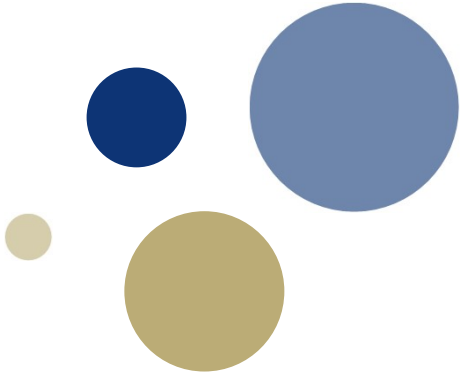
NTNU

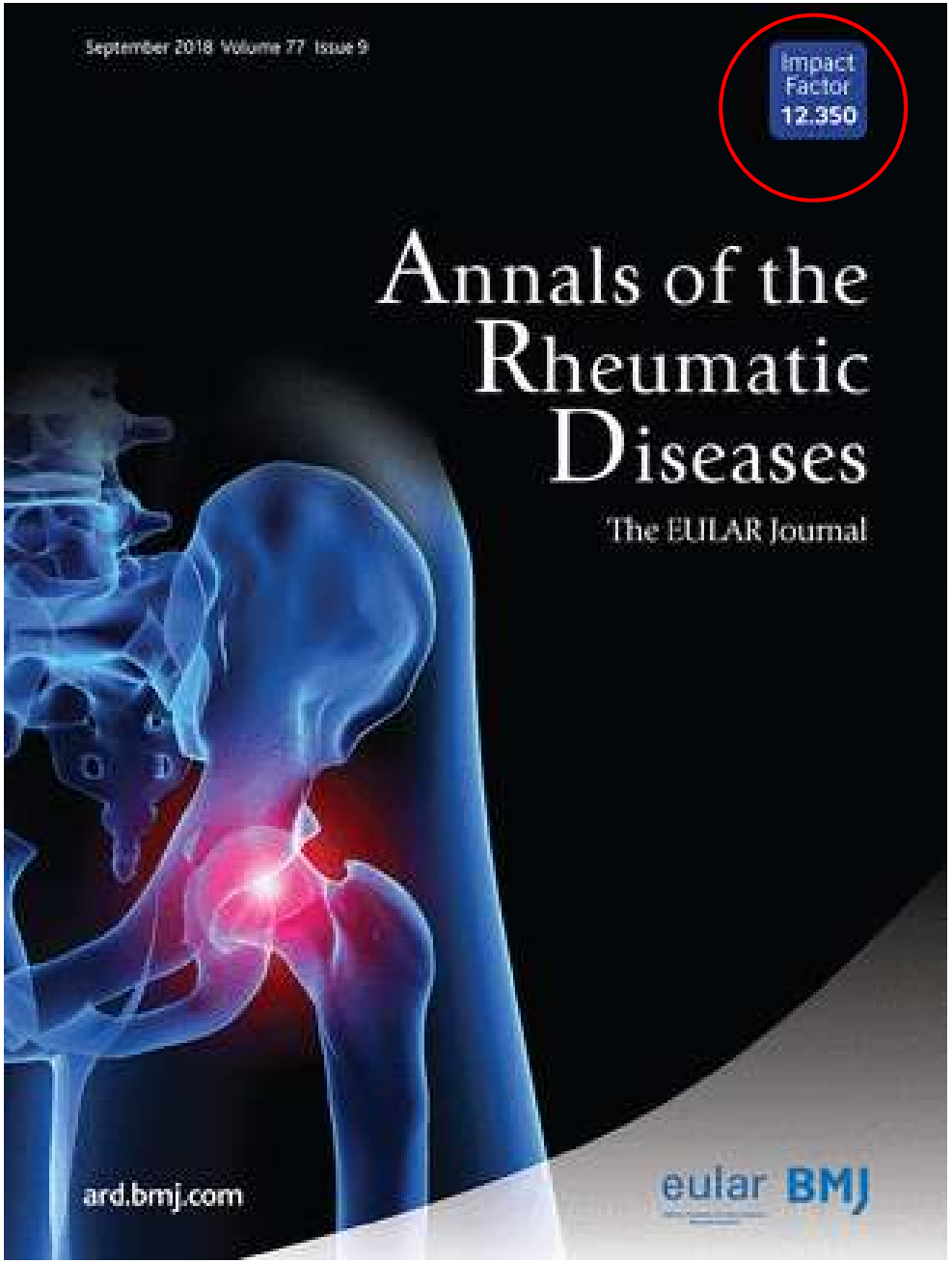
Kunnskap for en bedre verden



Statistical review: Frequently given comments.

Statistisk fagfellevurdering for Annals of the Rheumatic Diseases

av
Stian Lydersen
Presentasjon på forskningslunsj,
RKBU, 19 september 2018


September 2018 Volume 77 Issue 9
Impact Factor 12.350
Annals of the Rheumatic Diseases
The EULAR Journal
ard.bmj.com
eular BMJ

http://ard.bmj.com/pages/authors/#statistical_analysis

From the «Vancouver guidelines» www.icmje.org:

Statistical analysis

- Describe statistical methods with enough detail to enable the reader to judge its appropriateness for the study and to verify the reported results.
- When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as P values.
- Distinguish prespecified from exploratory analyses, including subgroup analyses.

NEW: For further information on common statistical errors to avoid, please read the article published by ARD's Statistical Advisor, Stian Lydersen.

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

regression model, for example, 17 covariates in a study with 64 cases. Traditional rules of thumb state that the ratio of cases per covariate ought to be in the size of order 10. Some authors recommend 15, some 20, others state that 5 is sufficient. In logistic regression and Cox regression, 10 events per variable is usually sufficient[2] and in many situations 5 events per variable is sufficient.[3] Note that in logistic regression this is not the total number of observations, but the smallest of the two outcome groups. Similarly, in Cox regression, only the number of events excluding censored observations is counted as cases in this context.

