NTNU
Regional Centre for Child and Youth
Mental Health and Child Welfare

Missing data:
The problem and possible solutions.

by
Stian Lydersen

Presentation at NTNU, 3 October 2014, 0900 to 1100
Updated 9 Oct 2014

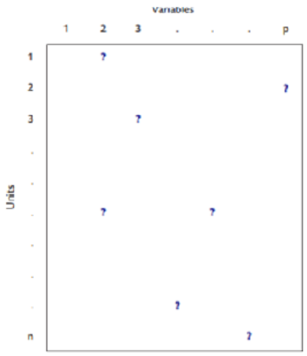
http://folk.ntnu.no/slyderse/medstat/Missing%20data%203%20Oct%202014.pdf

Contents

- Definitions
- Methods overview
- Examples
- Methods more closely
- Concluding remarks
- Litterature

Missing data:

- ”Holes” in the data matrix which ideally should be complete
- Usually, these are data we intended to collect, but for some reason did not.
- There exists a meaningful value which was not recorded.



Missing data mechanism

Let R denote what is missing, for example 0 (1) if the corresponding value is observed (missing).

The probability distribution of R has been called

- Missing data mechanism
- Probability of nonresponse
- response mechanism
- missingness mechanism
- probability of missingness
- distribution of missingness

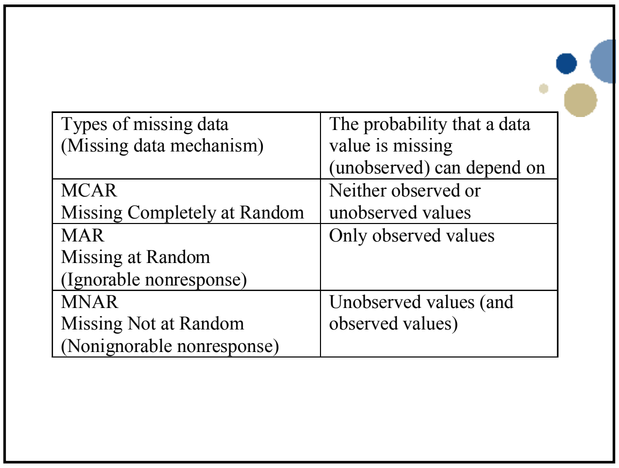
Valid analyses should provide unbiased estimates of:

1. The quantities of interest, such as population means or regression coefficients.
2. The variance (or SE) of our estimates.

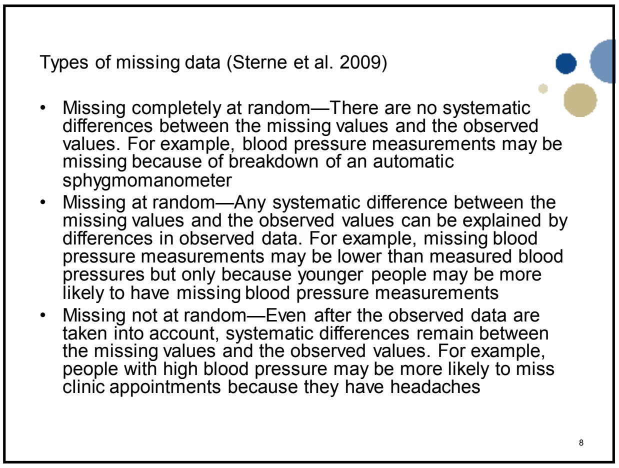
The second criterion is needed to obtain:

a. Confidence intervals with an actual coverage close to the nominal coverage (usually 95%).
b. Hypotheses tests with an actual significance level close to the nominal significance level.

(Bjørnstad & Lydersen 2012)



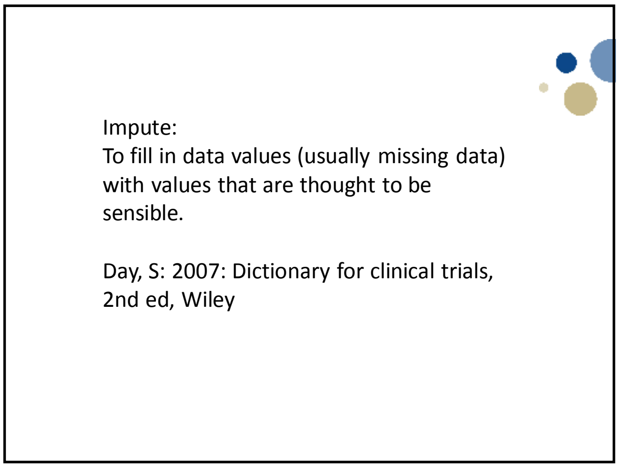
Types of missing data (Missing data mechanism)	The probability that a data value is missing (unobserved) can depend on
MCAR Missing Completely at Random	Neither observed or unobserved values
MAR Missing at Random (Ignorable nonresponse)	Only observed values
MNAR Missing Not at Random (Nonignorable nonresponse)	Unobserved values (and observed values)



Types of missing data (Sterne et al. 2009)

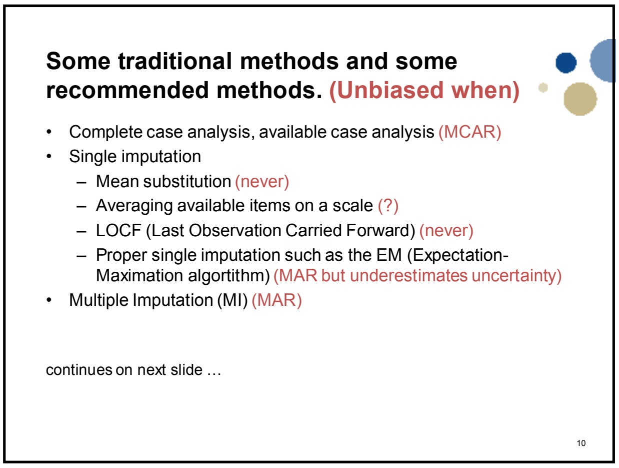
- Missing completely at random—There are no systematic differences between the missing values and the observed values. For example, blood pressure measurements may be missing because of breakdown of an automatic sphygmomanometer
- Missing at random—Any systematic difference between the missing values and the observed values can be explained by differences in observed data. For example, missing blood pressure measurements may be lower than measured blood pressures but only because younger people may be more likely to have missing blood pressure measurements
- Missing not at random—Even after the observed data are taken into account, systematic differences remain between the missing values and the observed values. For example, people with high blood pressure may be more likely to miss clinic appointments because they have headaches

8



Impute:
To fill in data values (usually missing data) with values that are thought to be sensible.

