Systematic handling of missing data in complex study designs – Experiences from the Health 2000 and 2011 Surveys

Tommi Härkänen[1] J Karvanen H Tolonen R Lehtonen K Djerf T Juntunen S Koskinen

National Institute for Health and Welfare / Public Health Solutions

Statistical Days, May 19, 2017

[1]E-mail: Tommi.Harkanen@thl.fi

Contents

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Aim 1: Study desing and statistical methods to handle missing data

What kind of design is useful?

We compare

- Cross-sectional design and
- Repeated measures design

Statistical methods

- Weighting
- Multiple imputation
- Doubly robust method

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Aim 2: Challenges in communicating different modeling assumptions

Statistical models Causal assumptions on the variables of interest.

Sampling designs Cluster sampling, nested case-control, varying sampling probabilities, etc.

Missing data Assumptions on missing data mechanism

How to communicate the assumptions to other researchers?

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Causal model with design [2]

A graphical model

Causal node X Variables of scientific interest in the population, possibly unobserved.

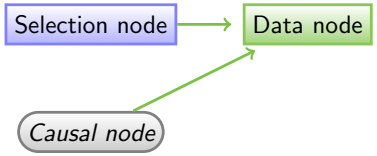
Selection node $\mathfrak{R}$ has the possible values 1 selected and 0 not selected.

Common nodes are

sampling r corresponding to sampling design and

participation R of the sample members.

Data node X^* is defined deterministically $X^* := \begin{cases} X, & \text{if } \mathfrak{R} = 1 \\ \text{NA}, & \text{if } \mathfrak{R} = 0. \end{cases}$



[2](Karvanen, 2015)

Population distribution of outcome Y

Different probabilities:

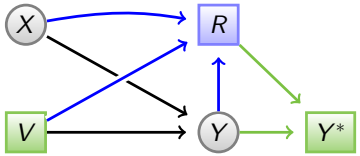
Distribution of outcome Causal model $\mathbb{P}\{ Y \mid V, X \}$.

Selection into sample $\mathbb{P}\{ r = 1 \mid V \}$ where V denotes fully observed (register) causal node.

Participation in survey $\mathbb{P}\{ R = 1 \mid Y, V, X, r = 1 \}$ where X denote partially observed causal node.

Data node of outcome: Y^*. Missing data assumptions:

- Missing compelety at random (MCAR) $\Rightarrow \mathbb{P}\{ R = 1 \mid Y, V, X, r = 1 \} = \mathbb{P}\{ R = 1 \mid r = 1 \}$.
- Missing at random (MAR) $\Rightarrow \mathbb{P}\{ R = 1 \mid Y, V, X, r = 1 \} = \mathbb{P}\{ R = 1 \mid V, r = 1 \}$.
- Missing not at random (MNAR) $\Rightarrow$ $\mathbb{P}\{ R = 1 \mid Y, V, X, r = 1 \} = \mathbb{P}\{ R = 1 \mid Y, V, X, r = 1 \}$.



Sampling design of the Health 2000 and 2011 Surveys in Finland

The Health 2000 Survey in 2000 (aged 18 or older)

- Stratified two-stage sampling design.
- Systematic sampling of individuals with double inclusion probabilities of people aged 80 and older.
- Total sample size was 10,000.

The Health 2011 Survey in 2011 (Lundqvist & Mäki-Opas, 2016; Härkänen et al., 2016)

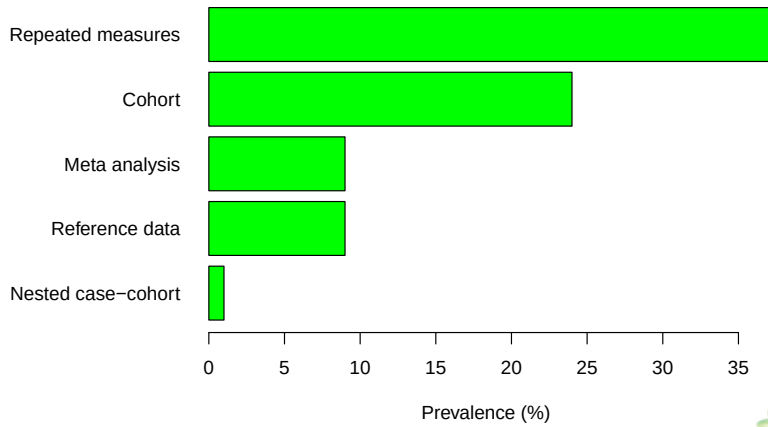
Health 2000 Survey data (aged 29 or older)

- Repeated measurements on the members of the Health 2000 sample
- 7,964 were invited in the age group 30 years or older

New sample of 1,994 young adults (aged 18 to 28)

Study designs in the Health 2000 and 2011 research plans

135 research plans between 2012 and 1/2016



Missing data in the Health 2000 and 2011 Surveys

Participation rates (%) in age group 30 years and above:

Section of the survey	2000	2011	Difference
Health examination	85	5959	-26
Any part of the survey	93	73	-20

Comparison with cross-sectional Finrisk 2012 survey (age 25-74 years):
59 % participated in health examination.

Factors which are often associated with nonresponse

- Low social activity, low education
- Oldest age groups: Illnesses, disabilities, weak functional capacity
- Young age groups: Male

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Administrative register data for all sample members

Linking of the survey sample using the personal ID numbers to several administrative registers with a good coverage contain

Socio-demographics

Age, gender, marital status, education, address, ...