3. Do not use stepwise selection of covariates

Automated variable selection procedures like stepwise selection used to be very popular. Today an increasing number of analysts criticise such methods. For example,[4] page 419 states: "There are several systematic, mechanical, and traditional algorithms for finding models (such as stepwise and best-subset regression) that lack logical and statistical justification and that perform poorly in theory, simulations and case studies … One serious problem is that the P-values and standard errors … will be downwardly biased, usually to a large degree".

Selection of covariates should be based on the research question at hand and on substantial knowledge such as what is biologically plausible. Chapter 10 'Predictor selection' in the book[5] gives good

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.

Lydersen, ARD 2015.

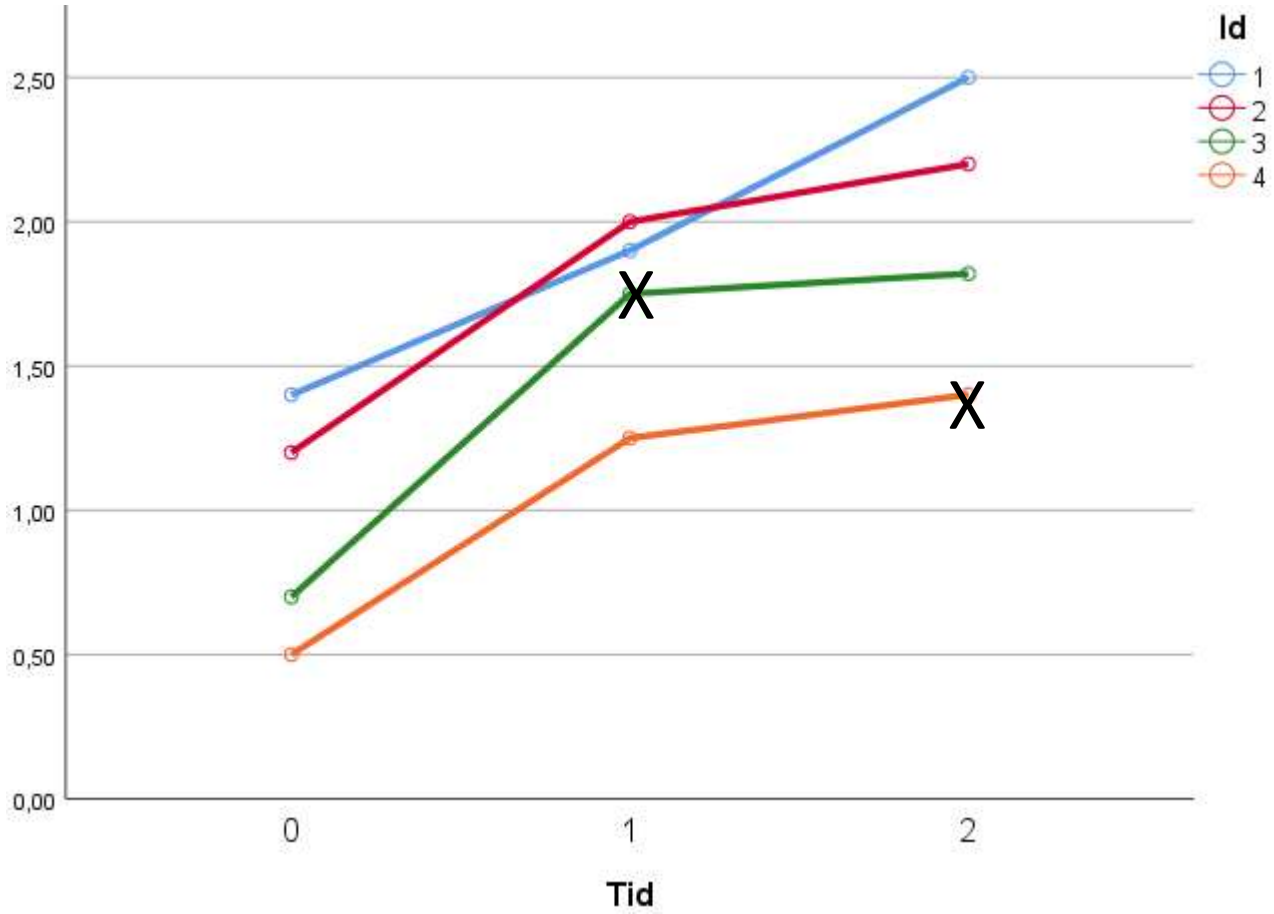
STROBE Statement—checklist of items that should be included in reports of observational studies:

- 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)

