Supplement to "Systematic handling of missing data in complex study designs - experiences from the Health 2000 and 2011 Surveys"

Tommi Härkänen[a,*] Juha Karvanen[b†] Hanna Tolonen[c‡]
Risto Lehtonen[d§] Kari Djerf[e¶] Teppo Juntunen[f‖]
Seppo Koskinen[g**]

February 1, 2016

A Data and sampling designs

A.1 The Health 2000 Survey

The Health 2000 Survey was a national health examination survey carried out in 2000–2001. For this survey, the target population encompassed persons aged 18 years or older living in mainland Finland on July 1, 2000.

The design was a stratified two-stage cluster sampling design. 20 geographical strata were based on the 15 largest towns, while the rest of continental Finland was divided into its 5 university hospital regions.

A total of 80 health center districts (HCD) were selected from a sample, including the 15 largest towns with probability 1, and a systematic probability proportional to size (PPS) sampling of smaller HCDs as the primary sampling units (PSUs), in such a way that the sample contained 16 HCDs in each university hospital region.

[*a,*]Corresponding author; National Institute for Health and Welfare, P.O.Box 30, FI-00271 Helsinki, Finland, Email: Tommi.Harkanen@thl.fi, Tel.: +358295248719

[†b]University of Jyväskylä, P.O.Box 35, FI-40014 Jyväskylä, Finland, Email: juha.karvanen@jyu.fi, Tel.: +358295248719

[‡c]National Institute for Health and Welfare, P.O.Box 30, FI-00271 Helsinki, Finland, Email: hanna.tolonen@thl.fi, Tel.: +358295248638

[§d]University of Helsinki, P.O.Box 68, FI-00014 Helsinki, Finland, Email: risto.lehtonen@helsinki.fi, Tel.: +358919151404

[¶e]Statistics Finland, Työpajankatu 13, FI-00580 Helsinki, Finland, Email: kari.djerf@stat.fi, Tel.: +358917343425

[‖f]National Institute for Health and Welfare, P.O.Box 30, FI-00271 Helsinki, Finland, Email: teppo.juntunen@thl.fi, Tel.: +358295247215

[**g]National Institute for Health and Welfare, P.O.Box 30, FI-00271 Helsinki, Finland, Email: seppo.koskinen@thl.fi, Tel.: +358295248762

Systematic sampling of persons was applied so that the sample size in each stratum was proportional to the corresponding population base. Element-level sampling was used in the case of the 15 largest towns so that the PSUs were persons (one-stage sampling). The sample sizes in the 65 smaller HCDs were equal within each stratum (two-stage sampling), thus the sampling design was based on Equal Probability Selection Method (EPSEM). Oversampling of persons aged 80 and older was applied using double inclusion probabilities. The total sample size was 9,922. For further details, see (8).

For details on participation in the Health 2000 Survey, see (7).

Because the interviews and questionnaires were available in Finnish and Swedish, most immigrants who were not fluent in these official languages were likely to drop out of the survey. A separate health examination study has been conducted on the immigrant population (4).

A.2 The Health 2011 Survey

In addition to the old sample, a new sample of 1,994 young adults aged 18 to 28 years were sampled as the original sample had aged by 11 years during the follow-up. The HCDs changed between 2000 and 2011, due to administrative changes e.g. amalgamation of municipalities, and breaking up and formation of various HCDs. Some of the new amalgamations were both geographically and in terms of the population size large, and the original sampling probabilities of the HCDs would not represent the possible samples from the new municipalities well, thus we decided to select the new sample of young adults using the original HCDs in 2000. In the largest amalgamated municipalities the traveling distances to the research localities would have been long, which could have increased nonparticipation. The subsetting of the amalgamated municipalities required special attention, because 65 HCDs were sampled using PPS based on the municipal boundaries and populations. Two HCDs were amalgamated with others, which were also included in the Health 2000 Survey (Pyhäselkä with Joensuu, and Perniö with Salo). In three cases, old HCDs were split, and some of the municipalities were amalgamated with another municipality (Piikkiö with Kaarina) or with a larger town (Korpilahti with Jyväskylä, and Savonranta with Savonlinna). In these cases a large sample was taken from the new town, and geographic coordinates were utilized in selecting an appropriate subset of study subjects within the amalgamated municipalities. However, sampling based on actual boundaries was not possible. On the other hand, an overlap or underlap is possible between the municipal boundaries and those that were approximated, but we considered this to represent minor differences, since underlap or overlap areas near boundaries were sparsely populated in most cases. Based on the municipal boundaries in 2000, population sizes in 2011 were obtained from Statistics Finland.