Day, S: 2007: Dictionary for clinical trials, 2nd ed, Wiley

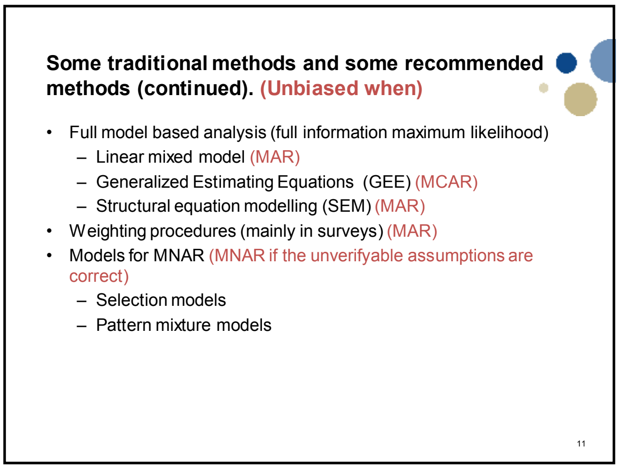


Some traditional methods and some recommended methods. (Unbiased when)

- Complete case analysis, available case analysis (MCAR)
- Single imputation
  - Mean substitution (never)
  - Averaging available items on a scale (?)
  - LOCF (Last Observation Carried Forward) (never)
  - Proper single imputation such as the EM (Expectation-Maximation algortithm) (MAR but underestimates uncertainty)
- Multiple Imputation (MI) (MAR)

continues on next slide …

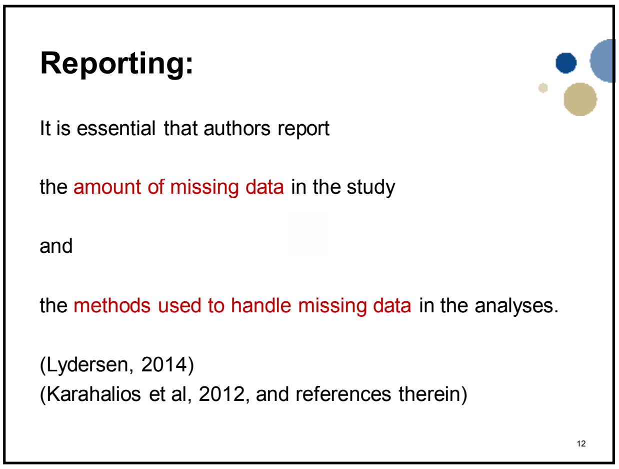
10



Some traditional methods and some recommended methods (continued). (Unbiased when)

- Full model based analysis (full information maximum likelihood)
  - Linear mixed model (MAR)
  - Generalized Estimating Equations (GEE) (MCAR)
  - Structural equation modelling (SEM) (MAR)
- Weighting procedures (mainly in surveys) (MAR)
- Models for MNAR (MNAR if the unverifyable assumptions are correct)
  - Selection models
  - Pattern mixture models

11



Reporting:

It is essential that authors report

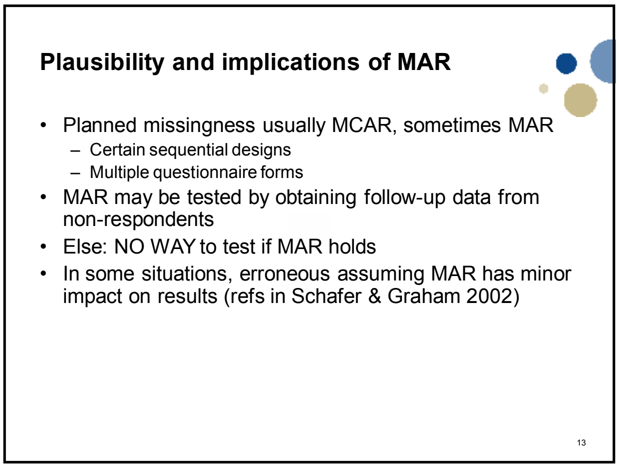
the amount of missing data in the study

and

the methods used to handle missing data in the analyses.

(Lydersen, 2014)
(Karahalios et al, 2012, and references therein)

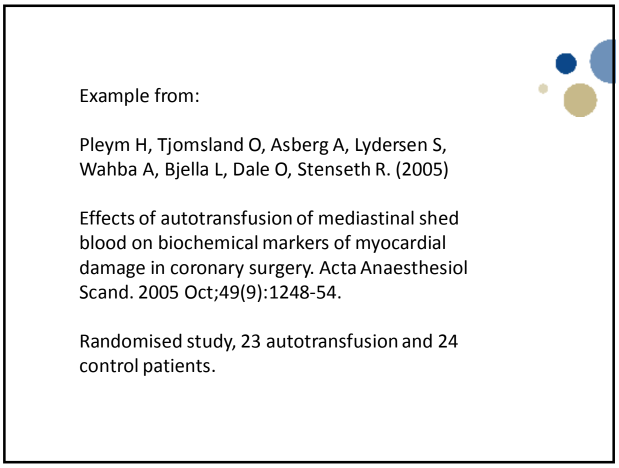
12



Plausibility and implications of MAR

- Planned missingness usually MCAR, sometimes MAR
  - Certain sequential designs
  - Multiple questionnaire forms
- MAR may be tested by obtaining follow-up data from non-respondents
- Else: NO WAY to test if MAR holds
- In some situations, erroneous assuming MAR has minor impact on results (refs in Schafer & Graham 2002)

13

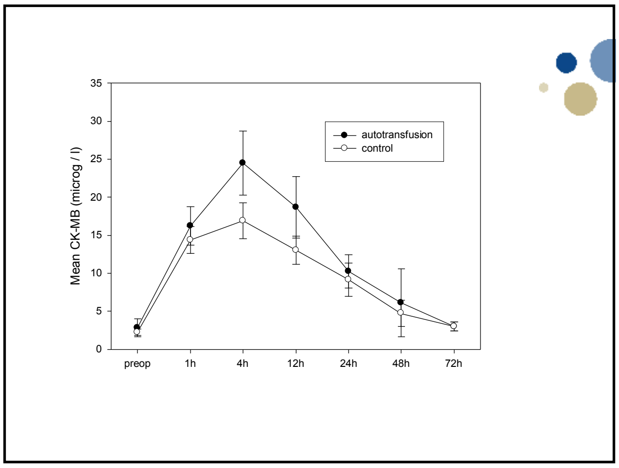


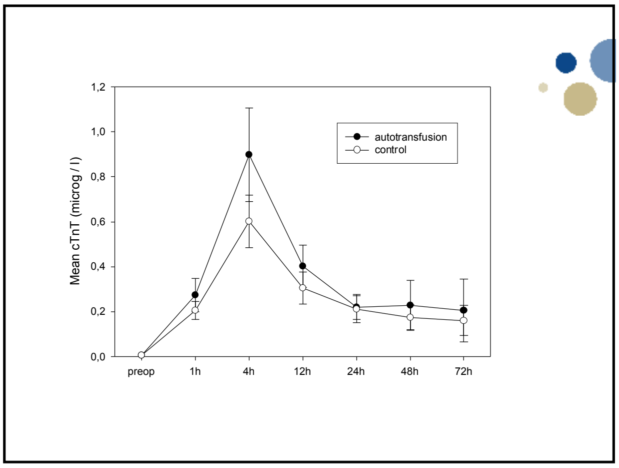
Example from:

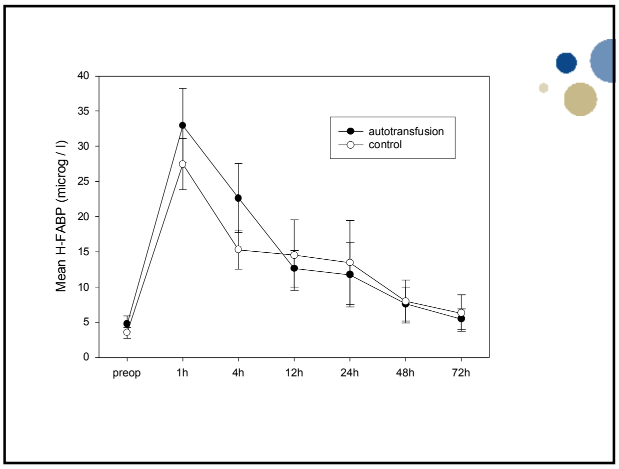
Pleym H, Tjomsland O, Asberg A, Lydersen S, Wahba A, Bjella L, Dale O, Stenseth R. (2005)