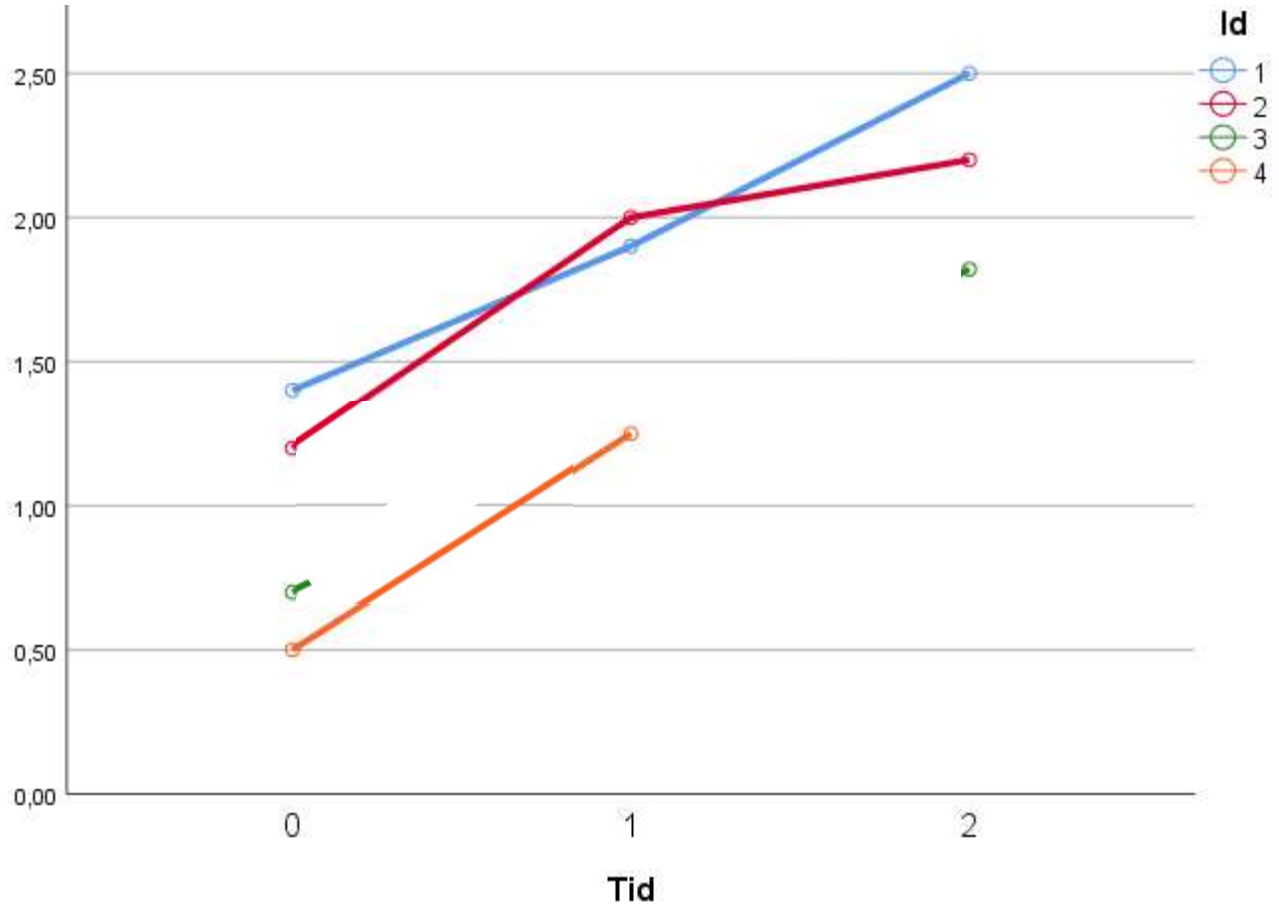
https://www.strobe-statement.org/index.php?id=available-checklists