B Statistical methods

B.1 Comparison of missing data methods

B.1.1 Weighted analysis

The complete case analysis, i.e. the removal of incomplete records, is probably the method most commonly applied to missing data. A weighted analysis is a modification of the complete case analysis, whereby the complete observations are weighted unequally in order to mitigate bias. These weights are obtained by multiplying the sampling weights, based on the sampling probabilities, by the inverse of the participation probabilities. In the literature, approaches of this kind of are referred to as inverse probability weighting (IPW) (3; 15; 18) or propensity score weighting (16; 9).

The expansion weights, which account for both the inclusion probabilities and non-participation, are obtained by calibrating the inverse probability weights v_{1i} based on (9):

$$v_{1i} := R_{1i}/\mathbb{P}\{\, R_{1i} = 1 \mid V_{1i}^*,\, R_{0i},\, X_{0i}^* \,\}\,. \tag{I}$$

Note that v_{1i} equal zero for those who did not participate in 2011, and $v_{1i} = 1 + \exp\{-\alpha_1 V_{1i}^* - \beta X_{0i}^*\}$ for those who participated in 2011. It should also be noted that the term $\exp\{-\alpha_1 V_{1i}^* - \beta X_{0i}^*\}$ can be interpreted as the number of individuals, who did not participate and who were assumed to be similar to the i^{th} individual, who participated in the survey, in terms of data nodes V_{1i}^* and X_{0i}^*.

A weighted analysis of variable Y with the weights (I) is unbiased if the quasi-randomization assumption $p(\, Y_i \mid R_{1i},\, V_{1i},\, R_{0i},\, X_{0i}\,) = p(\, Y_i \mid V_{1i},\, R_{0i},\, X_{0i}\,)$ holds, i.e., Y is independent of the participation status in the terms set by the observed covariates (10; 13). In many cases, a weighted analysis can be performed using standard software but unbiased variance estimates are difficult to obtain analytically. These weights are random variables because the non-participation probabilities are estimated based on the data. Bootstrapping is recommended with respect to the variance estimation (10).

The participants of the Health 2000 Survey had item nonresponse, which was handled using the single imputation method implemented in the `transcan` function with default settings of the R package Hmisc (5), in order to calculate the estimated participation probabilities based on equation 6. An excerpt from the documentation of `transcan`: "*`transcan` is a nonlinear additive transformation and imputation function, and there are several functions for using and operating on its results. `transcan` automatically transforms continuous and categorical variables to have maximum correlation with the best linear combination of the other variables. . . . Continuous variables are expanded as restricted cubic splines and categorical variables are expanded as contrasts (e.g., dummy variables). . . . By default, `transcan` imputes NAs with 'best guess' expected values of transformed variables, back transformed to the original scale. Values thus imputed are most like conditional medians assuming the transformations* *make variables' distributions symmetric (imputed values are similar to conditionl modes for categorical variables)."*

It should be noted that population distributions estimates based on the weights v_{1i} in (I) are not necessarily unbiased. This can occur if a parametric method such as a logistic regression model was applied when estimating the participation probabilities. We therefore perform calibration based on post-stratification, in which the weights are further adjusted according to the realized sample. In Health 2011, post-stratification can be applied to adjusting the age distribution to correspond to the age distribution within the population. Because the sampling fraction varied between age groups, the calibration was conducted by age groups $\mathrm{AG}_i \in \{[18,29), [29,50), [50,75), [75,80), [80,110)\}$ and participation in 2000 R_{0i}. The population sizes (excluding immigration after year 2000) in each poststratum $\{i : \mathrm{AG}_i = k_1, R_{0i} = k_2\}$ for 2011 were approximated by $N_{1,\mathrm{AG}=k_1,R_{0i}=k_2} := \sum_{\ell:\, \mathrm{AG}_\ell=k_1, R_{0\ell}=k_2} r_{1\ell} w_{0i}$. The sum of weights in each poststratum was calibrated to equal the corresponding (approximated) population size:

$$w_{1i} := v_{1i} \frac{N_{1,\mathrm{AG}_i,R_{0i}}}{\sum_{\ell:\, \mathrm{AG}_\ell=\mathrm{AG}_i, R_{0\ell}=R_{0i}} v_{i\ell}} \tag{II}$$

