COMPARISON OF MISSING DATA METHODS USING REGISTER-BASED AUXILIARY DATA FOR HEALTH-RELATED SURVEY DATA OUTCOME

Oona Pentala, Tommi Härkänen, Risto Kaikkonen

THL

NATIONAL INSTITUTE FOR HEALTH AND WELFARE, FINLAND

Overview

Survey Data

- Regional Health and Wellbeing study (ATH) 2010
- National sample 5,000 + regional samples from Turku, Kainuu and Northern Ostrobothnia (N=31,000 altogether). The final data consisted of 14,799 observed cases.
- Stratified random sampling design
- Response rate varied from 37% to 65%, overall 49 %

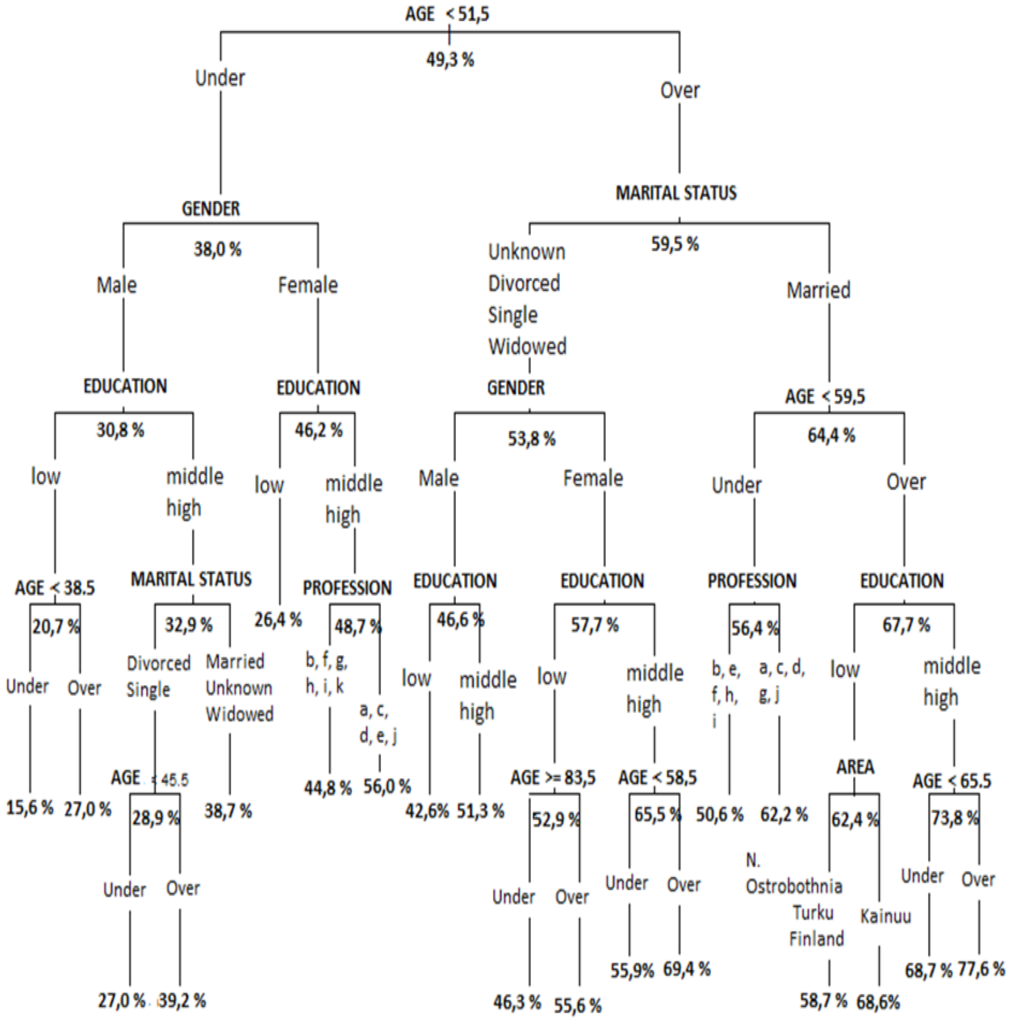
Register Data

- Age, gender, marital status, area, education, profession
- Special reimbursement of medication
- Compared with regional rates of people receiving reimbursement of depression medication