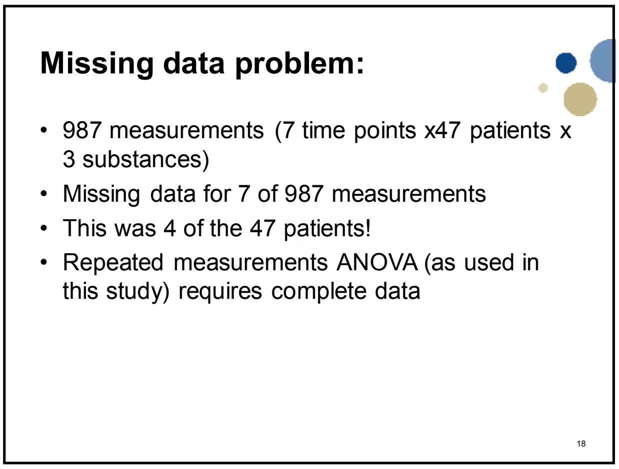
Effects of autotransfusion of mediastinal shed blood on biochemical markers of myocardial damage in coronary surgery. Acta Anaesthesiol Scand. 2005 Oct;49(9):1248-54.

Randomised study, 23 autotransfusion and 24 control patients.









Missing data problem:

- 987 measurements (7 time points x47 patients x 3 substances)
- Missing data for 7 of 987 measurements
- This was 4 of the 47 patients!
- Repeated measurements ANOVA (as used in this study) requires complete data

18

"A total of 7 out of 987 serum values were missing. Missing values were imputed using the EM algorithm with multivariate normal distribution on ln-transformed data. According to inspection of Q-Q plots, the ln-transformed data showed acceptable fit to the normal distribution, while the original data tended to be skewed. Repeated measurements ANOVA was used for joint analysis of the serum values of CK-MB, cTnT, and H-FABP, respectively, using the EM imputed ln-transformed values."

Example from:

Hallan, S. I., Ritz, E., Lydersen, S., Romundstad, S., Kvenild, K., & Orth, S. R. Combination of estimated glomerular filtration rate and albuminuria provides best prediction of kidney failure: Results of the HUNT II study, Norway. In press, Journal of the American Society of Nephrology, 2009.

Cox proportional hazards regression with time to kidney failure (CKD stage 5) as dependent variable.

HUNT II (Helseundersøkelsen i Nord-Trøndelag), 1995-1997. Follow-up until 2007.

92939 persons, 20 years and older, were invited. 65589 (70.6%) responded.
124 kidney failures.
8360 were hypertensive or had diabetes mellitus. These were asked to deliver urine samples, and 88.6% did so. In addition, a random 5% sample of non-diabetic non-hypertensive subjects (n=2,861) was also asked to deliver urine samples; 75.6% did so.

Hence: For 95% of the non-diabetic non-hypertensive subjects, urine samples were
Missing at random (MAR) by design.

Variable	n	% missing
Follow-up time	65589	0,0
Age	65589	0,0
Male sex	65589	0,0
Low education	61369	6,4
Depression	58423	10,9
Smoking	64395	1,8
Low physical activity	57881	11,8
Diabetes mellitus	64693	1,4
CVD	64624	1,5
BMI	64306	2,0
Waist circumference	64022	2,4

Variable	n	% missing
Systolic BP	64708	1,3
Diastolic BP	64708	1,3
Cholesterol	65158	0,7
HDL-Cholesterol	65155	0,7
GLUCOSE	65158	0,7
Triglycerides	65158	0,7
Creatinine	65158	0,7
eGFR [1]	65158	0,7
ACR [2]	9703	85,2

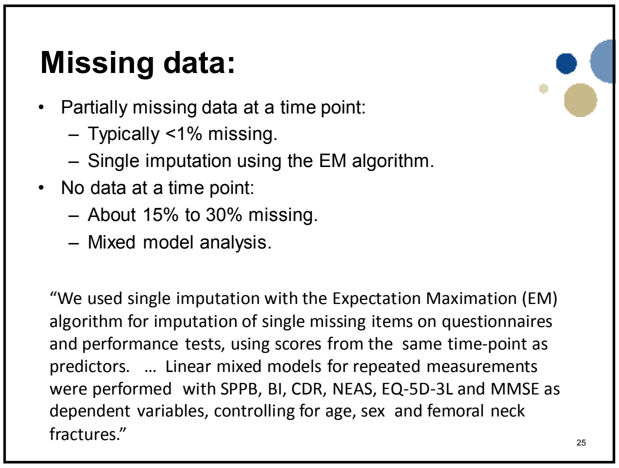
[1] estimated glomerular filtration rate
[2] Albumin creatinin ratio (from urine sample)
Not requested (Missing by design): 82,8 %
Requested, but not deliverd: 2,5%

Example from:

Prestmo, A., Hagen, G., Sletvold, O., Helbostad, J.L., Thingstad, P., Taraldsen, K., Lydersen, S., Halsteinli, V., Saltnes, T., Lamb, S.E., Johnsen, L.G., & Saltvedt, I. "A randomised trial of comprehensive geriatric care in hip-fracture patients." *The Lancet*, In press, 2014.

Hip fracture patients > 70 years. RCT of Comprehensive Geriatric care (CGC) versus usual ortopaedic care (OC)

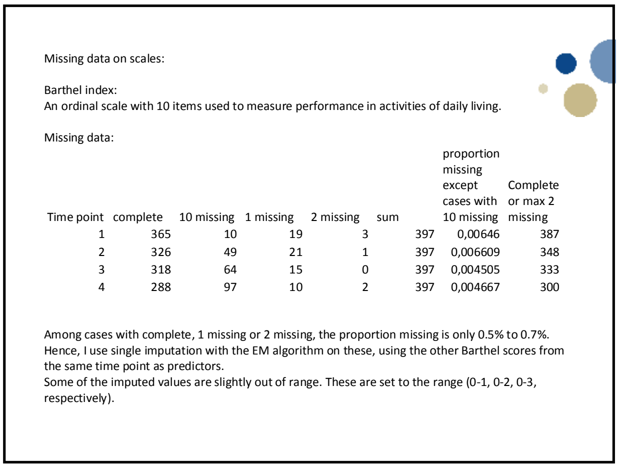
397 patients assessed at baseline, 1 month, 4 months and 1 year.



Missing data:

- Partially missing data at a time point:
  - Typically <1% missing.
  - Single imputation using the EM algorithm.
- No data at a time point:
  - About 15% to 30% missing.
  - Mixed model analysis.

"We used single imputation with the Expectation Maximation (EM) algorithm for imputation of single missing items on questionnaires and performance tests, using scores from the same time-point as predictors. … Linear mixed models for repeated measurements were performed with SPPB, BI, CDR, NEAS, EQ-5D-3L and MMSE as dependent variables, controlling for age, sex and femoral neck fractures."



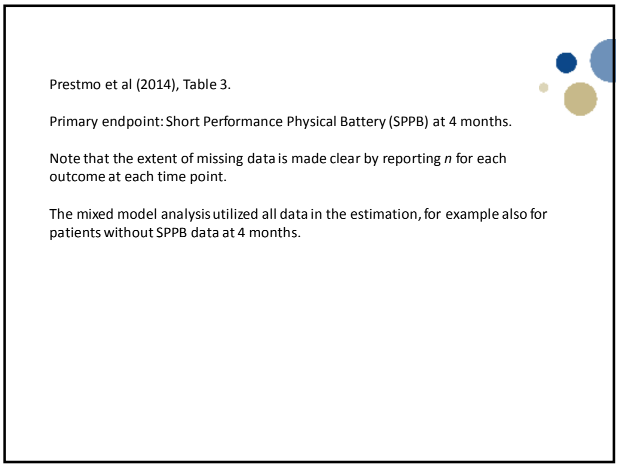
Missing data on scales:

Barthel index:
An ordinal scale with 10 items used to measure performance in activities of daily living.