The variance estimation was based on linerization and the survey package (11). As a comparison, prevalences were estimated also ignoring the clustering and participation probabilities, and denoted by SRS in the results.

B.1.2 Regression estimation with weights: doubly robust methods and propensity scores

So-called doubly robust methods introduced by Bang and Robins (2) are based on two models. The analysis model models the outcome variable, while the weighting model models the participation indicator using a (logistic) regression model. It has been demonstrated that, in order to obtain unbiased results, it is sufficient if one of the aforementioned models is correct. This methodology has also been applied when adjusting for missing data by Wirth *et al.* (23), as we have done in this work. We have used the same covariates, for both the outcome and weighting models, as in the IPW model above.

Kang and Schafer (6) conducted a simulation study to compare different doubly robust methods; the conclusion was that regression estimation using propensity-based covariates is less vulnerable, in cases where both weighting and analysis models are wrong.

B.1.3 Multiple imputation

In multiple imputation (MI), the missing data values are imputed using a predictive distribution, which is based on the observed data and possible prior information (17; 19). Such an imputation model can differ from the analysis model, which is applied to the imputed data. In general, because imputed values are characterized by a considerable degree of uncertainty, a single imputation would underestimate the uncertainty of the results. For this reason, several copies of the original dataset are created in MI, with both the imputation procedure and statistical analyses being performed separately for each of them. Finally, the results are combined based on the imputed datasets.

In a typical item non-response case the analysis variables cannot be ordered into a monotonic missing data pattern. This restricts the number of applicable MI methods. In standard statistical software packages it is assumed that the variables, which contain missing values, follow a multinormal distribution, which is not well suited to categorical or other non-Gaussian variables. However, binary variables are often approximated through normal distribution in MI. Categorical and continuous variables can be imputed using multivariate imputation, based on the chained equations available in many statistical software packages.

In this case, MI was performed using the chained equations approach implemented as a package `mice` (22; 21) in the R software (14). In the three register-based outcome variables, in which the observed values were removed before imputation for those individuals who did not participate in the Health 2011 Survey. Three imputation models using the predictive mean matching (PMM) method were applied. The first model (MI1) contained only categorical age, gender, language and education. In addition to the contents of the first model, the second model (MI2) contained the same variables as the weighting model: self-reported health (SRH), work ability and participation frequency in clubs or associations. The third model (MI3) also contained some biological risk factors body mass index (BMI), systolic blood pressure (SBP) and smoking measured at baseline. In addition to the PMM method, also the recursive partitioning and regression trees ((20), option `cart` in package `mice`) were applied in the MI models MI4 containing the same variables as MI3, and MI5 containing a large number of variables using the Akaike information criterion (AIC, 1). Table I presents the additional variables and the corresponding AIC values. 50 imputed datasets were created using 15 iterations. No weights were used in the selection of the MI models or in the MI procedures.

The imputed data was then analyzed separately for each imputed data set using the `survey` package (11), applying the baseline sampling design and calibration weights when handling the unit non-response at the baseline. The results were finally combined using the `mitools` (12) package of R, in which the variance estimates were based on the classical results (17).

Table I: Variables selected for a large imputation model using the AIC and BIC.