STROBE: STrengthening the Reporting of OBservational studies in Epidemiology

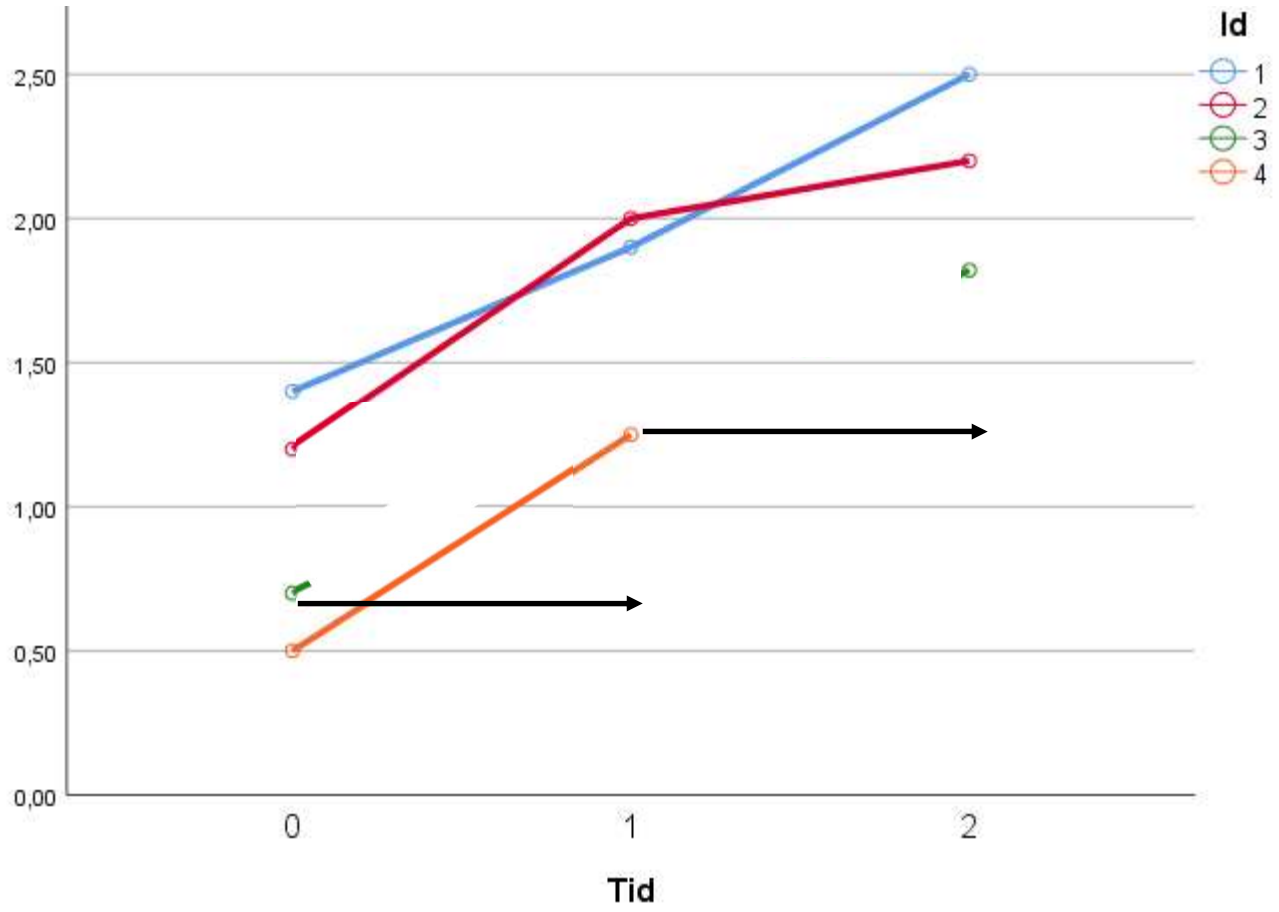
Longitudinell studie – manglende data



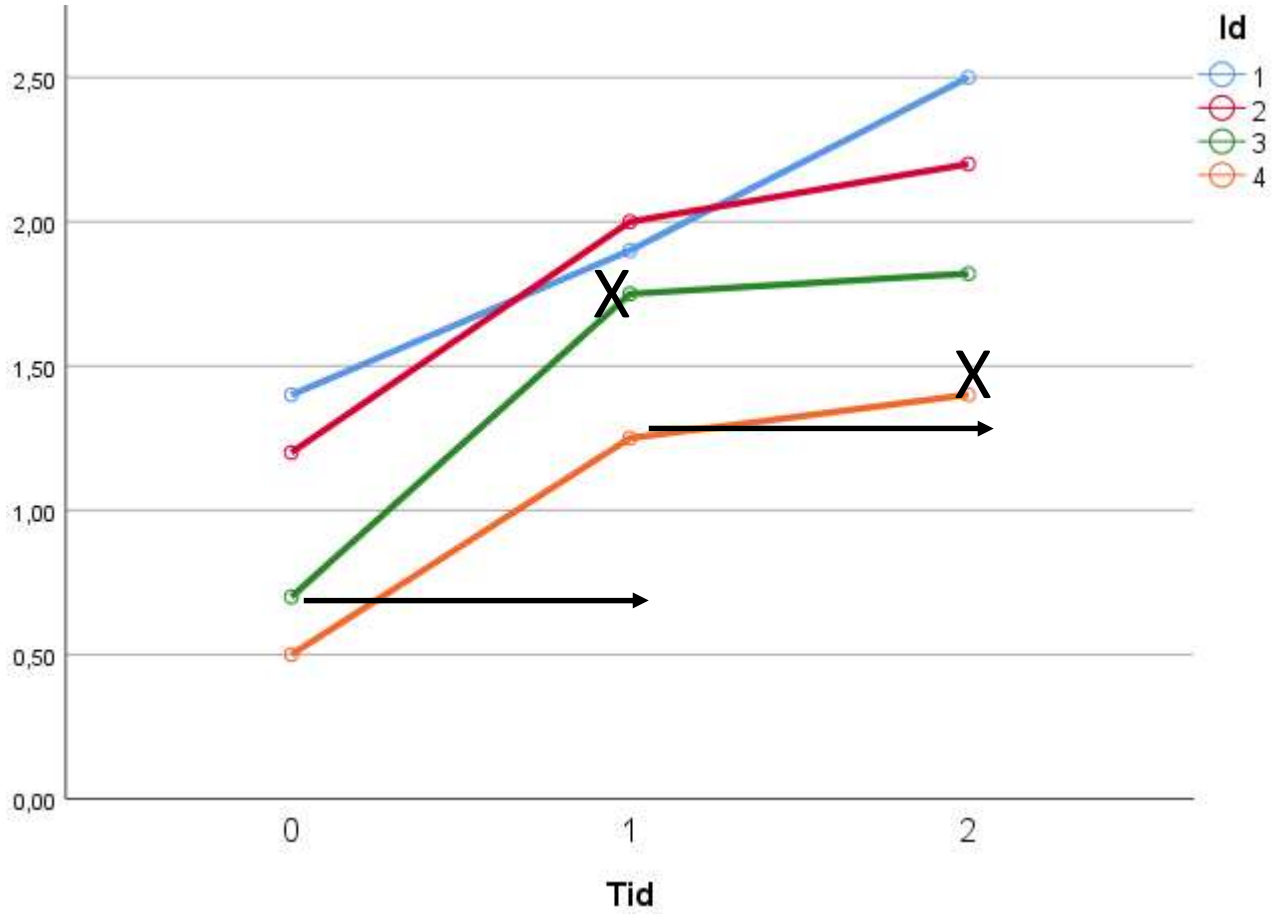
Longitudinell studie - manglende data



Longitudinell studie – last observation carried forward (LOCF)



Longitudinell studie – last observation carried forward (LOCF)



Last observation carried forward (LOCF, LVCF)

As LOCF is neither valid under general assumptions nor based on statistical principles, it is not a sensible method, and should not be used.

Carpenter, J. R. & Kenward, M. G. 2015, "Development of Methods for the Analysis of Partially Observed data and Critique of ad hoc Methods," In Handbook of Missing Data Methodology, G. Molenberghs et al., eds., CRC / Champan Hall, pp. 23-46.

"LOCF" is an assumption that is rarely clinically plausible."

O'Kelly, M. & Ratitch, B. 2014. Clinical trials with missing data a guide for practitioners Chichester, Wiley.

"This method is attractive because it is simple, but it has little else to recommend it."

Vickers, A.J. & Altman, D.G. 2013. Statistics notes: missing outcomes in randomised trials. BMJ, 346, f3438

Last observation carried forward (LOCF, LVCF)

" … LOCF is dubious. The method has long been used in clinical trials. The U.S. Food and Drug Administration (FDA) has traditionally viewed LOCF as the preferred method of analysis, considering it conservative and less prone to selection than listwise deletion. However, ((Molenberghs and Kenward 2007) pp 47 – 50) show that the bias can operate in both directions, and that LOCF can yield biased estimates even under MCAR."

