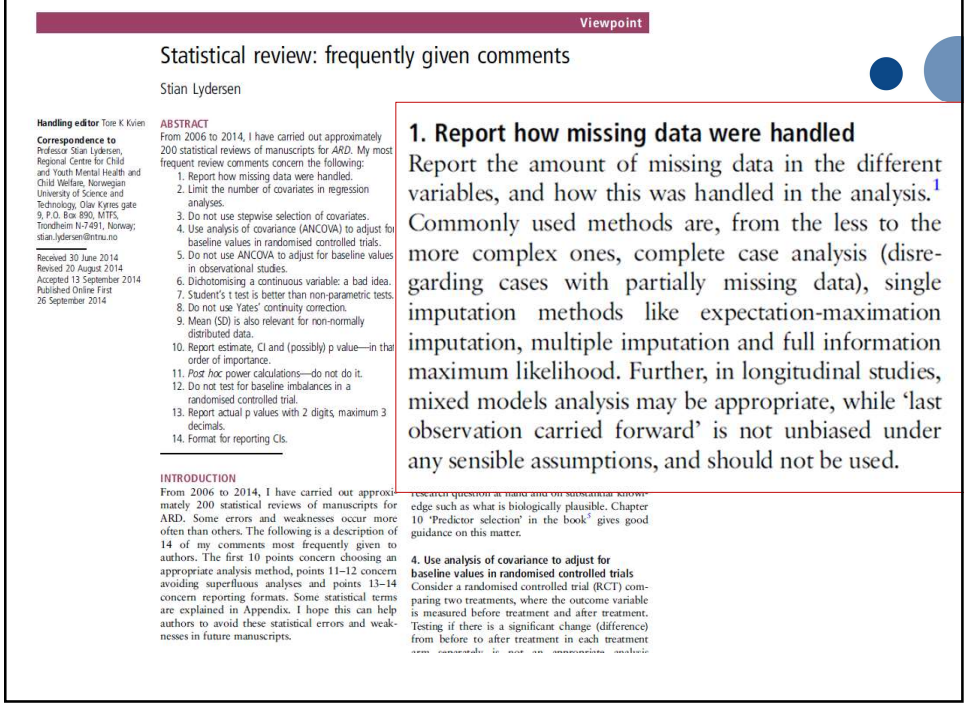
NTNU
Norwegian University of Science and Technology

# How to handle missing data

Stian Lydersen
Professor of Medical Statistics, NTNU

Presented at the Nor-COAST seminar
Hurtigruten Trondheim – Bergen 14 May 2018

http://folk.ntnu.no/slyderse/Hurtigruten14May2018.pdf



Viewpoint

Statistical review: frequently given comments

Stian Lydersen

**Handling editor** Tore K Kvien

**Correspondence to**
Professor Stian Lydersen, Regional Centre for Child and Youth Mental Health and Child Welfare, Norwegian University of Science and Technology, Olav Kyrres gate 9, P.O. Box 890, MTFS, Trondheim N-7491, Norway; stian.lydersen@ntnu.no

Received 30 June 2014
Revised 20 August 2014
Accepted 13 September 2014
Published Online First
26 September 2014

ABSTRACT
From 2006 to 2014, I have carried out approximately 200 statistical reviews of manuscripts for *ARD*. My most frequent review comments concern the following:

1. Report how missing data were handled.
2. Limit the number of covariates in regression analyses.
3. Do not use stepwise selection of covariates.
4. Use analysis of covariance (ANCOVA) to adjust for baseline values in randomised controlled trials.
5. Do not use ANCOVA to adjust for baseline values in observational studies.
6. Dichotomising a continuous variable: a bad idea.
7. Student's t test is better than non-parametric tests.
8. Do not use Yates' continuity correction.
9. Mean (SD) is also relevant for non-normally distributed data.
10. Report estimate, CI and (possibly) p value—in that order of importance.
11. *Post hoc* power calculations—do not do it.
12. Do not test for baseline imbalances in a randomised controlled trial.
13. Report actual p values with 2 digits, maximum 3 decimals.
14. Format for reporting CIs.

INTRODUCTION

From 2006 to 2014, I have carried out approximately 200 statistical reviews of manuscripts for ARD. Some errors and weaknesses occur more often than others. The following is a description of 14 of my comments most frequently given to authors. The first 10 points concern choosing an appropriate analysis method, points 11–12 concern avoiding superfluous analyses and points 13–14 concern reporting formats. Some statistical terms are explained in Appendix. I hope this can help authors to avoid these statistical errors and weaknesses in future manuscripts.

