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

Institute for Health and Welfare / Department of Health

BaNoCoSS, August 25, 2015

NATIONAL INSTITUTE FOR HEALTH AND WELFARE FINLAND

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

Contents

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Challenges in communicating assumptions of statistical models

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

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Challenges in communicating assumptions of statistical models

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

Missing data Assumptions on missing data mechanism:

- **MCAR** Missing completely at random,
- **MAR** missing at random, or
- **NMAR** not missing at random.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Challenges in communicating assumptions of statistical models

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

Missing data Assumptions on missing data mechanism:

- **MCAR** Missing completely at random,
- **MAR** missing at random, or
- **NMAR** not missing at random.

Statistical models Causal assumptions on the variables of interest.

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Challenges in communicating assumptions of statistical models

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

Missing data Assumptions on missing data mechanism:

- **MCAR** Missing completely at random,
- **MAR** missing at random, or
- **NMAR** not missing at random.

Statistical models Causal assumptions on the variables of interest.

How to communicate assumptions of steps above to other researchers?

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Causal model with design

A graphical model

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

Causal node

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Causal model with design

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.

Selection node

Causal node

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Causal model with design

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

Selection node

Data node

Causal node

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Causal model with design

A graphical model

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

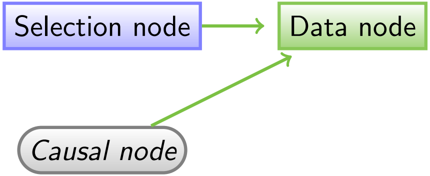
Selection node $\mathfrak{R}$ has the possible values 1 selected and 0 not selected.

Common nodes are

sampling r corresponding to sampling design and

participation R of the sample members.

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



NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Causal model with design

A graphical model

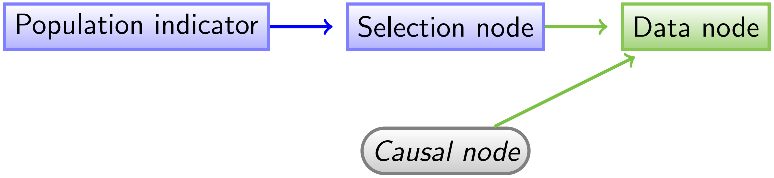
Causal node X Variables of scientific interest in the population, possibly unobserved.

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

Population indicator $r_\Omega \equiv 1$.

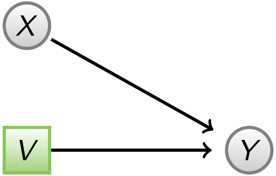


NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Population distribution of outcome Y

Different probabilities:

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

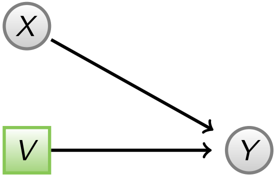


Population distribution of outcome Y

Different probabilities:

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

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



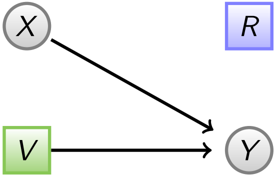
Population distribution of outcome Y

Different probabilities:

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

Selection for 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.



NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Population distribution of outcome Y

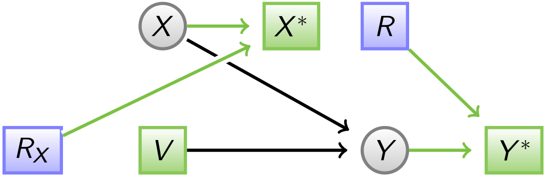
Different probabilities:

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

Selection for 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 nodes of outcome and covariate: Y^* and X^*.



NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Population distribution of outcome Y

Different probabilities:

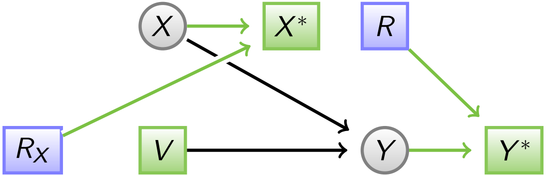
Distribution of outcome Causal model $\mathbb{P}\{Y \mid V, X\}$.

Selection for 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 nodes of outcome and covariate: Y^* and X^*. Missing data assumptions:

- Missing compelety at random (MCAR) $\Rightarrow \mathbb{P}\{R=1 \mid Y, V, X, r=1\} = \mathbb{P}\{R=1 \mid r=1\}$.



NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Population distribution of outcome Y

Different probabilities:

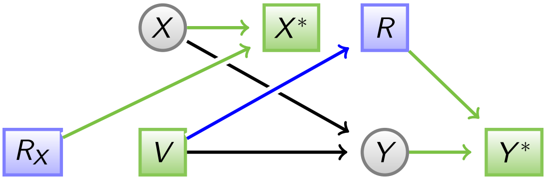
Distribution of outcome Causal model $\mathbb{P}\{Y \mid V, X\}$.

Selection for 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 nodes of outcome and covariate: Y^* and X^*. 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) $\Rightarrow \mathbb{P}\{R = 1 \mid Y, V, X, r = 1\} = \mathbb{P}\{R = 1 \mid V, r = 1\}$.



NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Population distribution of outcome Y

Different probabilities:

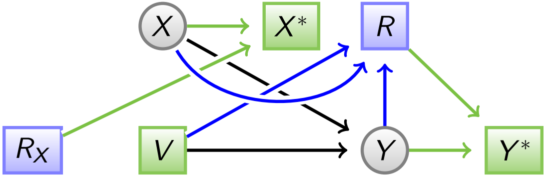
Distribution of outcome Causal model $\mathbb{P}\{Y \mid V, X\}$.