Missing data:

Time point	complete	10 missing	1 missing	2 missing	sum	proportion missing except cases with 10 missing	Complete or max 2 missing
1	365	10	19	3	397	0,00646	387
2	326	49	21	1	397	0,006609	348
3	318	64	15	0	397	0,004505	333
4	288	97	10	2	397	0,004667	300

Among cases with complete, 1 missing or 2 missing, the proportion missing is only 0.5% to 0.7%. Hence, I use single imputation with the EM algorithm on these, using the other Barthel scores from the same time point as predictors.
Some of the imputed values are slightly out of range. These are set to the range (0-1, 0-2, 0-3, respectively).

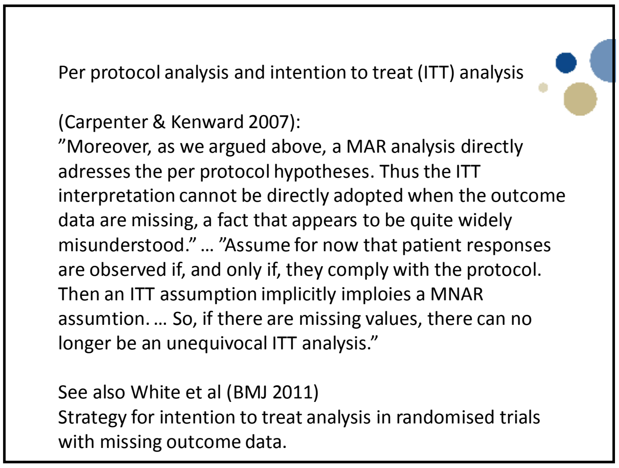


Prestmo et al (2014), Table 3.

Primary endpoint: Short Performance Physical Battery (SPPB) at 4 months.

Note that the extent of missing data is made clear by reporting *n* for each outcome at each time point.

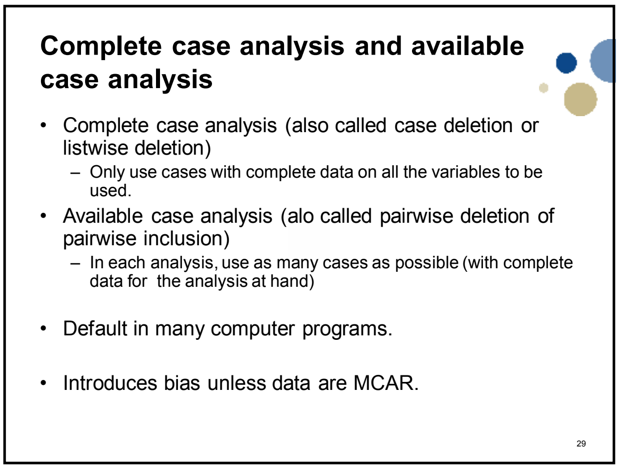
The mixed model analysis utilized all data in the estimation, for example also for patients without SPPB data at 4 months.



Per protocol analysis and intention to treat (ITT) analysis

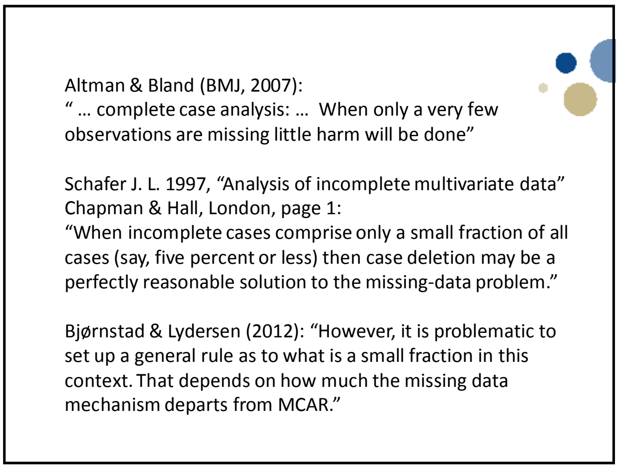
(Carpenter & Kenward 2007):
"Moreover, as we argued above, a MAR analysis directly adresses the per protocol hypotheses. Thus the ITT interpretation cannot be directly adopted when the outcome data are missing, a fact that appears to be quite widely misunderstood." … "Assume for now that patient responses are observed if, and only if, they comply with the protocol. Then an ITT assumption implicitly imploies a MNAR assumtion. … So, if there are missing values, there can no longer be an unequivocal ITT analysis."

See also White et al (BMJ 2011)
Strategy for intention to treat analysis in randomised trials with missing outcome data.



Complete case analysis and available case analysis

- Complete case analysis (also called case deletion or listwise deletion)
  - Only use cases with complete data on all the variables to be used.
- Available case analysis (alo called pairwise deletion of pairwise inclusion)
  - In each analysis, use as many cases as possible (with complete data for the analysis at hand)
- Default in many computer programs.
- Introduces bias unless data are MCAR.



Altman & Bland (BMJ, 2007):
" … complete case analysis: … When only a very few observations are missing little harm will be done"

Schafer J. L. 1997, "Analysis of incomplete multivariate data" Chapman & Hall, London, page 1:
"When incomplete cases comprise only a small fraction of all cases (say, five percent or less) then case deletion may be a perfectly reasonable solution to the missing-data problem."

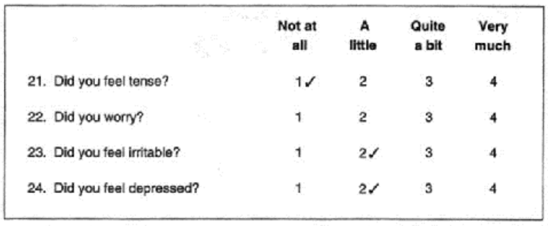
Bjørnstad & Lydersen (2012): "However, it is problematic to set up a general rule as to what is a small fraction in this context. That depends on how much the missing data mechanism departs from MCAR."

Mean substitution:

- For subject missing data on a variable, fill in the mean for the subjects with data on the variable.
- NEVER OK to do this
- Note that this means averaging across subjects. Averaging *within* subjects (items on a scale) can be OK)

Averaging available items on a scale

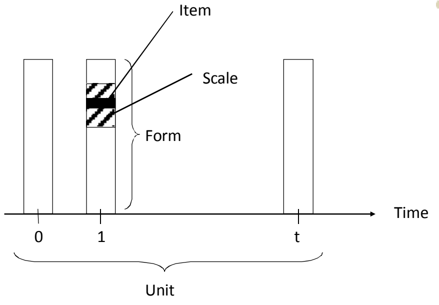
European organization for research and treatment of cancer QLQ-C30 (EORTC QLQ-C30)



	Not at all	A little	Quite a bit	Very much
21. Did you feel tense?	1✓	2	3	4
22. Did you worry?	1	2	3	4
23. Did you feel irritable?	1	2✓	3	4
24. Did you feel depressed?	1	2✓	3	4

Figure 15.3 The emotional functioning scale of the EORTC QLQ-C30

Example: Quality of Life questionnaires



Beregning av gjennomsnitt i SPSS

Mean(q21, q22, q23, q24).
beregner hvis minst en av variablene er gitt

(q21+ q22 + q23 + q24)/4
beregner bare hvis alle variablene er gitt

Mean.2(q21, q22, q23, q24).
beregner hvis minst 2 av verdiene er gitt

Last observation carried forward (LOCF, LVCF)

Suppose a trial has longitudinal follow up, and a patient withdraws from the study or temporarily fails to attend one of the visits. If the missing value is set equal to the last observed value, this is called last observation carried forward (LOCF). This is a popular method for handling missing data, due to its simplicity. But it is not valid under any sensible assumptions, and it should not be used, see (Carpenter & Kenward 2007) and references therein.