Label	AIC	BIC
Time use: read books, listen music	5134	5560
BMI	5121	5428
Marital status	5121	5540
Self-assessed change of work ability (past 12 months)	5120	5453
Time use: meeting relatives, friends or neighbours	5118	5411
university hospital district (childhood)	5117	5490

physical exercise at least 30 min. (leisure time)	5116	5423
Eyeglasses or other visual aid	5115	5382
Smoking, ever	5115	5361
urbalization level of childhood home municipality	5114	5407
Previously married	5113	5360
physical activity	5113	5373
Time use: read newspapers or magazines	5113	5380
university hospital district (adulthood)	5112	5392
Time use: travelling abroad	5112	5379
urbalization level of adulthood home municipality	5112	5378
Main occupation	5112	5398
Time use: restaurants, bars, clubs, dancing	5110	5404
Time use: inviting relatives, friends etc. to visit	5109	5376
Self-rated health (5 classes)	5109	5403
frequency of alcohol use	5109	5402
Self-rated health (good)	5107	5354
Fresh vegetables excl. potatoes, per week	5107	5367
Self-rated oral health	5106	5372
physical exercise (leisure time)	5106	5365
language (Finnish or Swedish)	5105	5352
Self-rated health status (10-class VAS)	5104	5351
Time use: handicrafts, play music, hobbies	5104	5370
work ability score	5100	5354
Self-reported work ability	5100	5353
Time use: theater, movies, etc.	5096	5389
Time use: clubs or associations	5083	5350

C Results

In 2000, participants appeared to have higher hospitalization prevalences than non-participants under 65 years, but prevalences were lower among older participants (Table II). In 2011 these differences appeared only in those older than 65. Non-participants had higher disability pension prevalences in 2011.

Table II: *Register data prevalences (%), participants vs. non-participants $R_{\cdot i}$ in the age group '30 years and older'.*

Year	Age	Gender	$R_{\cdot i}$	Disability pension (%)	SE	Hospitalization (%)	SE	Reimbursement (%)	SE
2000	[30,45)	Female	0	1.4	1.4	9.6	3.5	12.3	3.9
2000	[30,45)	Female	1	1.7	0.4	17.6	1.1	13.5	1.0
2000	[30,45)	Male	0	1.9	1.4	5.8	2.3	10.6	3.0
2000	[30,45)	Male	1	2.9	0.5	8.1	0.8	8.7	0.8
2000	[45,65)	Female	0	16.4	4.8	4.9	2.8	26.2	5.7
2000	[45,65)	Female	1	10.7	0.8	13.4	0.9	29.0	1.2

2000	[45,65)	Male	0	14.6	3.9	4.9	2.4	28.0	5.0
2000	[45,65)	Male	1	13.3	0.9	14.1	1.0	25.4	1.2
2000	[65,75)	Female	0	0.0	0.0	25.0	11.2	37.5	12.5
2000	[65,75)	Female	1	2.0	0.7	18.6	1.9	57.2	2.5
2000	[65,75)	Male	0	0.0	0.0	0.0	0.0	62.5	18.3
2000	[65,75)	Male	1	2.6	1.0	16.2	2.4	56.0	3.3
2000	[75,Inf)	Female	0	0.0	0.0	37.5	18.3	50.0	18.9
2000	[75,Inf)	Female	1	0.6	0.6	24.0	3.2	62.3	3.7
2000	[75,Inf)	Male	1	0.0	0.0	35.0	6.2	55.0	6.5
2011	[30,45)	Female	0	4.5	2.6	7.6	3.3	19.7	4.9
2011	[30,45)	Female	1	2.6	1.0	12.7	2.0	15.3	2.2
2011	[30,45)	Male	0	4.1	2.0	9.2	2.9	10.2	3.1
2011	[30,45)	Male	1	5.7	1.6	7.7	1.8	18.7	2.7
2011	[45,65)	Female	0	17.7	2.1	12.9	1.8	37.8	2.7
2011	[45,65)	Female	1	12.0	0.9	14.1	0.9	32.4	1.2
2011	[45,65)	Male	0	18.4	1.8	13.6	1.6	34.8	2.2
2011	[45,65)	Male	1	12.1	0.9	11.6	0.9	29.1	1.3
2011	[65,75)	Female	0	10.9	3.0	21.8	4.0	63.6	4.6
2011	[65,75)	Female	1	7.6	1.1	18.4	1.6	52.3	2.1
2011	[65,75)	Male	0	17.1	3.6	27.0	4.2	60.4	4.7
2011	[65,75)	Male	1	8.2	1.3	18.6	1.9	54.6	2.4
2011	[75,Inf)	Female	0	0.4	0.4	38.8	3.1	76.5	2.7
2011	[75,Inf)	Female	1	0.0	0.0	32.3	2.4	75.8	2.2
2011	[75,Inf)	Male	0	0.0	0.0	42.4	5.2	83.7	3.9
2011	[75,Inf)	Male	1	0.0	0.0	35.7	3.0	77.1	2.7