Outcome

- Self-reported diagnosis or treatment of depression: "Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months?"
- Overall item nonresponse was 9.4 %

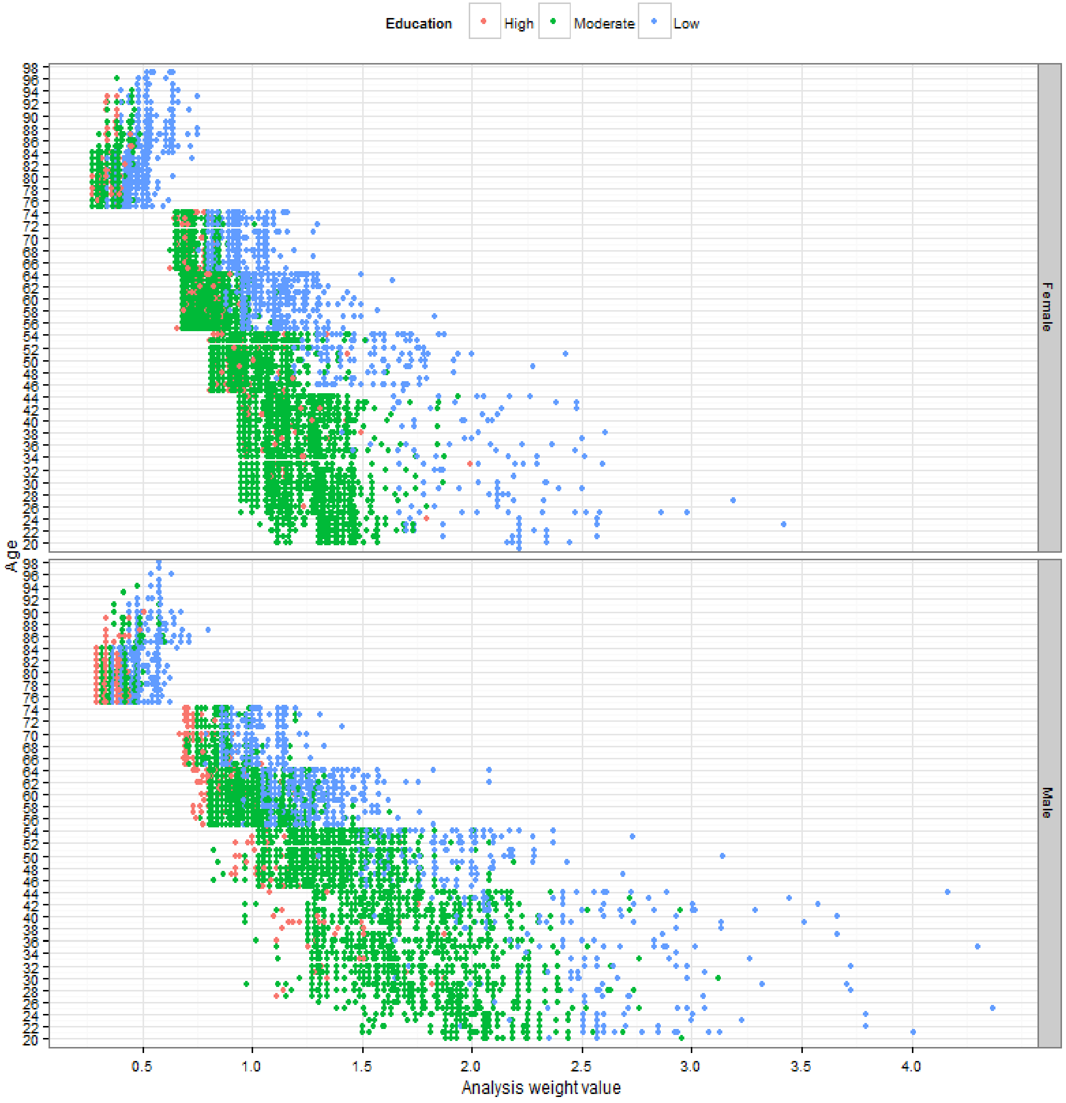
Unit nonresponse



Coefficients:	Estimate	Sig.
(Intercept)	-1.04	**
Agegoup (10 year)	0.25	***
Gender: Female	0.54	***
Area: Finland	-0.21	***
Area: Northern Ostrobothnia	-0.21	***
Area: Turku	-0.19	***
Marital status: Unknown	-0.6	**
Marital status: Divorced	-0.27	***
Marital status: Widowed	-0.49	***
Marital status: Single	-0.36	***
Language: Swedish	-0.04	
Language: Finnish	0.02	
Language: Russian	-0.29	
Profession: Unknown	-0.06	
Profession: Technicians	0.04	
Profession: Managers	-0.14	.
Profession: Agricultural workers	-0.27	**
Profession: Elementary workers	-0.25	***
Profession: Service and Sales workers	-0.15	**
Profession: Craft related trades workers	-0.22	**
Profession: Plan and machine operators	-0.39	***
Profession: Armed forces	0.51	*
Profession: Clerical support workers	-0.17	*
Eduction: High	0.33	***
Education: Low	-0.35	***
Gender:Female*education high	-0.32	**
Gender:Female*education basic	-0.37	***
Medicine reimbursement: Other	-0.24	**
Medicine reimbursement: Psyche	-0.51	***
Medicine reimbursement: Diabetes	-0.05	
Medicine reimbursement: Overall	0.1	**

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Inverse Probability Weighting (IPW)



- In IPW individuals are weighted by the inverse of their probability of being completely observed.
- The model for probability most often uses only completely observed variables (Seaman and White, 2013).
- IPW is applicable if the data are MAR, since then it is possible to obtain suitable estimates for the probabilities using for example logistic regression. (Molenberghs and Kenward, 2007).

- Best model found according to Bayesian Information Criterion was Main effects+Gender*Education.
- IPW used to handle both the differential sampling probabilities and missing data
- More on using IPW in ATH study in Härkänen et al: *Inverse probability weighting and doubly robust methods in correcting the effects of non-response in the reimbursed medication and self-reported turnout estimates in the ATH survey* (2014).

Item nonresponse & Answer wave

Table 1. Response rates of self-reported depression by answer wave

Answer wave	Percent	95% Confidence Limits	
Early	91.5	91.0	92.1
Late	87.8	86.9	88.8