Defining «missing» as a data value

For example, if smoking has the categories 0 (no) and 1 (yes), one could add an additional category 2 (missing), and regard this as three nominal categories with no missing answers. Such approaches have the potential to introduce bias and are not recommended, see Horton and Kleinman (2007) and references therein.

Using logical structures in the questionnaire

Example: The HUNT 2 questionnaire includes several questions about smoking habits

"Do you smoke daily at present?"

"If you smoked earlier, how long ago did you quit smoking?"

If the first question is unanswered, and the second question is answered, one can deduce that the person does not smoke daily at present. Originally, 15% of the subjects did not answer the question about daily smoking. Assuming that the answers were internally consistent, it was possible to fill in most of the missing values, resulting in only 2% missing in daily smoking (Hallan et al 2009).

Single imputation:
The EM (Expectation – Maximation) Algorithm for ML estimation

- Assume a multivariate distribution (usually normal)
- Fill in missing data with a best guess
- Estimate the parameters for the complete data set
- Re-guess missing data with the estimated parameters
- Repeat until convergence

- May need many iterations

- Available in many statistical software packages

- Unbiased if MAR but underestimates uncertainty

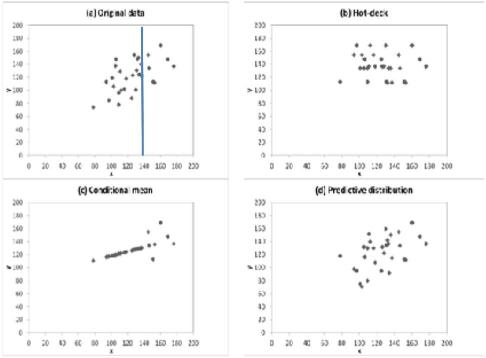
Multiple imputation (MI)

Example
(Bjørnstad & Lydersen 2012, Example 13.4 page 445 – 447, Schafer 2002)

Record systolic blood pressure (x) in January. Only those with x > 140 measure blood pressure in February.

30 fictious observations with mean=125, SD=26, correlation=0.6.

x	*y*	
	Complete	**MAR**
169	148	148
126	123	--
132	149	--
160	169	169
105	138	--
116	102	--
125	88	--
112	100	--
133	150	--
94	113	--
109	96	--
109	78	--
106	148	--
176	137	137
128	155	--
131	131	--
130	101	--
145	155	155
136	140	--
146	134	134
111	129	--
97	85	--
134	124	--
153	112	112
118	118	--
137	122	--
101	119	--
103	106	--
78	74	--
151	113	113
Summary data: mean (standard deviation)		
125.7 (23.0)	121.9 (24.7)	138.3 (21.1)



MI (Multiple Imputation), Rubin (1987)

- Create m > 1 (for example m=20) data sets by single imputation from the conditional distribution (Imputation model)
- Analyse each data set by a complete data method (Analysis model)
- Combine the results using simple artihmetric to obtain overall estimates reflecting missing data uncertainty and finite-sample variations.

MI - advantages

- Retains the attractive of single imputation from conditional distribution
- A single imputed set may be randomly atypical
- Does not underestimate uncertainty
- Unlike other Monte Carlo methods, few repetitions are needed.

Rubin's (1987) rules for combining estimates and variances

Q = the population quantity of interest, $U = Var(\hat{Q})$

m estimates $\hat{Q}^{(j)}$, $U^{(j)}$, for j = 1, …, m

Estimate for Q:

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}^{(j)}$$

Average within-imputation variance

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U^{(j)}$$

Between-imputation variance

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left[\hat{Q}^{(j)} - \bar{Q}\right]^2$$

Total variance:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

Student's t approximation for confidence intervals and tests for Q

$$\frac{\bar{Q} - Q}{\sqrt{T}} \sim t_{\upsilon}$$

where

$$\upsilon = (m-1)\left[1 + \frac{\bar{U}}{(1+m^{-1})B}\right]^2$$

Proper MI reflecting the uncertainty in the model parameters

A single imputation is drawn from $P(Y_{mis} \mid Y_{obs}; \hat{\theta})$

MI:

simulate m plausible values $\theta^{(1)}, \ldots, \theta^{(m)}$
draw $Y_{mis}^{(t)}$ from $P[Y_{mis} \mid Y_{obs}; \theta^{(t)}]$ for t=1,…,m

Bayesian approach with a prior distribution for θ
is natural but not essential

How many imputations *m*?

- The classic advice was *m* = 3 to 5.
- Bjørnstad & Lydersen (2012) generally recommend *m* = 20. But a higher number may be required to report p-values with, say, 2 digits accuracy.
- Van Buuren (2012) reviews relevant work. «It could be beneficial to set *m* higher, in the range 20 to 100.»
- If you use *m*=100, you are on the safe side.

Multiple imputation using chained equations (“MICE”)

Analysis model:

$y \mid x_1\, x_2\, x_3 \cdots x_p$

Imputation model:
A set of regression equations
(usually linear, binary logistic regression, or ordinal logistic regression)

$$x_1 \mid y \quad x_2\, x_3 \cdots x_p$$
$$x_2 \mid y\, x_1 \quad x_3 \cdots x_p$$
$$\vdots$$
$$x_p \mid y\, x_1\, x_2 \cdots x_{p-1}$$

MI using chained equations

- Idea: Mimic the conditional distribution of the missing values given the joint distribution
- Automatic procedure:
  - Insert initial guesses for the missing values.
  - Use the equations to improve the predicted missing values
  - Repeat until convergence (not always achieved)
- Even when convergence is achieved, it can happen that it converges at some other distribution
- Simulation studies confirm that the procedure still works (surprisingly) well

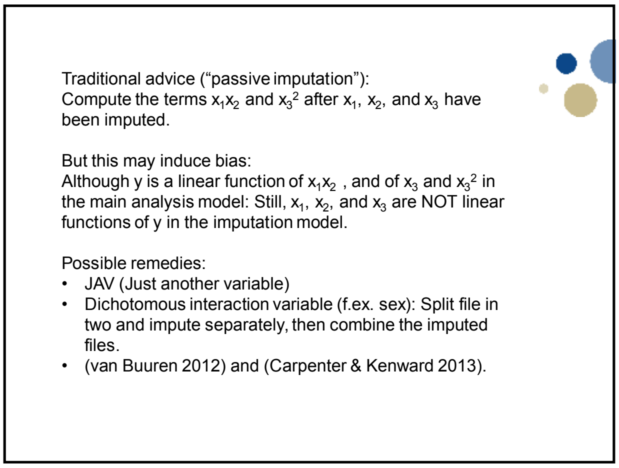
Predictors in the imputation model

- Include all variables to be used in the main analysis model(s). Failure to do so may bias the analysis.
- Possible interactions and nonlinear effects must be handled appropriately.
- Include predictors of missingness.
- Include variables explaining much of the variance in the target variable (to be imputed)
- Some algorithms require categorical variables to be coded 0, 1, …, k-1
- In MI algorithms, continuous variables are typically assumed normally distributed. This may be handled in alternative ways. (See later slide).
- The outcome variable in the analysis model must be included as a predictor in the imputation model
- When the main analysis is lifetime analysis such as Cox regression, include both the time t and censoring indicator as predictors.

