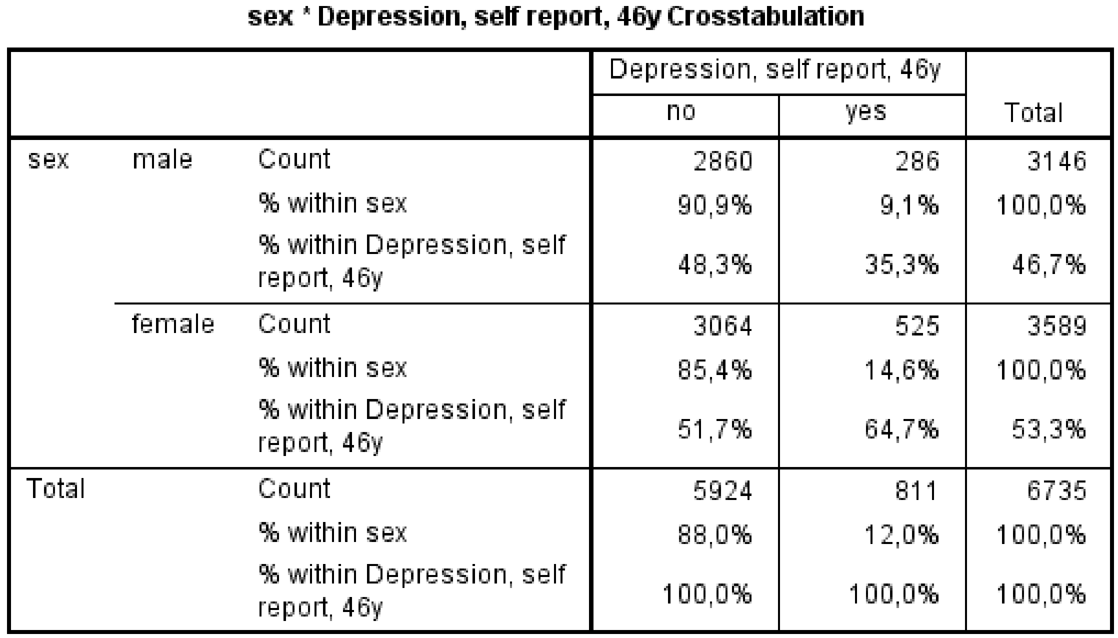
Unweighted



sex * Depression, self report, 46y Crosstabulation

			Depression, self report, 46y		Total
			no	yes	
sex	male	Count	2821	254	3075
		% within sex	91,7%	8,3%	100,0%
		% within Depression, self report, 46y	47,2%	33,3%	45,7%
	female	Count	3151	508	3659
		% within sex	86,1%	13,9%	100,0%
		% within Depression, self report, 46y	52,8%	66,7%	54,3%
Total		Count	5972	762	6734
		% within sex	88,7%	11,3%	100,0%
		% within Depression, self report, 46y	100,0%	100,0%	100,0%