Molenberghs, G. & Kenward, M.G. 2007. *Missing data in Clinical Studies* Chichester, Wiley.
van Buuren, S. 2018. *Flexible imputation of missing data*, 2 ed. Boca Raton, FL, CRC Press.

2. Limit the number of covariates in regression analyses

Some authors attempt to include too many covariates compared with the number of cases in a regression model, for example, 17 covariates in a study with 64 cases. Traditional rules of thumb state that the ratio of cases per covariate ought to be in the size of order 10. Some authors recommend 15, some 20, others state that 5 is sufficient. In logistic regression and Cox regression, 10 events per variable is usually sufficient[2] and in many situations 5 events per variable is sufficient.[3] Note that in logistic regression this is not the total number of observations, but the smallest of the two outcome groups. Similarly, in Cox regression, only the number of events excluding censored observations is counted as cases in this context.

3. Do not use stepwise selection of covariates

Automated variable selection procedures like stepwise selection used to be very popular. Today an increasing number of analysts criticise such methods. For example,[4] page 419 states: "There are several systematic, mechanical, and traditional algorithms for finding models (such as stepwise and best-subset regression) that lack logical and statistical justification and that perform poorly in theory, simulations and case studies … One serious problem is that the P-values and standard errors … will be downwardly biased, usually to a large degree".

Selection of covariates should be based on the research question at hand and on substantial knowledge such as what is biologically plausible. Chapter 10 'Predictor selection' in the book[5] gives good guidance on this matter.

Stepwise procedures give biased regression coefficients (the coefficients for remaining variables are too large); see (Tibshirani 1996)

(Katz 2006) In Preface: “Writing a second edition has given me the privilege of updating my thinking on multivariable analysis. The biggest change from the prior edition is that I have gone from being an “agnostic” on the topic of automatic variable selection algorithms (e.g. forward stepwise selection) to being against using them for explanatory models”

Katz, M.H. 2006. Multivariable analysis a practical guide for clinicians, 2nd ed. Cambridge, Cambridge University Press.
Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society.Series B, 58, (1) 267-288

Annals of Internal Medicine state in their instruction to authors (http://annals.org/aim/pages/AuthorInformationStatisticsOnly):

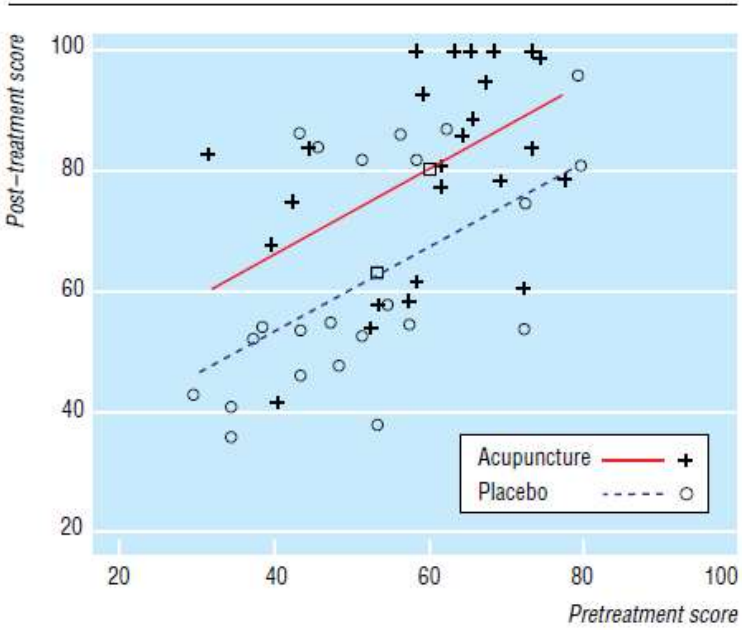
"Model building: Authors should avoid stepwise methods of model building, except for the narrow application of hypothesis generation for subsequent studies. Stepwise methods include forward, backward, or combined procedures for the inclusion and exclusion of variables in a statistical model based on predetermined P value criteria. Better strategies than P value driven approaches for selecting variables are those that use external clinical judgment. …"

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 method. One can compare the mean change between the treatment arms. But an even better approach is regression with outcome after treatment as dependent variable, and baseline value and treatment group as covariates.[6] This method is often called analysis of covariance (ANCOVA).

$$Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \beta_2 Intervention + \ldots + \varepsilon_i$$

Vickers & Altman, BMJ, 2001:



Pretreatment and post-treatment scores in each group showing fitted lines. Squares show mean values for the two groups. The estimated difference between the groups from analysis of covariance is the vertical distance between the two lines

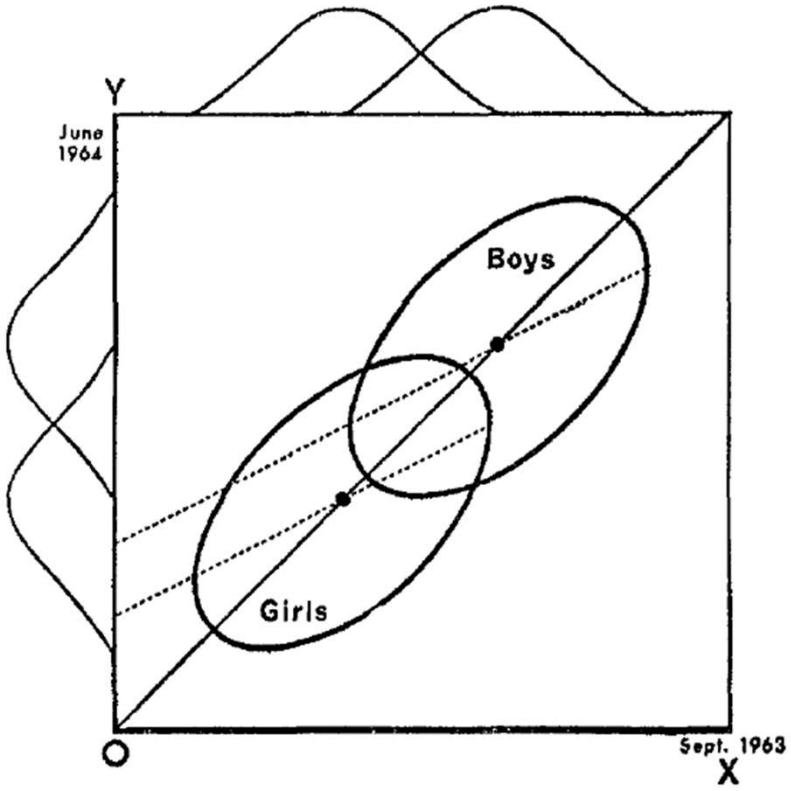
5. Do not use ANCOVA to adjust for baseline values in observational studies