Table 2. Prevalences of self-reported depression by answer wave

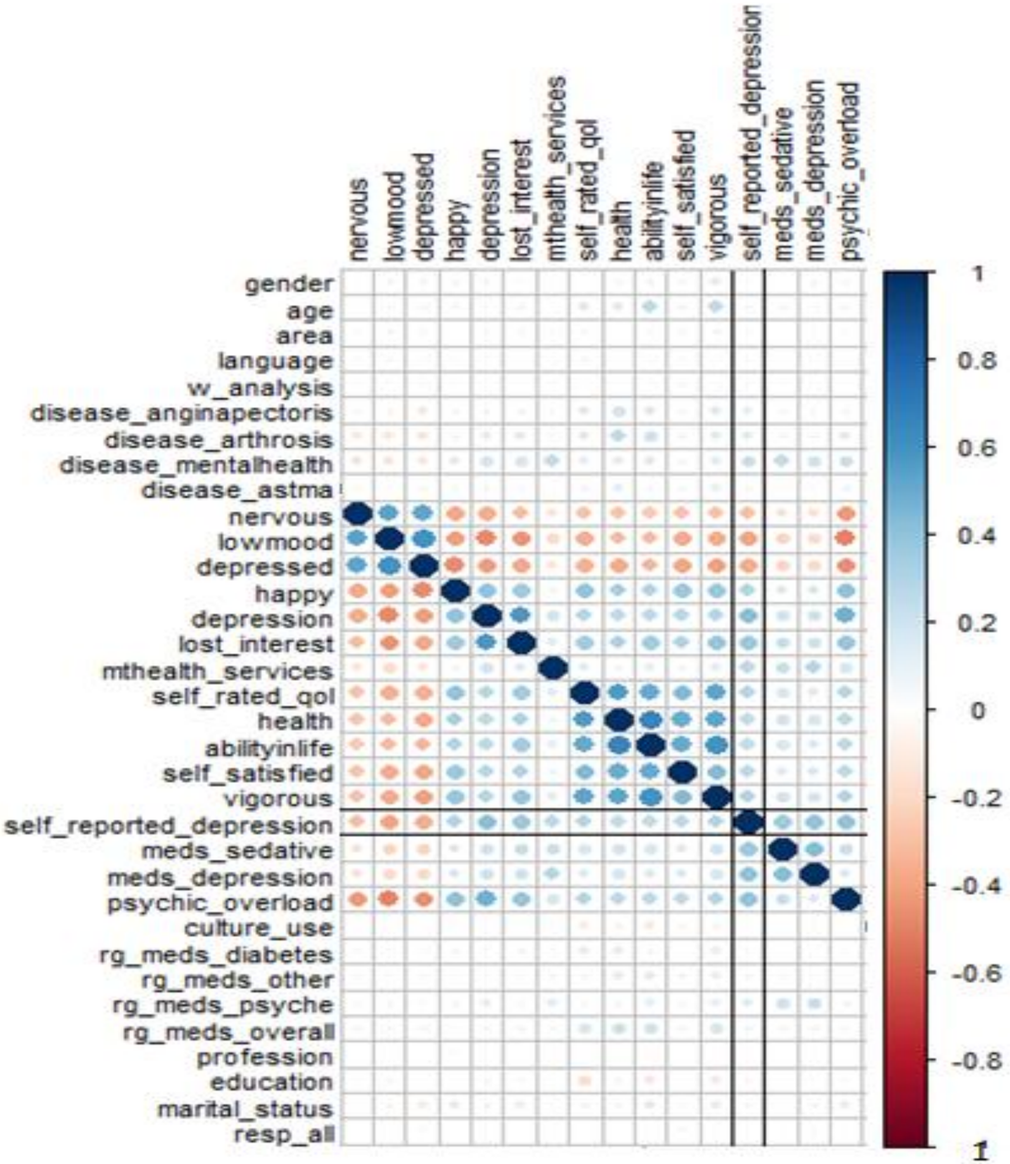
Answer wave	Percent	95% Confidence Limits	
Early	10.6	9.9	11.2
Late	13.1	12.0	14.1

Frequency Missing = 1417 which equals 9.4 % of the observed cases

Weighted Sequential Hot Deck Imputation

- WSHDI algorithm corrects the potential bias of simple hot deck imputation by using weighting technique.
- Sampling and/or analysis weights are used to restrict the number of times a respondent value can be used for imputation (Cox, 1980)
- Model was same as for IPW but also response wave was added to the model
- SUDAAN PROC HOTDECK was used, it can not handle missing values in donor determining variables
- Missingness in other than sample based predictors of depression associated highly with the missingness in the outcome variable

Outcome and its dependents



- Questionnaire items such as nervousness, feelings of low mood, depression, happiness, lost of interest on daily life, self-rated quality of life, satisfaction in health and self, use of mental health services and medication use correlate quite strongly (0,3) with depression.

Multiple Imputation

- Predictors were selected according to van Buuren and Groothuis-Oudshoorn (2011) as follows:
  - all the variables of the “complete data model” are included as predictors.
  - all the variables associated with the nonresponse, both unit and item nonresponse are included.
  - all the variables that explain the target variables variance are included. This is determined by examining the correlations with the target variables and choosing those exceeding certain correlation.
  - variables with too many missing values in different subgroups of incomplete cases are then removed from the model due to preserving the efficiency.
- The final predictor set contained 32 variables. Predictors with missing values were not imputed.
- The MI procedure was carried out with R Software package MICE and iterations were set to 20. Imputation method used was logistic regression with 5 multiple imputations.
- The final prevalence of self-reported depression is a weighted mean (with previously calculated IPW weights) of the prevalences in the imputed data sets.