The participation rate was very low among men with low educational attainment and amongst the '29 to 34 years' age group as shown by the OR estimates of the main effects (Table III). Poor self-reported work ability, language other than Finnish, and either no or high activity in clubs or associations also decreased participation in 2011.

Table III: *Odds ratio estimates of the baseline predictors and the corresponding p-values using participation in the Health 2011 Survey as the outcome in multivariate analysis. Selection of the predictors was based on the BIC. This analysis was based on participants in the Health 2000 Survey, for which predictors were available.*

Parameter	OR	Pr(>|z|)
Intercept	0.24	0.00
Age group 35-44	2.07	0.09
Age group 45-54	7.80	0.00
Age group 55-64	9.22	0.00
Age group 65-74	14.91	0.00
Age group 75-101	9.19	0.00
Gender female	5.59	0.00
Education secondary	5.49	0.00
Education tertiary	9.32	0.00
Self-reported work ability partially disabled	0.77	0.01

Self-reported work ability disabled	0.44	0.00
language Swedish	0.50	0.00
Self-rated health status (10-class VAS)	1.04	0.10
Time use: clubs or associations Once or twice a year	1.39	0.00
Time use: clubs or associations Once or twice a month	1.72	0.00
Time use: clubs or associations Once or twice a week	1.57	0.00
Time use: clubs or associations Every day or during most days	0.45	0.00
Age group 35-44 : Gender female	0.41	0.20
Age group 45-54 : Gender female	0.21	0.02
Age group 55-64 : Gender female	0.23	0.02
Age group 65-74 : Gender female	0.24	0.02
Age group 75-101 : Gender female	0.13	0.00
Age group 35-44 : Education secondary	0.58	0.25
Age group 45-54 : Education secondary	0.22	0.00
Age group 55-64 : Education secondary	0.25	0.00
Age group 65-74 : Education secondary	0.38	0.06
Age group 75-101 : Education secondary	0.44	0.15
Age group 35-44 : Education tertiary	0.46	0.13
Age group 45-54 : Education tertiary	0.26	0.01
Age group 55-64 : Education tertiary	0.25	0.00
Age group 65-74 : Education tertiary	0.22	0.00
Age group 75-101 : Education tertiary	0.54	0.36
Gender female : Education secondary	0.25	0.03
Gender female : Education tertiary	0.28	0.05
Age group 35-44 : Gender female : Education secondary	2.20	0.30
Age group 45-54 : Gender female : Education secondary	6.09	0.01
Age group 55-64 : Gender female : Education secondary	6.22	0.01
Age group 65-74 : Gender female : Education secondary	2.51	0.23
Age group 75-101 : Gender female : Education secondary	1.99	0.38
Age group 35-44 : Gender female : Education tertiary	2.59	0.22
Age group 45-54 : Gender female : Education tertiary	2.56	0.21
Age group 55-64 : Gender female : Education tertiary	3.86	0.07
Age group 65-74 : Gender female : Education tertiary	3.76	0.11
Age group 75-101 : Gender female : Education tertiary	1.43	0.70

Acknowledgements

The Version of Record of the article has been published and is available in Journal of Applied Statistics, February 20 2016
http://www.tandfonline.com/10.1080/02664763.2016.1144725.

References

[1] H. Akaike, *Maximum likelihood identification of gaussian autoregressive moving average models*, Biometrika 60 (1973), pp. 255–265.

[2] H. Bang and J.M. Robins, *Doubly robust estimation in missing data and causal inference models*, Biometrics 61 (2005), pp. pp. 962–972, Available at `http://www.jstor.org/stable/3695907`.