In an observational study, on the other hand, use of ANCOVA cannot be generally recommended[7] (page 126). In fact, ANCOVA can produce different conclusions than analysing a score difference (after score minus before score), a phenomenon also known as Lord's paradox.[8] A central issue is that in a randomised trial, the treatment is applied after measuring the baseline score. Hence the treatment cannot have affected the baseline score. In an observational study, the exposure may also have been present before the baseline score was measured. Then, ANCOVA would generally introduce bias. See also ref 9.

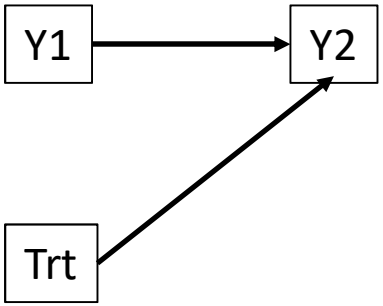
Psychological Bulletin
1967, Vol. 68, No. 5, 304–305

A PARADOX IN THE INTERPRETATION OF GROUP COMPARISONS

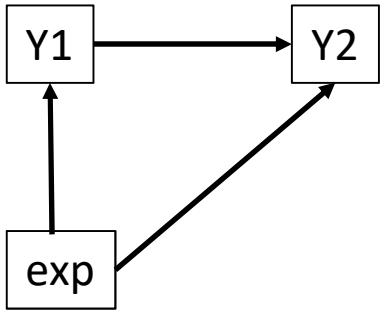
FREDERIC M. LORD



RCT:



Observational study:



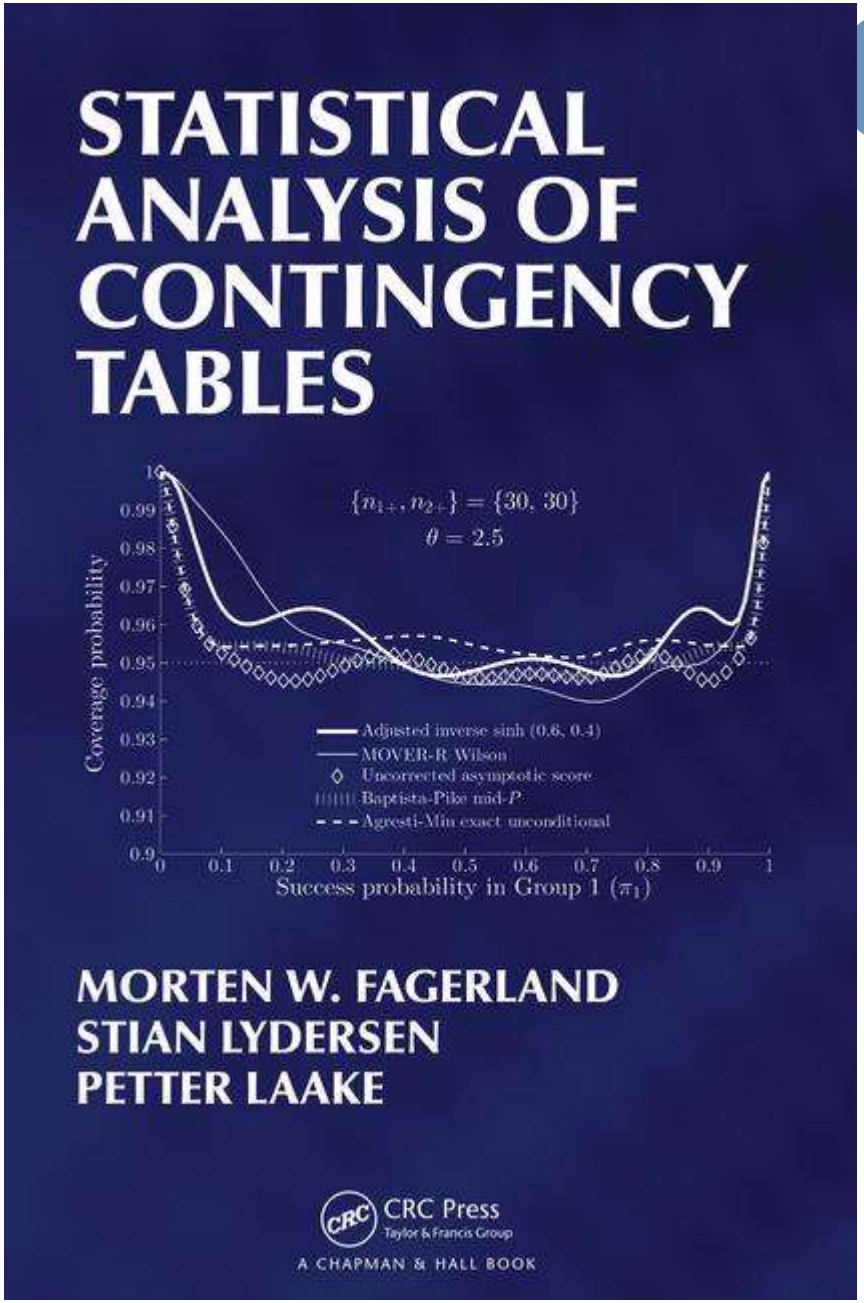
6. Dichotomising a continuous variable: a bad idea

Avoid dichotomising continuous variables if possible.[10–12] Dichotomising implies loss of information and hence loss of statistical power. Moreover, dichotomizing a covariate implies that the effect of that covariate is a step-function changing only at the threshold. In reality, most effects are smooth functions of the covariate. However, sometimes it can be sensible to dichotomise according to some predefined clinical threshold. Data-driven categorisation such as above/below the median of the observations is never a good idea. The same arguments are valid for categorising into more than two categories, although the harm is then somewhat less than by dichotomising.