Interactions and nonlinear effects in the analysis model:

Interaction:
Includes the term x_1x_2 in addition to the main effect x_1 and x_2.

Nonlinear effect:
For example, x_3 and x_3^2.

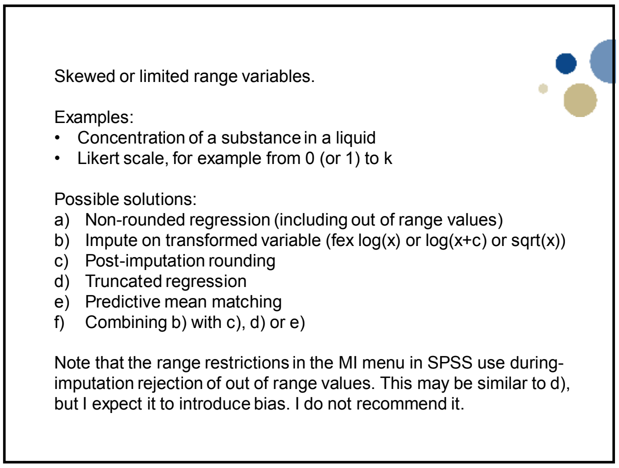


Traditional advice (“passive imputation”):
Compute the terms x_1x_2 and x_3^2 after x_1, x_2, and x_3 have been imputed.

But this may induce bias:
Although y is a linear function of x_1x_2 , and of x_3 and x_3^2 in the main analysis model: Still, x_1, x_2, and x_3 are NOT linear functions of y in the imputation model.

Possible remedies:

- JAV (Just another variable)
- Dichotomous interaction variable (f.ex. sex): Split file in two and impute separately, then combine the imputed files.
- (van Buuren 2012) and (Carpenter & Kenward 2013).



Skewed or limited range variables.

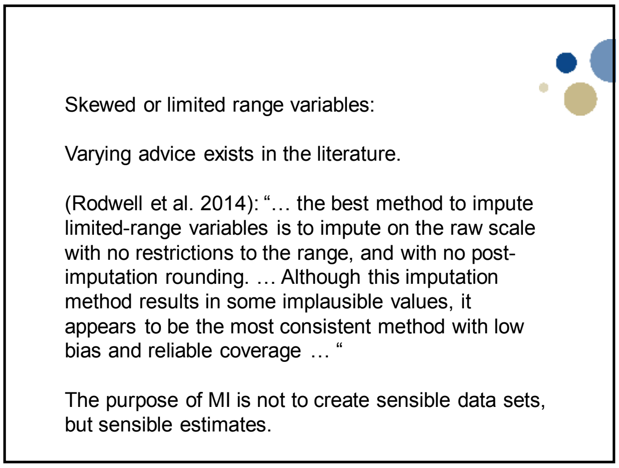
Examples:

- Concentration of a substance in a liquid
- Likert scale, for example from 0 (or 1) to k

Possible solutions:

a) Non-rounded regression (including out of range values)
b) Impute on transformed variable (fex log(x) or log(x+c) or sqrt(x))
c) Post-imputation rounding
d) Truncated regression
e) Predictive mean matching
f) Combining b) with c), d) or e)

Note that the range restrictions in the MI menu in SPSS use during-imputation rejection of out of range values. This may be similar to d), but I expect it to introduce bias. I do not recommend it.

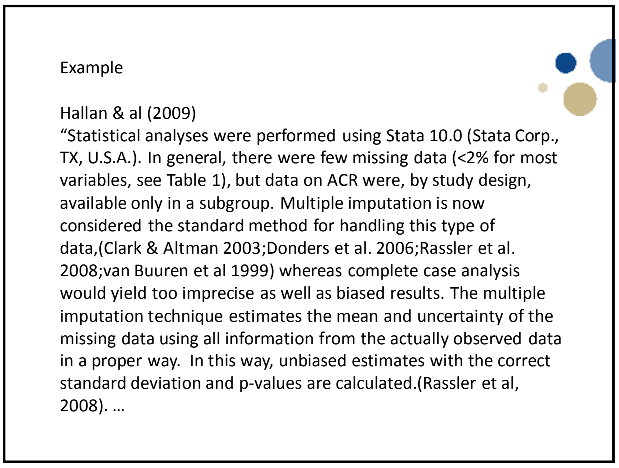


Skewed or limited range variables:

Varying advice exists in the literature.

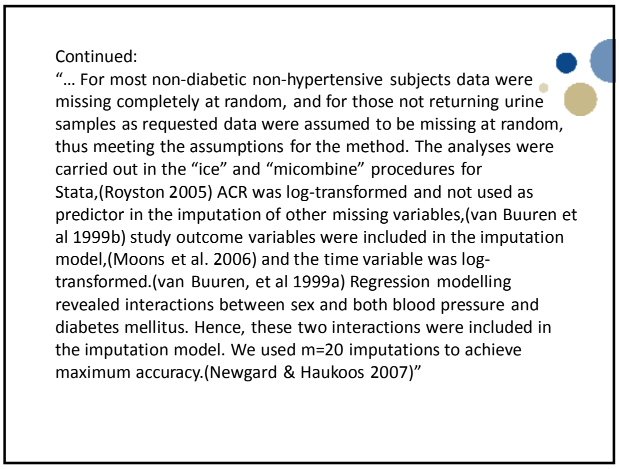
(Rodwell et al. 2014): “… the best method to impute limited-range variables is to impute on the raw scale with no restrictions to the range, and with no post-imputation rounding. … Although this imputation method results in some implausible values, it appears to be the most consistent method with low bias and reliable coverage … “

The purpose of MI is not to create sensible data sets, but sensible estimates.

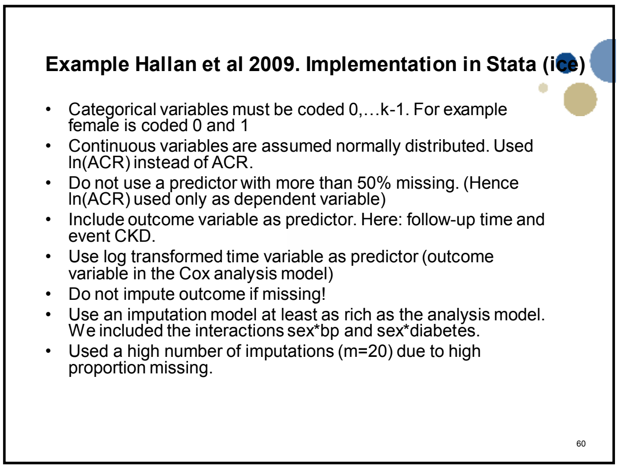


Example

Hallan & al (2009)
“Statistical analyses were performed using Stata 10.0 (Stata Corp., TX, U.S.A.). In general, there were few missing data (<2% for most variables, see Table 1), but data on ACR were, by study design, available only in a subgroup. Multiple imputation is now considered the standard method for handling this type of data,(Clark & Altman 2003;Donders et al. 2006;Rassler et al. 2008;van Buuren et al 1999) whereas complete case analysis would yield too imprecise as well as biased results. The multiple imputation technique estimates the mean and uncertainty of the missing data using all information from the actually observed data in a proper way. In this way, unbiased estimates with the correct standard deviation and p-values are calculated.(Rassler et al, 2008). …