Table 3. Original and imputed values of depression (0= no depression, 1= depression)

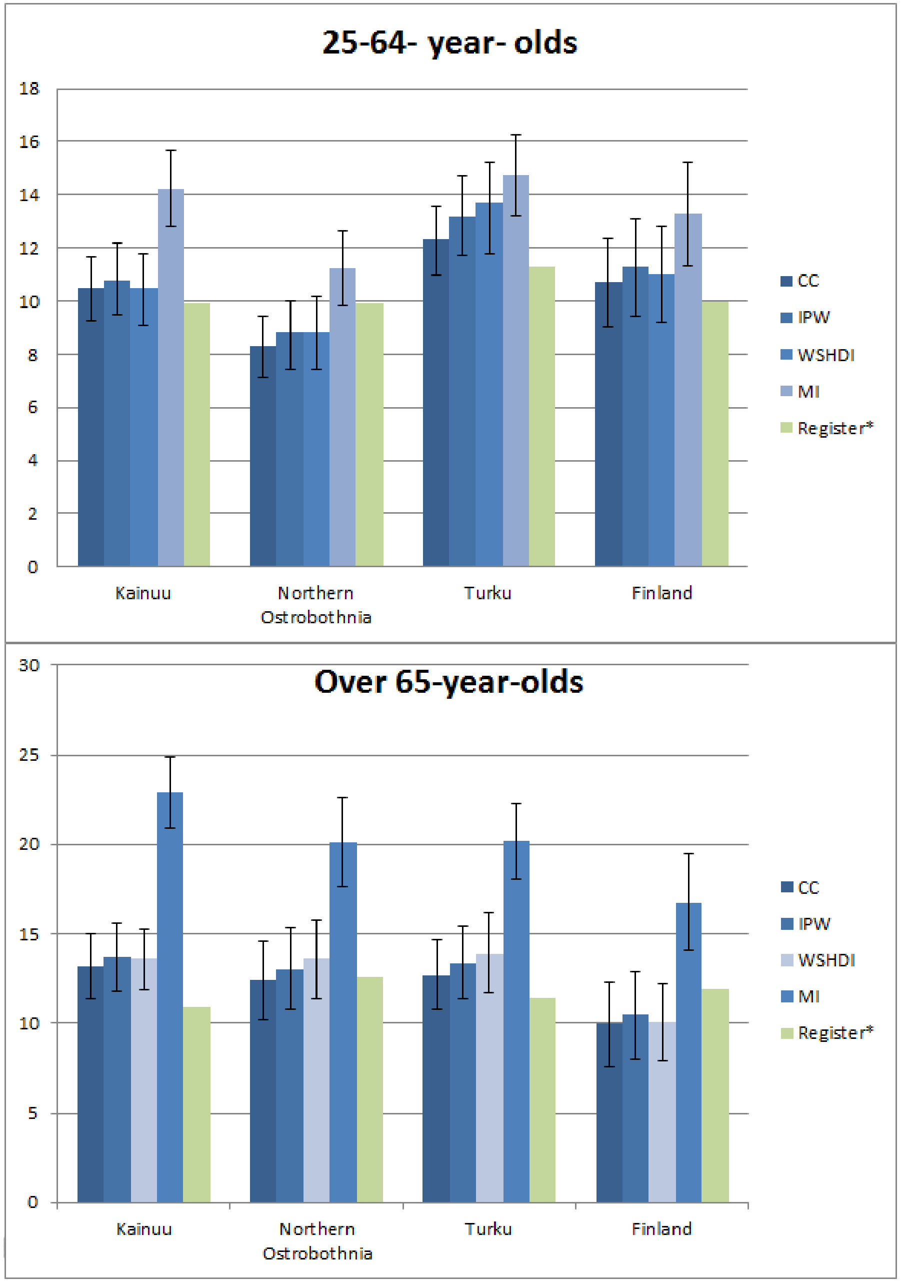
	agegroup 25-64			agegroup 65 and over		
original values	imputed values		Total	imputed values		Total
	0	1		0	1	
0	7435	0	7435	3729	0	3729
	100%	0	87%	100%	0	69%
1	0	869	869	0	579	579
	0	100%	**10%**	0	100%	**11%**
missing	32	252	284	78	1046	1124
	11%	89%	**3%**	7%	93%	**21%**