See also:

Fagerland, M., Lydersen, S., & Laake, P. 2017. *Statistical Analysis of Contingency Tables.* Chapman and Hall/CRC.

Section 13.5: Categorization of Continuous Variables



7. Student's t test is better than non-parametric tests

Student's t test has major advantages over non-parametric tests such as the Wilcoxon test[13]: First, the method allows to compute a CI for the mean of interest, not only a p value. Second, Student's t test is more powerful, particularly in small samples.[14] A widespread misunderstanding is that Student's t test should not be used in small samples. Third, Student's t test is readily generalised into regression analysis and other analyses.

Student's t test is rather robust to deviations from normality[15] as long as there are no residuals extremely distant, say much more than 4–5 SDs, from zero. Visual inspection of Q-Q plots is well suited to detect such deviations. Visual inspection of P-P plots is *not* suited for detecting such deviations. When the data deviate substantially from the normal distribution, one can for example, use bootstrapping to obtain CIs and p values.[16] Bootstrapping has been available in standard statistical software for several years, and is an underused technique in many applications of statistics.

See also:
Fagerland, M.W. 2012. t-tests, non-parametric tests, and large studies-a paradox of statistical practice?
BMC.Med.Res.Methodol., 12, 78

8. Do not use Yates' continuity correction

Many methods have been proposed for testing equality of two proportions. A traditional recommendation is to use Pearson's asymptotic χ^2 test without Yates' correction in 'large' samples, say all expected cell counts are at least five, else, use a small sample method such as Fisher's exact test. Some authors use Pearson's test with Yates' correction. But Yates' correction should be regarded as a historic curiosity from the time before computers were commonly available, and it should never be used.[17 18] Similarly, the version of Yates correction for CIs should never be used.[19] Further recommendations are given in refs 20 and 21.

See also:

Lydersen, S., Fagerland, M.W., & Laake, P. 2009. Recommended tests for association in 2 x 2 tables. *Stat.Med.*, 28, (7) 1159-1175

Highly Cited Paper

Lydersen, S., Langaas, M., & Bakke, Ø. 2012. The Exact Unconditional z-pooled Test for Equality of Two Binomial Probabilities: Optimal Choice of the Berger and Boos Confidence Coefficient. *Journal of Statistical Computation and Simulation*, 82, (9) 1311-1316

Fagerland, M.W., Lydersen, S., & Laake, P. 2015. Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*, 42, (2) 224-254

And …

Fagerland, M., Lydersen, S., & Laake, P. 2017. *Statistical Analysis of Contingency Tables.* Chapman and Hall/CRC.