**1. Report how missing data were handled**

Report the amount of missing data in the different variables, and how this was handled in the analysis.[1] Commonly used methods are, from the less to the more complex ones, complete case analysis (disregarding cases with partially missing data), single imputation methods like expectation-maximation imputation, multiple imputation and full information maximum likelihood. Further, in longitudinal studies, mixed models analysis may be appropriate, while 'last observation carried forward' is not unbiased under any sensible assumptions, and should not be used.

research question at hand and on substantial knowledge such as what is biologically plausible. Chapter 10 'Predictor selection' in the book[3] gives good guidance on this matter.

**4. Use analysis of covariance to adjust for baseline values in randomised controlled trials**

Consider a randomised controlled trial (RCT) comparing two treatments, where the outcome variable is measured before treatment and after treatment. Testing if there is a significant change (difference) from before to after treatment in each treatment arm separately is not an appropriate analysis

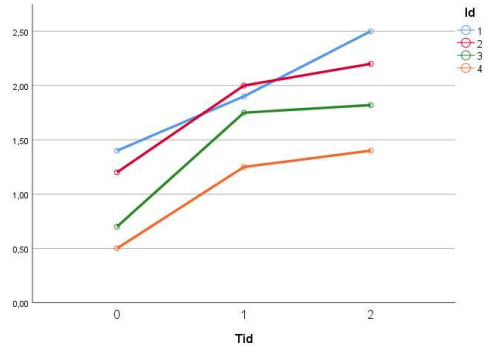
**STROBE Statement—checklist of items that should be included in reports of observational studies:**

- Missing data:
  - Explain how missing data were addressed. (STROBE, Statistical Methods, 12c)
  - Indicate the number of participants with missing data for each variable of interest (STROBE, Descriptive data, 14b)

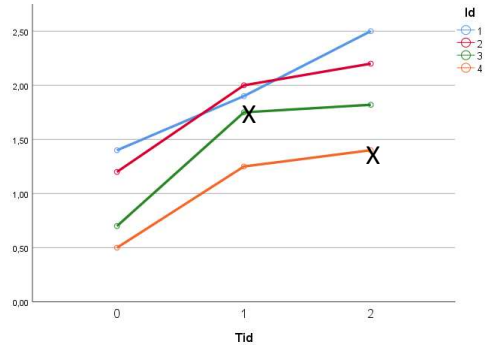
https://www.strobe-statement.org/index.php?id=available-checklists

3

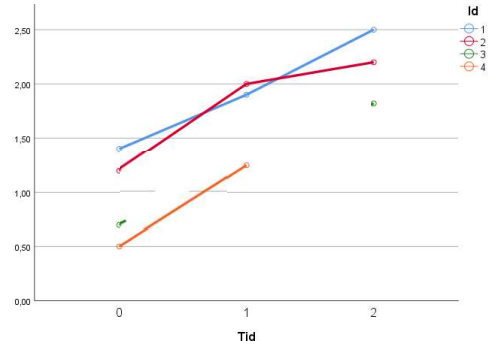
Longitudinal study – complete data



Longitudinal study – missing data



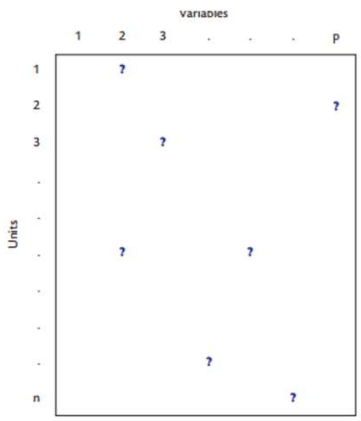

Longitudinal study – missing data



# 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.

7



Types of missing data (Sterne et al. 2009)

- MCAR: 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
- MAR: 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
- MNAR: 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

9

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

## Some methods for handling missing data. (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)

Based on (Bjørnstad & Lydersen 2012)

continues on next slide …

11

## Some methods for handling missing data(continued). (Unbiased when)

- Full model based analysis (for example 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

Based on (Bjørnstad & Lydersen 2012)

12

Missing values and Intention to treat (ITT)

(Carpenter and Kenward 2007), page 24, state: “Thus the ITT interpretation cannot be directly adopted when outcome data are missing (Hollis and Campbell 1999), a fact that appears to remain quite widely misunderstood (Wood et al. 2004).”

References

Carpenter, J. R. & Kenward, M. G.. 21-11-2007. “ Missing data in randomised controlled trials - a practical guide.” Birmingham, National Institute for Health Research. Publication RM03/JH17/MK.
Ref Type: Serial (Book,Monograph)

Hollis, S. & Campbell, F. 1999. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*, 319, (7211) 670-674

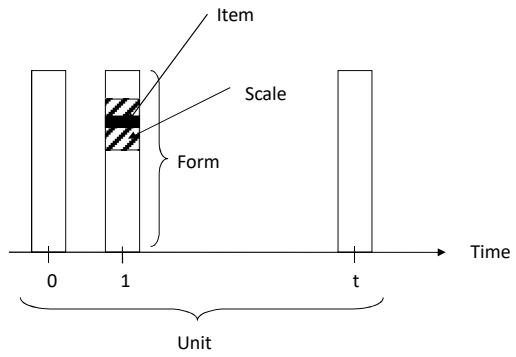
Wood, A.M., White, I.R., & Thompson, S.G. 2004. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin.Trials*, 1, (4) 368-376