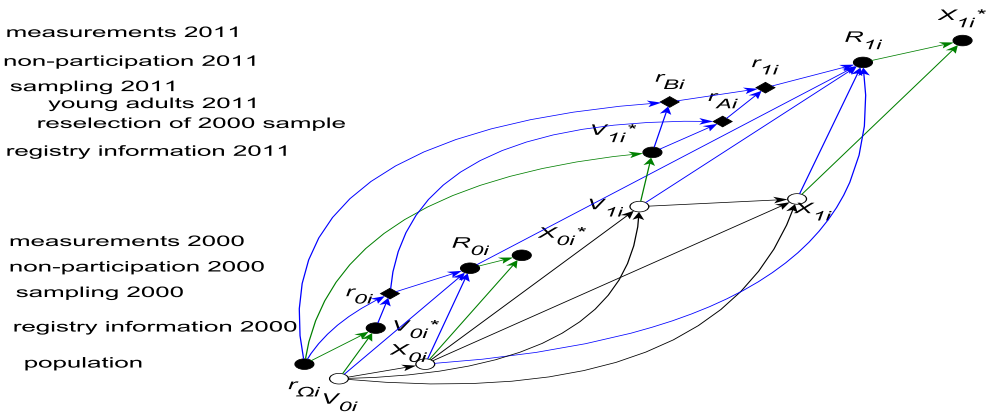
Health-related registers

Care Register from which hospitalization in 2010.

Reimbursement of medical expenses from which medication in 2011.

Disability benefits and services from which disability pension in 2009.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE



	Observed Symbol	2000	2011	Unobserved Symbol	2000	2011
Selection status	♦	r_{0i}	r_{1i}			
Participation status	•	R_{0i}	R_{1i}			
Empirical data	•	X^*_{0i}	X^*_{1i}	○	X_{0i}	X_{1i}
Register data	•	V^*_{0i}	V^*_{1i}	○	V_{0i}	V_{1i}

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Different methods to handle nonparticipation in 2011

(Härkänen et al., 2016)

Inverse probability weights (IPW)

Separate models for participation

Participants of Health 2000 Register data and observed Health 2000 Survey data were used. Weighting model was selected using the Bayesian Information Criterion: self-reported health and work ability, and participation frequency in clubs or associations measured in 2000.

Nonparticipants of Health 2000 Only register data were used.

Multiple imputation

Imputation model 1 (MI1) contained categorical age, gender, language and education

Imputation model 3 (MI3) In addition to variables in MI1 and IPW, also body mass index (BMI), systolic blood pressure and smoking measured in 2000.

Doubly robust

The same weighting model as for the IPW method was used (Wirth et al., 2010).

.FARE

Doubly robust (DR) method

1. Based on two models:
   Outcome regression for outcome Y_i, covariates X_i and regression coefficients β.
   Participation probability for R_i, covariates Z_i and regression coefficients α.
2. Calculate predictive values for outcome $\tilde{Y}_i(X_i, \hat{\beta})$ and participation probability $\pi_i := 1/\left(1+\exp\{-Z_i\hat{\alpha}\}\right)$.
3. Define pseudo outcome

$$\hat{Y}_i^{\mathrm{DR}} := \frac{R_i}{\pi_i} Y_i - \frac{R_i}{\pi_i} \tilde{Y}_i(X_i, \hat{\beta}) + \tilde{Y}_i(X_i, \hat{\beta}).$$

4. DR estimator for the mean outcome is average of the pseudo outcomes $\hat{\mu}_{\mathrm{DR}} := n^{-1} \sum_{i=1}^{n} \hat{Y}_i^{\mathrm{DR}}$.

Large sample bias of $\hat{\mu}_{\mathrm{DR}}$ is zero if
Outcome regression model is correct $\tilde{Y}_i(X_i, \hat{\beta}) \to \mathbb{E}[\, Y_i \,|\, X_i \,]$ **or**
Participation probability model is correct $\pi_i \to \mathbb{P}\{\, R_i \,|\, Z_i \,\}$.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Conclusion

Sampling design

Repeated measures design can

- improve performance of multiple imputation and other methods to correct for effects of missing data (more individuals participate in at least one measurement point) and
- provide more reliable results for assessing causal hypotheses

when compared to cross-sectional design.

Statistical methods to handle missing data

Our empirical analyses suggest that the multiple imputation methods managed to remove most bias caused by the non-response.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

References

T. Härkänen, et al. (2016). 'Systematic handling of missing data in complex study designs–experiences from the Health 2000 and 2011 Surveys'. *Journal of Applied Statistics* **43**(15):2772–2790.

J. Karvanen (2015). 'Study design in causal models'. *Scandinavian Journal of Statistics* **42**(2):361–377.

A. Lundqvist & T. Mäki-Opas (eds.) (2016). *Health 2011 Survey - Methods*. No. 2016_008 in Raportti. THL.

K. E. Wirth, et al. (2010). 'Adjustment for Missing Data in Complex Surveys Using Doubly Robust Estimation: Application to Commercial Sexual Contact Among Indian Men'. *Epidemiology* **21**(6):863–871.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Afternoon seminar

Our NoPaHes project will present more results in the Afternoon Seminar of the Finnish Statistical Society.

Place National Institute for Health and Welfare (THL), Mannerheimintie 166 A, Helsinki

Time August 30, 2017 at 12:30

More details will be available at the web site of the Society:

`http://tilastoseura.fi/`

Welcome!

NATIONAL INSTITUTE FOR HEALTH AND WELFARE