Selection for 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 nodes of outcome and covariate: Y^* and X^*. 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) $\Rightarrow \mathbb{P}\{R = 1 \mid Y, V, X, r = 1\} = \mathbb{P}\{R = 1 \mid V, r = 1\}$.
- Not missing at random (NMAR) $\Rightarrow$
  $\mathbb{P}\{R = 1 \mid Y, V, X, r = 1\} = \mathbb{P}\{R = 1 \mid Y, V, X, r = 1\}$.



NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Sampling design of the Health 2000 and 2011 Surveys in Finland

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

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.

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

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.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

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.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

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

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

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

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

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)

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

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	59	**-26**
Any part of the survey	93	73	**-20**

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

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	59	**-26**
Any part of the survey	93	73	**-20**

Factors which are often associated with nonresponse

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	59	**-26**
Any part of the survey	93	73	**-20**

Factors which are often associated with nonresponse

- Low social activity, low education

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

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	59	**-26**
Any part of the survey	93	73	**-20**

Factors which are often associated with nonresponse

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

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

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	59	**-26**
Any part of the survey	93	73	**-20**

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

Available administrative register data

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

Available administrative register data

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, . . .

Available administrative register data

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 (former Hospital Discharge Register) from which hospitalization in 2010.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Available administrative register data

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 (former Hospital Discharge Register) from which hospitalization in 2010.

Reimbursement of medical expenses from which medication in 2011.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Available administrative register data

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, . . .

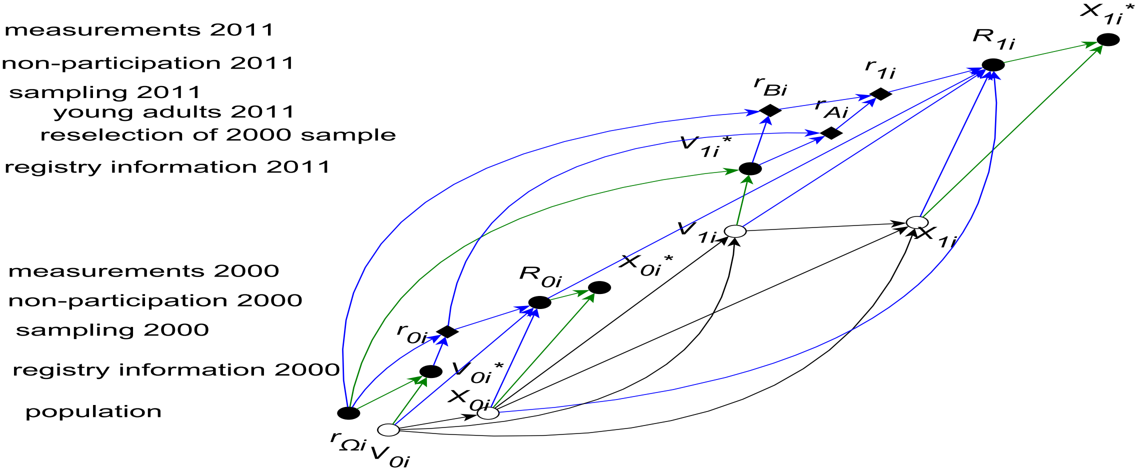
Health-related registers

Care Register (former Hospital Discharge 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......			Unobserved		
	Symbol	2000	2011	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

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.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Different methods to handle nonparticipation in 2011

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.

Nonparticipants of Health 2000 Only register data were used.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Different methods to handle nonparticipation in 2011

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.

Nonparticipants of Health 2000 Only register data were used.

Doubly robust

The same weighting model as for the IPW method was used.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Different methods to handle nonparticipation in 2011

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.

Nonparticipants of Health 2000 Only register data were used.

Doubly robust

The same weighting model as for the IPW method was used.

Multiple imputation

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

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

Results in the Health 2011 Survey

Variable	Clustering	Missing data method	Prev. (%)	SE
Disability pension	SRS	Complete case analysis	8.8	0.4
	Complex	IPW weights	9.3	0.4
	Complex	Doubly Robust	9.3	0.5
	Complex	MI1	9.4	0.5
	Complex	MI3	**9.5**	0.4
	Complex	Full sample prevalence	**9.5**	0.4
Hospitalization	SRS	Complete case analysis	16.6	0.5
	Complex	IPW weights	16.9	0.5
	Complex	Doubly Robust	17.2	0.6
	Complex	MI1	**17.4**	0.9
	Complex	MI3	17.2	0.5
	Complex	Full sample prevalence	**17.6**	0.5
Medication	SRS	Complete case analysis	40.2	0.7
	Complex	IPW weights	40.8	0.8
	Complex	Doubly Robust	41.4	0.7
	Complex	MI1	41.0	0.9
	Complex	MI3	**41.8**	0.7
	Complex	Full sample prevalence	**41.9**	0.6

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE

Conclusion

Graphical models

provide a useful tool to describe complex statistical (causal) models involving a sampling design and non-response.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Conclusion

Graphical models

provide a useful tool to describe complex statistical (causal) models involving a sampling design and non-response.

Assumptions on missing data mechanism

MCAR and MAR can be handled easily, but researcher should consider also possible NMAR mechanisms.

NATIONAL INSTITUTE FOR HEALTH AND WELFARE

Conclusion

Graphical models

provide a useful tool to describe complex statistical (causal) models involving a sampling design and non-response.

Assumptions on missing data mechanism

MCAR and MAR can be handled easily, but researcher should consider also possible NMAR mechanisms.

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