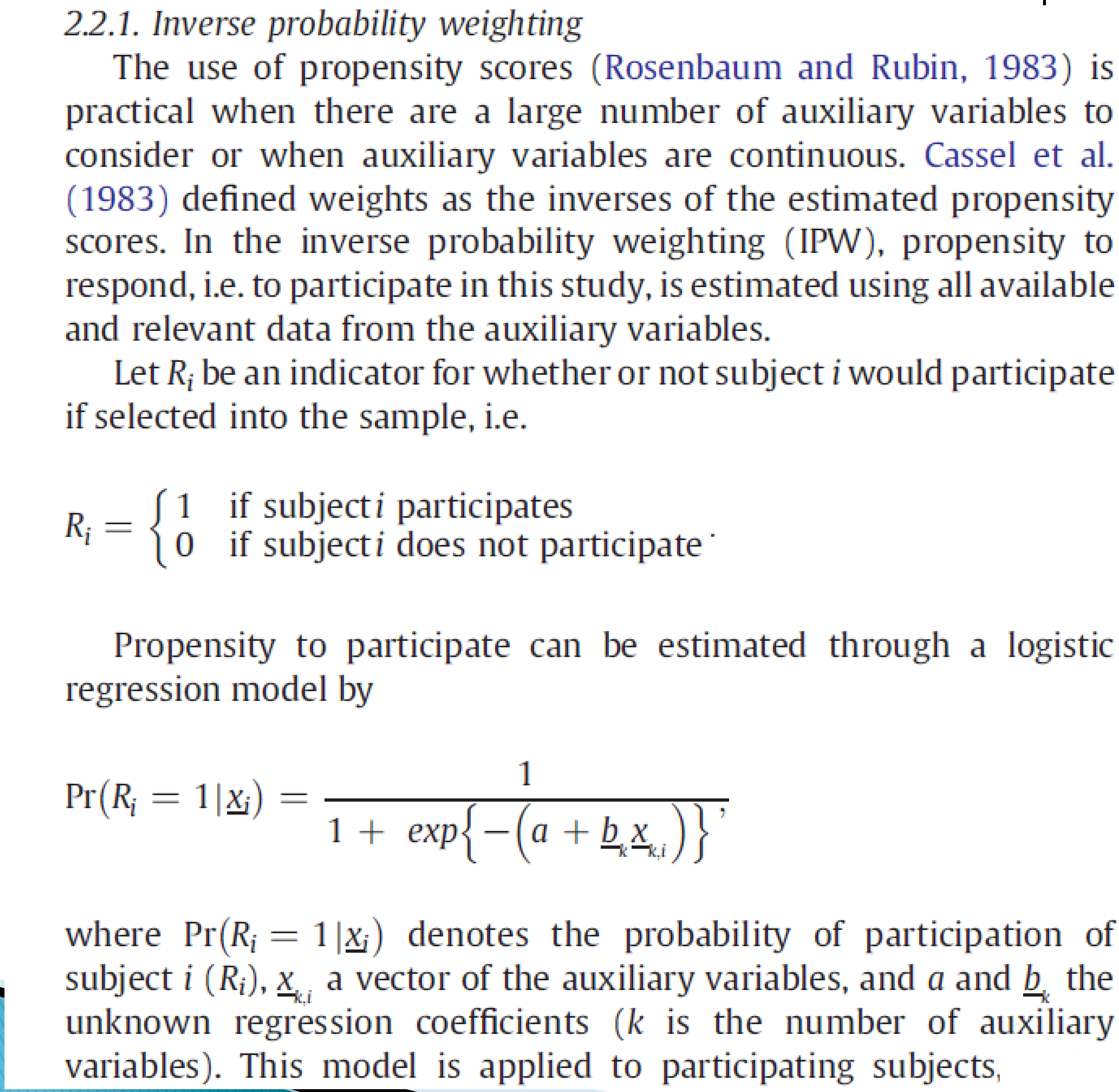
Weighted



sex * Depression, self report, 46y Crosstabulation

			Depression, self report, 46y		Total
			no	yes	
sex	male	Count	2860	286	3146
		% within sex	90,9%	9,1%	100,0%
		% within Depression, self report, 46y	48,3%	35,3%	46,7%
	female	Count	3064	525	3589
		% within sex	85,4%	14,6%	100,0%
		% within Depression, self report, 46y	51,7%	64,7%	53,3%
Total		Count	5924	811	6735
		% within sex	88,0%	12,0%	100,0%
		% within Depression, self report, 46y	100,0%	100,0%	100,0%

Haapea et al. 2011



2.2.1. Inverse probability weighting

The use of propensity scores (Rosenbaum and Rubin, 1983) is practical when there are a large number of auxiliary variables to consider or when auxiliary variables are continuous. Cassel et al. (1983) defined weights as the inverses of the estimated propensity scores. In the inverse probability weighting (IPW), propensity to respond, i.e. to participate in this study, is estimated using all available and relevant data from the auxiliary variables.

Let R_i be an indicator for whether or not subject i would participate if selected into the sample, i.e.

$$R_i = \begin{cases} 1 & \text{if subject } i \text{ participates} \\ 0 & \text{if subject } i \text{ does not participate} \end{cases}.$$

Propensity to participate can be estimated through a logistic regression model by

$$\Pr(R_i = 1 | \underline{x}_i) = \frac{1}{1 + exp\{-(a + \underline{b}_k \underline{x}_{k,i})\}},$$

where $\Pr(R_i = 1 | \underline{x}_i)$ denotes the probability of participation of subject i (R_i), $\underline{x}_{k,i}$ a vector of the auxiliary variables, and a and $\underline{b}_k$ the unknown regression coefficients (k is the number of auxiliary variables). This model is applied to participating subjects,

NORTHERN FINLAND BIRTH COHORT 1966

Self-reported diagnoses at 46 year									
	UNWEIGHTED					WEIGHTED WITH MARITAL AND WORK STATUS			
	Male		Female			Male		Female	
	N	%	N	%		N	%	N	%
Type II Diabetes	92	3,0	101	2,7		96	3,1	106	2,9
Epilepsy	35	1,1	58	1,6		41	1,3	62	1,7
Psychosis	41	1,3	48	1,3		55	1,8	65	1,8
Depression	256	8,2	519	14,0		289	9,3	565	15,3
Other mental health problem	100	3,2	181	4,9		124	4,0	195	5,3
Alcohol use problem	139	4,5	51	1,4		170	5,4	61	1,7
Hypertension, high BP	705	22,6	701	18,8		720	23,1	714	19,3

Attrition questions

- Are the good register data that could be used for weighting?
  - work, medications, hospitalisations, etc.
- Low number of participants per weighting group gives unreliable weight estimates
- Multiple imputation is also a good option, some data (e.g. 66% of variables) is needed for the method to be reliable

EXAMPLE DATA SET

This sample data set is an **anonymised**, **randomly-selected**, 1000 participant sample from the cohorts managed by the Northern Finland Birth Cohort Project Center. Any identifying information has been removed from the sample. Some variables have been recoded for purpose of this analysis and missingness of the data has been exaggerated.

Please respect the confidentiality of the data and delete all related files before you leave the room. Please ensure that you do not save or send any of the data provided today.