Continued:
“… For most non-diabetic non-hypertensive subjects data were missing completely at random, and for those not returning urine samples as requested data were assumed to be missing at random, thus meeting the assumptions for the method. The analyses were carried out in the “ice” and “micombine” procedures for Stata,(Royston 2005) ACR was log-transformed and not used as predictor in the imputation of other missing variables,(van Buuren et al 1999b) study outcome variables were included in the imputation model,(Moons et al. 2006) and the time variable was log-transformed.(van Buuren, et al 1999a) Regression modelling revealed interactions between sex and both blood pressure and diabetes mellitus. Hence, these two interactions were included in the imputation model. We used m=20 imputations to achieve maximum accuracy.(Newgard & Haukoos 2007)”



Example Hallan et al 2009. Implementation in Stata (ice)

- Categorical variables must be coded 0,…k-1. For example female is coded 0 and 1
- Continuous variables are assumed normally distributed. Used ln(ACR) instead of ACR.
- Do not use a predictor with more than 50% missing. (Hence ln(ACR) used only as dependent variable)
- Include outcome variable as predictor. Here: follow-up time and event CKD.
- Use log transformed time variable as predictor (outcome variable in the Cox analysis model)
- Do not impute outcome if missing!
- Use an imputation model at least as rich as the analysis model. We included the interactions sex*bp and sex*diabetes.
- Used a high number of imputations (m=20) due to high proportion missing.

60

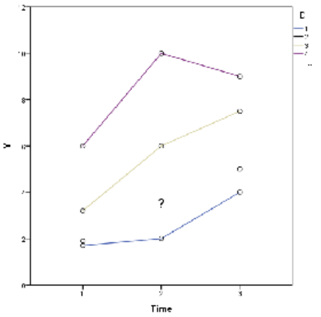
(results from 5 of the imputations in Hallan et al 2009)
Results of Cox Proportional hazard regression. Regression coefficient estimate (SE), p-value, and FMI (Fraction missing information). Results from Stata with commands ice and mim.

	Imputation number					Total, by Rubin's rules	p-value	FMI
Age, years	1 0.0707 (0.0067)	2 0.0705 (0.0067)	3 0.0701 (0.0067)	4 0.0701 (0.0067)	5 0.0707 (0.0067)	0.0704 (0.0067)	<0.001	0.002
Female sex	-0.612 (0.189)	-0.580 (0.190)	-0.570 (0.190)	-0.582 (0.190)	-0.589 (0.190)	-0.587 (0.191)	0.002	0.008
ACR	0.0276 (0.0013)	0.0282 (0.0013)	0.0285 (0.0013)	0.0283 (0.0013)	0.0280 (0.0013)	0.0281 (0.0014)	<0.001	0.082

Full model based analysis:

Full information maximum likelihood and similar

Longitudinal study with missing data:



Longitudinal study with missing data:

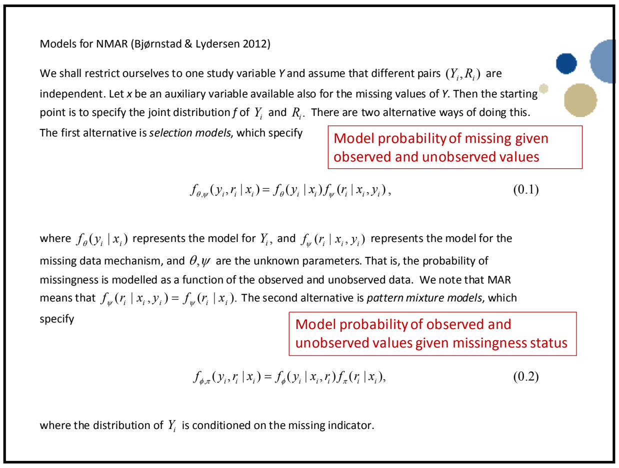
- Repeated measurements ANOVA:
  - Uses data only from subjects with complete data.
  - Not recommended.
- Mixed model:
  - Includes data from the available time points for each individual
  - Uses a full infomation maximum likelihood method (ML) or restricted maximum likelihood (REML) method
  - Recommended method and unbiased if MAR.
- Generalized estimating equations (GEE)
  - May perform better than mixed models when categorical outcomes.
  - Unbiased only if MCAR (Fitzmaurize et al, 2009, page 59)

Full information maximum likelihood (FIML)

- Can be used in a general setting
- Available (almost) only for multivariate normal models.
- Available in structural equation modelling (SEM) software such as Stata and Mplus. FIML is not necessarily default – make sure you use the correct options.

Models for NMAR

- Need to make unverifyable assumptions about the degree of departure from MAR.
- Alternative approaches:
  - Selection models: Specify how the probability of missing depends on the unobserved variable(s)
  - Pattern mixture models: Specify how the distribution of the variable(s) depends on the missingness indicator.
- This is difficult.



Models for NMAR (Bjørnstad & Lydersen 2012)

We shall restrict ourselves to one study variable *Y* and assume that different pairs (Y_i, R_i) are independent. Let *x* be an auxiliary variable available also for the missing values of *Y*. Then the starting point is to specify the joint distribution *f* of Y_i and R_i. There are two alternative ways of doing this. The first alternative is *selection models*, which specify

Model probability of missing given observed and unobserved values

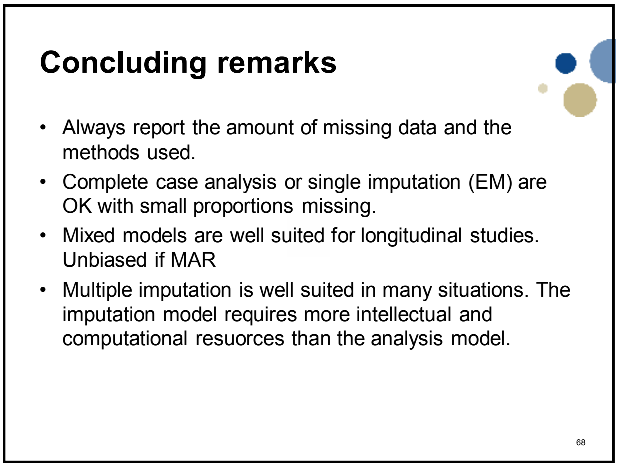
$$f_{\theta,\psi}(y_i, r_i \mid x_i) = f_\theta(y_i \mid x_i) f_\psi(r_i \mid x_i, y_i)\,, \tag{0.1}$$

where $f_\theta(y_i \mid x_i)$ represents the model for Y_i, and $f_\psi(r_i \mid x_i, y_i)$ represents the model for the missing data mechanism, and θ, ψ are the unknown parameters. That is, the probability of missingness is modelled as a function of the observed and unobserved data. We note that MAR means that $f_\psi(r_i \mid x_i, y_i) = f_\psi(r_i \mid x_i)$. The second alternative is *pattern mixture models*, which specify

Model probability of observed and unobserved values given missingness status

$$f_{\phi,\pi}(y_i, r_i \mid x_i) = f_\phi(y_i \mid x_i, r_i) f_\pi(r_i \mid x_i), \tag{0.2}$$

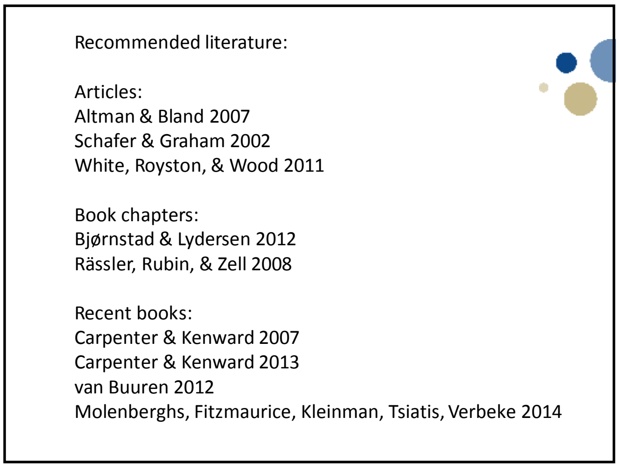
where the distribution of Y_i is conditioned on the missing indicator.