Results

Table 4. Odds ratio for self-reported depression

	CC	IPW	WSHDI	MI
Area Finland	1	1	1	1
Area Kainuu	1.08	1.04	1.05	**1.44**
Area Northern Ostrobothnia	0.86	0.84	0.89	1.02
Area Turku	1.21	1.22	1.32	1.06
Education high	1.00	1.00	1.00	1.00
Education low	1.27	1.51	**1.72**	**2.27**
Education intermediate	1.46	0.92	1.11	**0.83**
Single	1.00	1.00	1.00	1.00
Unknown	1.64	1.46	1.23	1.76
Married	1.17	**0.66**	**0.55**	**0.94**
Divorced	1.36	1.43	1.18	1.44
Widowed	1.22	1.38	1.33	**2.96**
Gender Male	1.00	1.00	1.00	1.00
Gender Female	1.19	1.22	1.01	**1.34**
20s	1.00	1.00	1.00	1.00
30s	0.76	0.99	0.91	0.93
40s	1.21	1.12	0.94	1.13
50s	1.10	**1.32**	1.07	**1.46**
60s	0.72	0.96	0.81	**1.60**
70s	0.95	1.29	1.12	**3.30**
80s	1.32	**2.04**	**1.74**	**5.26**
90s	0.87	**1.98**	**1.96**	**4.71**

Plot 2. Prevalences of self-reported depression



25-64- year- olds

Over 65-year-olds

Discussion

- Overall the results of self-reported depression prevalences are quite similar in CC analysis, IPW and WSHDI.
- All the methods seem to respect the relation between areas.
- Results indicate that CC analysis underestimates slightly the rate of depression in every area studied.
- MI rates stand out significantly especially in over 65-year-olds.
  - In MI also non-complete predictors were used to correct the nonresponse
  - These results may be biased and overestimate the depression rates in the older group since predictors included information about the mood in the last month, self-rated health and quality of life which are known to differ from the younger age group without signs of diagnosis-based depression (Saarela and Stenberg, 2011).
  - In this case the imputation model may not preserve the connection of age and depression and for further consideration it would be useful to construct predictor sets separatively for the age groups and take into account the interactions between age and the predictors.
- Register based rates only describe the situation from another point of view which is the reimbursement of depression medication.
- Overall the methods’ results also suggest that the willingness to answer the depression question depends on depression itself which means that the data is MNAR and the nonresponse corrected estimates may still be biased.

Thank you for listening, any questions?

References: ATH study, results & statistics

- Kaikkonen R., Murto J., Pentala O., Koskela T., Virtala E., Härkänen T., Koskenniemi T., Ahonen J., Vartiainen E. & Koskinen S. *Results of Regional Health and Well-being Study 2010-2014*. Internet publication: www.thl.fi/ath. 2010-2014.
- Härkänen T., Kaikkonen R., Virtala E. and Koskinen S. *Inverse probability weighting and doubly robust methods in correcting the effects of non-response in the reimbursed medication and self-reported turnout estimates in the ATH survey*. doi:10.1186/1471-2458-14-1150. BMC Public Health 2014, 14:1150.
- Pentala O. *Väestötutkimusaineiston tilastolliset kadonhallintamenetelmät - Alueellisen terveys- ja hyvinvointitutkimuksen kyselyaineisto 2010 (Statistical Nonresponse Methods in Population Surveys – Regional Health and Wellbeing study 2010)*. Master’s Thesis. University of Helsinki. http://hdl.handle.net/10138/144287, 2014.

References: statistics and software

- S. Seaman and I. White. *Review of inverse probability weighting for dealing with missing data*. Statistical Methods in Medical Research, 22(3):278–295, 2013. doi: 10.1177/0962280210395740.
- G. Molenberghs and M. Kenward. *Missing Data in Clinical Studies*. Wiley, 2007.
- B. Cox. *The weighted sequential hot deck imputation procedure.* ASA Proc Section on Survey Res Methods, pages 721–726, 1980. http://www.amstat.org/sections/srms/proceedings/papers/1980_152.pdf.
- S. van Buuren and K. Groothuis-Oudshoorn. *mice: Multivariate imputation by chained equations in r.* Journal of Statistical Software, 45(3):1–67, 2011. URL http://www.jstatsoft.org/v45/i03/.

References: depression in surveys

- M. Haapea. *Non-response and information bias in population-based psychiatric research.* The Northern Finland 1966 Birth Cohort Study. PhD thesis, University of Oulu, 2010.
- T. Saarela and J. Stenberg. *Kun mikään ei kelpaa vanhukselle - taustalla persoonallisuushäiriö?* Lääketieteellinen Aikakauskirja Duodecim, 127(4):397–405, 2011.