Chapter 4: The 2x2 Table

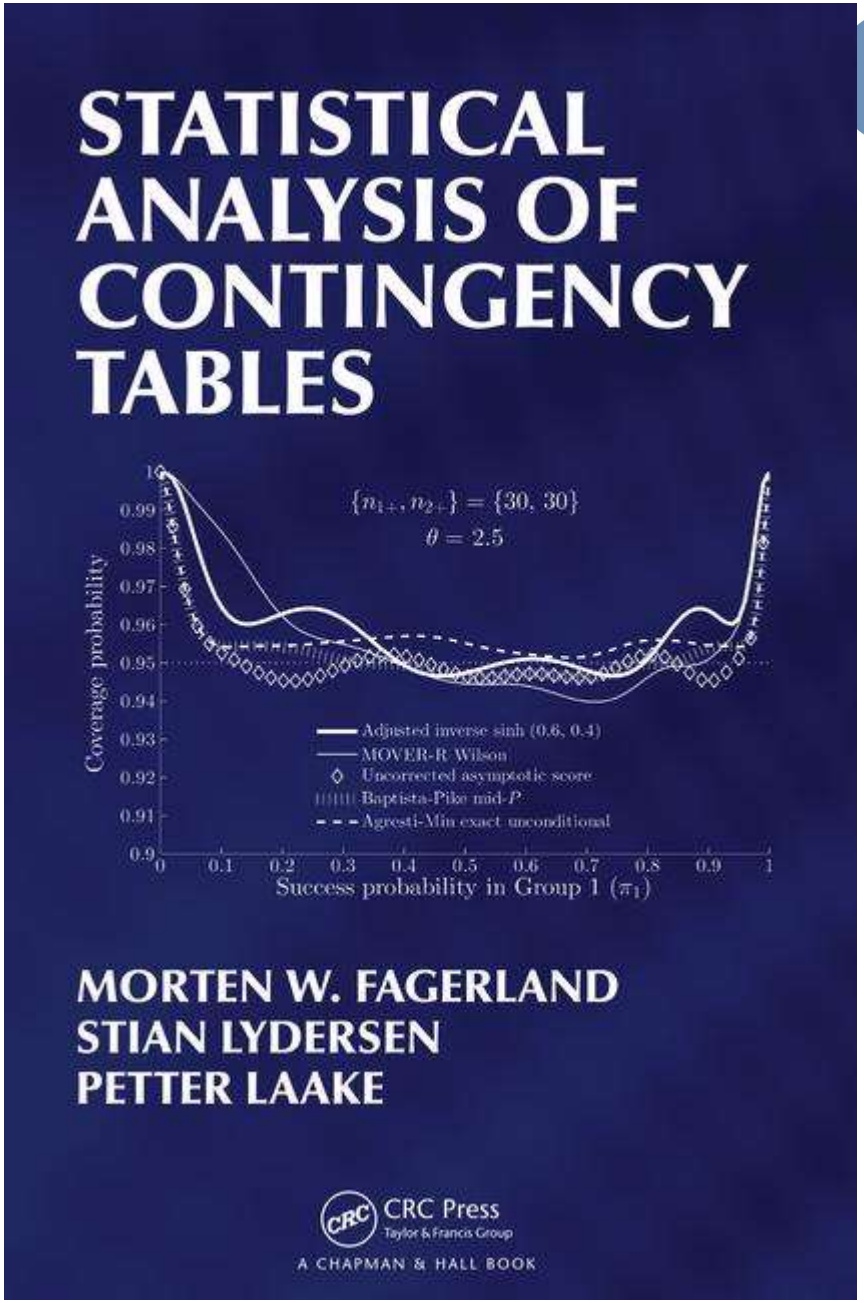


TABLE 4.24

Recommended tests and confidence intervals (CIs) for 2×2 tables

Analysis	Recommended methods	Sample sizes
Tests for association	Fisher mid-P*	all
	Suissa-Shuster exact unconditional†	small/medium
	Fisher-Boschloo exact uncond.†	small/medium
	Pearson chi-squared*	large
CIs for difference between probabilities	Agresti-Min exact unconditional†	small/medium
	Agresti-Caffo*	medium/large
	Newcombe hybrid score*	medium/large
	Miettinen-Nurminen asympt. score	medium/large
	Wald*	large
CIs for number needed to treat	The reciprocals of the limits of the recommended intervals for the difference between probabilities	
CIs for ratio of probabilities	Adjusted inverse sinh*	all
	MOVER-R Wilson*	all
	Koopman asymptotic score	all
	Agresti-Min exact unconditional†	small/medium
	Katz log*	large
CIs for odds ratio	Adjusted inverse sinh*	all
	MOVER-R Wilson*	all
	Baptista-Pike mid-P	all
	Agresti-Min exact unconditional†	small/medium
	Woolf logit*	large

*These methods have closed-form expression

†Preferably with the Berger and Boos procedure ($\gamma = 0.0001$)

Fagerland, Lydersen, Laake (2017)

9. Mean (SD) is also relevant for non-normally distributed data

The mean and SD are meaningful descriptive statistics for data following all types of continuous distributions and sometimes even for ordinal data, not only the normal distribution. A widespread misunderstanding is that one *must* use other measures such as median and IQR if data do not follow the normal distribution. In fact, the mean and SD have several favourable properties. For example, the mean and SD from different studies can readily be combined in a possible later meta-analysis. This is not the case for the quantile-related measures.

10. Report estimate, CI and (possibly) p value—in that order of importance

p Values are overused and overemphasised in medical research as well as many other applied sciences. This problem is well described in a recent article in *Nature*[22] and its accompanying editorial.[23] Sometimes authors report *only* the p value, for example: "Patients exposed to E were more likely than the unexposed to develop the disease D (p=0.04)". The 'Vancouver'-guidelines http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html#d state the following: "When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as p values, which fail to convey important information about effect size and precision of estimates".

11. *Post hoc* power calculations—do not do it

Post hoc power calculations are futile, although it has been recommended by some journals. Power is the probability of rejecting the null hypothesis in a (future) study. Once the study has been conducted, this probability is either 1 (if the null hypothesis was rejected) else 0. *Post hoc* power is fundamentally flawed.[24] After the study, meaningful quantifications of uncertainty are CIs and p values.[24 25]

12. Do not test for baseline imbalances in a RCT

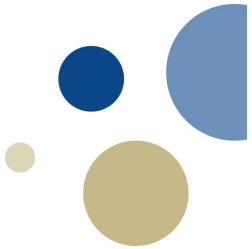
When reporting a RCT, it is recommended to show a table with baseline demographic and clinical characteristics for each treatment group. But testing for baseline imbalances in a properly randomised trial is futile, although reported in some medical journal articles. Such testing is discouraged by the CONSORT guidelines.[26] Assuming that randomisation has been done properly, we can expect 5% of the baseline variables to differ significantly between the groups (at level 5%), see also refs 27 and 28.

13. Format for reporting CIs

Commonly used separators between confidence limits are comma(,), semicolon(;) and hyphen(-). The comma and hyphen should be avoided, since they resemble a decimal separator, a thousands separator, or a minus sign. A good choice is to use 'to', for example, (0.16 to 0.25), as recommended by refs 29 and 30. The same advice applies for other intervals, such as IQR and minimum to maximum values.

14. Report actual p values with 2 digits, maximum 3 decimals

Avoid reporting p values as n.s. or $p<0.05$ or $p<0.01$. The exception is extremely small p values, which ought to be reported as, for example, $p<0.001$. A much used recommendation is to report p values with up to 2 significant digits and maximum 3 decimals, such as $p=0.12$, $p=0.035$, $p=0.006$ and $p<0.001$.



Other statistical issues?

Adjust for multilple hypotheses? When and how?

Multiplicity adjustment continues to be a field of much research and controversy (Senn 2007). In fact, the influential epidemiologist Kenneth Rothman argues against multiplicity adjustment in many settings (Rothman 1990;Rothman 2014).

Rothman, K.J. 1990. No adjustments are needed for multiple comparisons. *Epidemiology*, 1, (1) 43-46

Rothman, K.J. 2014. Six persistent research misconceptions. *J.Gen.Intern.Med.*, 29, (7) 1060-1064

Senn, S. 2007. *Statistical issues in drug development*, 2nd ed. Chichester, England, John Wiley & Sons.

Make data visible / available?

Give numeric results not only as derivatives (for example, percentages) but also as the absolute numbers from which the derivatives were calculated …

www.icmje.org

Availability of data file(s)?


"That's
all
folks!"