**Example: Averaging available items on a scale:**

- 36-Item Short Form Survey (SF-36) is a generic quality of life instrument.
- Consists of eight scales with 2 to 10 items each:
  - physical functioning
  - role limitations due to physical problems
  - bodily pain
  - general health perceptions
  - Vitality
  - Social functioning
  - role limitations due to emotional problems
  - mental health
- Recommended in the manual: On each scale, compute the average score if at least 50% of the items are available

14

Example: Quality of Life questionnaires




Calculating mean in SPSS

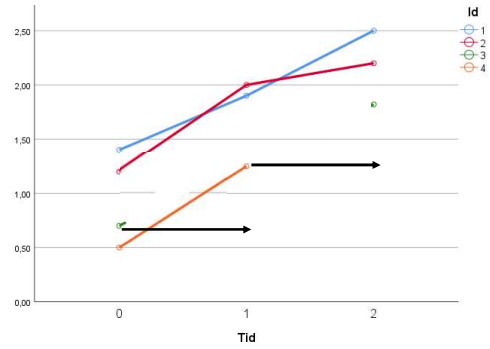
Mean(q21, q22, q23, q24).
Calculates mean if data on at least one variable

(q21+ q22 + q23 + q24)/4
Calculates mean if data are are available for all variables

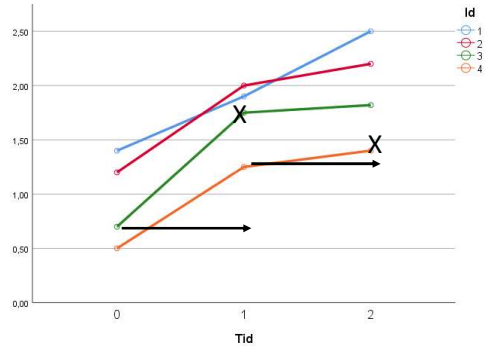
Mean.2(q21, q22, q23, q24).
Calculates mean if data are are available for at least 2 variables

Mean.m(q21, q22, q23, q24).
Calculates mean if data are are available for at least m variables

Longitudinell studie – last observation carried forward (LOCF)




Longitudinell studie – last observation carried forward (LOCF)



"Last observation carried forward" (LOCF) is unfortunately used in many publications, probably due to its simplicity.

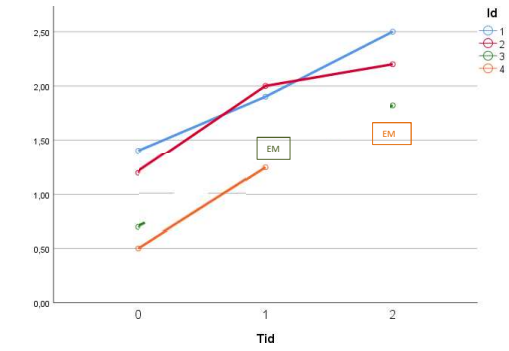
"LOCF" is an assumption that is rarely clinically plausible." (O'Kelly and Ratitch 2014)

"This method is attractive because it is simple, but it has little else to recommend it." (Vickers and Altman 2013)

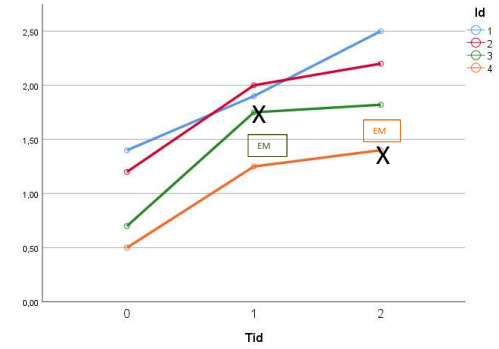
References

O'Kelly, M. & Ratitch, B. 2014. *Clinical trials with missing data a guide for practitioners* Chichester, Wiley.
Vickers, A.J. & Altman, D.G. 2013. Statistics notes: missing outcomes in randomised trials. *BMJ*, 346, f3438

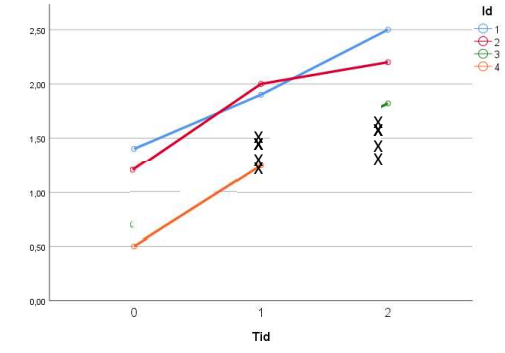
Single EM imputation



Single EM imputation




Multiple imputation



## 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.

23

## 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.

24

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

Bayesian approach with a prior distribution for $\theta$
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.