Concluding remarks

- Always report the amount of missing data and the methods used.
- Complete case analysis or single imputation (EM) are OK with small proportions missing.
- Mixed models are well suited for longitudinal studies. Unbiased if MAR
- Multiple imputation is well suited in many situations. The imputation model requires more intellectual and computational resuorces than the analysis model.

68

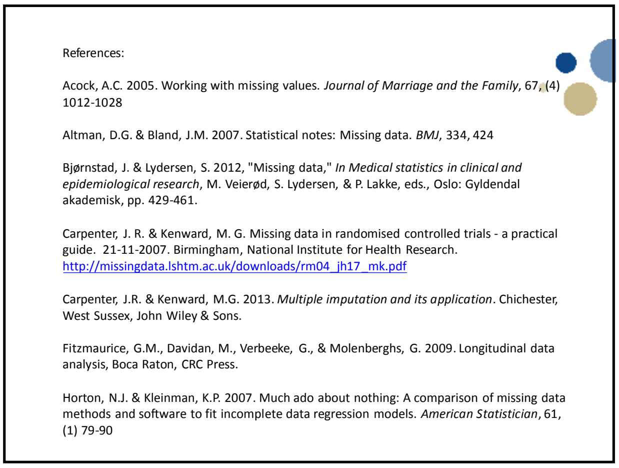


Recommended literature:

Articles:
Altman & Bland 2007
Schafer & Graham 2002
White, Royston, & Wood 2011

Book chapters:
Bjørnstad & Lydersen 2012
Rässler, Rubin, & Zell 2008

Recent books:
Carpenter & Kenward 2007
Carpenter & Kenward 2013
van Buuren 2012
Molenberghs, Fitzmaurice, Kleinman, Tsiatis, Verbeke 2014



References:

Acock, A.C. 2005. Working with missing values. *Journal of Marriage and the Family*, 67, (4) 1012-1028

Altman, D.G. & Bland, J.M. 2007. Statistical notes: Missing data. *BMJ*, 334, 424

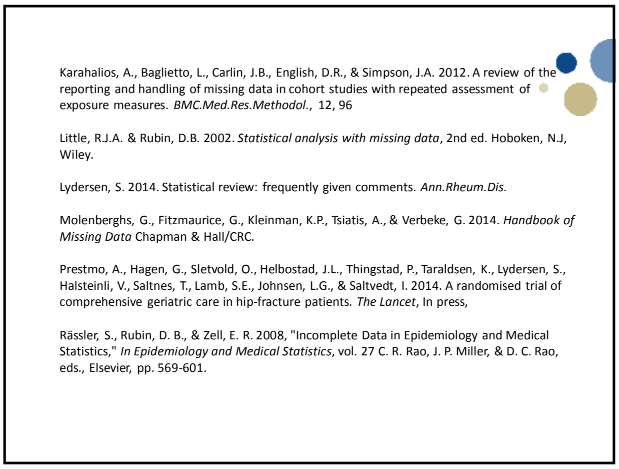
Bjørnstad, J. & Lydersen, S. 2012, "Missing data," *In Medical statistics in clinical and epidemiological research*, M. Veierød, S. Lydersen, & P. Lakke, eds., Oslo: Gyldendal akademisk, pp. 429-461.

Carpenter, J. R. & Kenward, M. G. Missing data in randomised controlled trials - a practical guide. 21-11-2007. Birmingham, National Institute for Health Research. http://missingdata.lshtm.ac.uk/downloads/rm04_jh17_mk.pdf

Carpenter, J.R. & Kenward, M.G. 2013. *Multiple imputation and its application*. Chichester, West Sussex, John Wiley & Sons.

Fitzmaurice, G.M., Davidan, M., Verbeeke, G., & Molenberghs, G. 2009. Longitudinal data analysis, Boca Raton, CRC Press.

Horton, N.J. & Kleinman, K.P. 2007. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician*, 61, (1) 79-90



Karahalios, A., Baglietto, L., Carlin, J.B., English, D.R., & Simpson, J.A. 2012. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC.Med.Res.Methodol.*, 12, 96

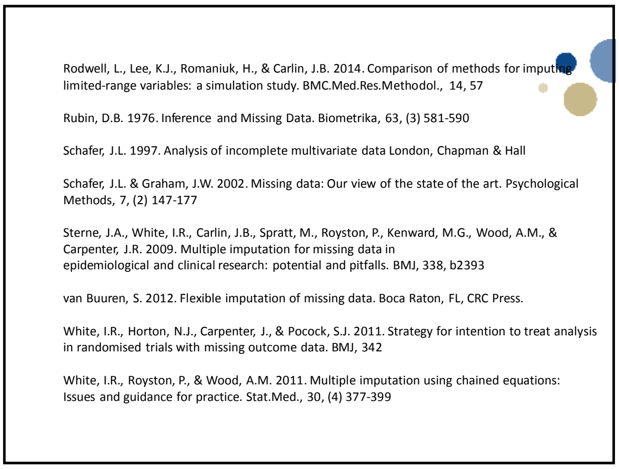
Little, R.J.A. & Rubin, D.B. 2002. *Statistical analysis with missing data*, 2nd ed. Hoboken, N.J, Wiley.

Lydersen, S. 2014. Statistical review: frequently given comments. *Ann.Rheum.Dis.*

Molenberghs, G., Fitzmaurice, G., Kleinman, K.P., Tsiatis, A., & Verbeke, G. 2014. *Handbook of Missing Data* Chapman & Hall/CRC.

Prestmo, A., Hagen, G., Sletvold, O., Helbostad, J.L., Thingstad, P., Taraldsen, K., Lydersen, S., Halsteinli, V., Saltnes, T., Lamb, S.E., Johnsen, L.G., & Saltvedt, I. 2014. A randomised trial of comprehensive geriatric care in hip-fracture patients. *The Lancet*, In press,

Rässler, S., Rubin, D. B., & Zell, E. R. 2008, "Incomplete Data in Epidemiology and Medical Statistics," *In Epidemiology and Medical Statistics*, vol. 27 C. R. Rao, J. P. Miller, & D. C. Rao, eds., Elsevier, pp. 569-601.



Rodwell, L., Lee, K.J., Romaniuk, H., & Carlin, J.B. 2014. Comparison of methods for imputing limited-range variables: a simulation study. BMC.Med.Res.Methodol., 14, 57

Rubin, D.B. 1976. Inference and Missing Data. Biometrika, 63, (3) 581-590

Schafer, J.L. 1997. Analysis of incomplete multivariate data London, Chapman & Hall

Schafer, J.L. & Graham, J.W. 2002. Missing data: Our view of the state of the art. Psychological Methods, 7, (2) 147-177

Sterne, J.A., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., & Carpenter, J.R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338, b2393

van Buuren, S. 2012. Flexible imputation of missing data. Boca Raton, FL, CRC Press.

White, I.R., Horton, N.J., Carpenter, J., & Pocock, S.J. 2011. Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ, 342

White, I.R., Royston, P., & Wood, A.M. 2011. Multiple imputation using chained equations: Issues and guidance for practice. Stat.Med., 30, (4) 377-399