[3] C.M. Cassel, C.E. Sarndal, and J.H. Wretman, *Some uses of statistical models in connection with the nonresponse problem*, in *Incomplete data in sample surveys*, W.G. Madow and I. Olkin, eds., Vol. 3, Academic Press, New York, 1983, pp. 143–160.

[4] A.E. Castaneda, S. Rask, P. Koponen, M. Mölsä, and S. Koskinen, *Maahanmuuttajien terveys ja hyvinvointi -tutkimus venäläis-, somalialais- ja kurditaustaisista Suomessa*, Tech. Rep. 61, National Institute for Health and Welfare (THL), 2012, Available at `http://urn.fi/URN:ISBN:%20978-952-245-739-4`.

[5] F.E. Harrell and C. Dupont, *Hmisc: Harrell Miscellaneous* (2013), Available at `http://CRAN.R-project.org/package=Hmisc`, r package version 3.12-2.

[6] J.D.Y. Kang and J.L. Schafer, *Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data*, Statistical Science 22 (2007), pp. pp. 523–539, Available at `http://www.jstor.org/stable/27645858`.

[7] S. Koskinen, J. Laiho, S. Rinne, N. Kuosmanen, P. Alha, H. Rissanen, and T. Martelin, *Methodology report: Health 2000 Survey*, chap. Participation, complementary data collection and other measures to increase participation rate; 10, National Public Health Institute (2008), pp. 132–149, Available at `http://www.terveys2000.fi/doc/methodologyrep.pdf`.

[8] J. Laiho, K. Djerf, and R. Lehtonen, *Methodology report: Health 2000 Survey*, chap. Sampling design, no. 26 in B, National Public Health Institute (2008), pp. 13–15, Available at `http://www.terveys2000.fi/doc/methodologyrep.pdf`.

[9] R.J. Little, *Survey nonresponse adjustments*, International Statistical Review 54 (1986), pp. 139–157.

[10] R.J. Little and D.B. Rubin, *Statistical analysis with missing data*, Wiley, 2002.

[11] T. Lumley, *Analysis of complex survey samples*, Journal of Statistical Software 9 (2004), pp. 1–19.

[12] T. Lumley, *mitools: Tools for multiple imputation of missing data* (2012), Available at `http://CRAN.R-project.org/package=mitools`, R package version 2.2.

[13] H. Oh and F.J. Scheuren, *Weighting adjustment for unit nonresponse*, in *Incomplete data in sample surveys*, W.G. Madow, I. Olkin, and D.B. Rubin, eds., Vol. 2, New York: Academic Press, 1983, pp. 143–184.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2012), Available at `http://www.R-project.org/`, ISBN 3-900051-07-0.

[15] J.M. Robins, A. Rotnitzky, and L.P. Zhao, *Estimation of regression coefficients when some regressors are not always observed*, Journal of the American Statistical Association 89 (1994), pp. 846–866, Available at `http://amstat.tandfonline.com/doi/full/10.1080/01621459.1994.10476818`.

[16] P.R. Rosenbaum and D.B. Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika 70 (1983), pp. 41–55.

[17] D.B. Rubin, *Multiple imputation for nonresponse in surveys*, Wiley, New York, 1987.

[18] C.E. Särndal and S. Lundström, *Estimation in surveys with nonresponse*, John Wiley & Sons, 2005.

[19] J.L. Schafer, *Multiple imputation: a primer.*, Statistical Methods in Medical Research 8 (1999), pp. 3–15.

[20] T. Therneau, B. Atkinson, and B. Ripley, *rpart: Recursive Partitioning and Regression Trees* (2015), Available at `http://CRAN.R-project.org/package=rpart`, r package version 4.1-10.

[21] S. van Buuren, *Flexible imputation of missing data*, CRC press, 2012.

[22] S. van Buuren and K. Groothuis-Oudshoorn, *mice: Multivariate imputation by chained equations in R*, Journal of Statistical Software 45 (2011), pp. 1–67, Available at `http://www.jstatsoft.org/v45/i03/`.

[23] K.E. Wirth, E.J. Tchetgen Tchetgen, and M. Murray, *Adjustment for missing data in complex surveys using doubly robust estimation: Application to commercial sexual contact among indian men*, Epidemiology 21 (2010), pp. 863–871.