29

Impute the outcome variable?

Under MAR, there are generally no benefits to impute the outcome, and for a low number of imputations the results may even be somewhat more variable because of simulation error. There is an important exception to this. If we have access to an auxiliary complete variable that is not part of the model and that is highly correlated with the outcome, imputation can be considerably more efficient than complete case analysis, resulting in more precise estimates and shorter confidence intervals. A common scenario where this occurs is if we have a cheap outcome measure for everyone, and an expensive measure for a subset.

In many data sets, missing data also occur in the independent variables. In these cases, we need to impute the outcome variable since its imputed version is needed to impute the independent variables.

https://stats.stackexchange.com/questions/46226/multiple-imputation-for-outcome-variables (Stef van Buuren, 13 Jan 2013)

## Alternative methods for logitudinal data:

- Repeated measures ANOVA
- Mixed models
- Generalized estimating equations (GEE)

Some structural equation models (SEM), including growth models, are generalizations of mixed models, and relevant with latent variables beyond random intercept and random slope models.

31

## Repeated measures ANOVA

- Only complete cases are included in the analysis
- Unbiased only if data are missing completely at random (MCAR)
- The underlying mathematical model is not transparent
- Was an attractive method before computers became powerful (Personal communication with Garrett M. Fitzmaurice)
- Ought to be in the museum.

32

Chance, vil 18, no 3, 2005

*Should we quit using repeated measures analysis of variance?*

# Repeated Measures ANOVA, R.I.P.?


Charles E. McCulloch

It is a difficult experiment to run and to analyze: What are the effects of alcohol on sleepiness and does a hormone, pregnenolone, which has been shown to enhance memory in rat experiments, help alleviate the sleepiness? Each person is tested under each of four conditions on four different visits in random order: a placebo for the drug and for the hormone, alcohol alone, hormone alone, and the combination. Each subject is also queried multiple times within a visit in the minutes after alcohol (or placebo) ingestion. Some subjects drop out of the protocol without completing all the conditions and some of the sleepiness scores are not recorded within a visit because of difficulties executing the protocol. How should the data be analyzed?

This is an opportunity for the professional statistician to pull out any of a number of impressive and more recent tools of the trade: generalized estimating equations, mixed model analyses, imputation, and inverse probability weighting. Gone

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + p_i + \varepsilon_{ijk}, \quad (1)$$

where $\mu$ is the overall mean, $\alpha_j$ represents the alcohol effect, $\beta_k$ the pregnenolone effect, $\alpha\beta_{jk}$ the interaction effect, $p_i$ is the person effect, and $\varepsilon_{ijk}$ is an error term. This model hypothesizes simple person effects that raise or lower (if the effect is negative) the average sleepiness in all four conditions. Interest focuses on the interaction, because the scientific question is whether pregnenolone helps to reduce the sleep-inducing effect of alcohol.

With a mean, error term and four explanatory factors in the model, the analysis of variance would partition the variability in $Y_{ijk}$ into four sources: person, alcohol, pregnenolone, and the interaction. An ANOVA would generate a table as outlined in Table 1. A typical assumption is that the $\varepsilon_{ijk}$ are all independent and follow a $N(0, \sigma^2)$ distribution and that the rest of the terms in (1) are fixed. If these assumptions are correct, the

# Mixed models

- Includes all subjects, also those with missing data at some time point(s)
- Unbiased under the less restrictive missing at random (MAR) assumption
- Transparent mathematical model

34

# GEE:
# Generalized estimating equations

- A useful alternative to Mixed models, especially for categorical outcome such as binary data (logistic regression) or count data (Poisson regression).
- Unbiased only if data are MCAR

35

References

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

Lydersen, S. 2015. Statistical review: frequently given comments. Ann.Rheum.Dis., 74, (2) 323-325

McCulloch, C.E. 2005. Repeated Measures ANOVA, R.I.P.? Chance, 18, (3) 29-33

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

Thoresen, M. 2012, "Longitudinal Analysis," In Medical statistics in clinical and epidemiological research, M. Veierød, S. Lydersen, & P. Laake, eds., Oslo: Gyldendal Akademisk, pp. 259-287.

Thoresen, M. & Gjessing, H. K. 2012, "Mixed Models," In Medical statistics in clinical and epidemiological research, M. Veierød, S. Lydersen, & P. Laake, eds., Oslo: Gyldendal Akademisk, pp. 231-258.

Twisk, J.W.R. 2013. Applied longitudinal data analysis for epidemiology. A practical guide, 2nd ed.

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