Reporting statistics: How to survive Statistical Peer Review

(and a World Cup penalty shootout)

Michael Schlüssel
Medical Statistican & Research Fellow @EQUATORNetwork

EQUATOR Lightning Workshops
4th July 2018

What it is and what it is not

It is not:

- To tell you the best/novel approaches to analyse data
- To teach how to write a response letter
- To show you how to score a goal

It is:

- Tips and advices from own and others experiences
- A (non-exhaustive) list of basic things we often forget
- A practical example of how to mislead with meaningless stats

2

The peer-review process

- Now often includes a review by a statistician, particularly for the large general medical journals
  - e.g. BMJ, Lancet, NEJM, PLoS Med, Ann Intern Med, etc.
- But also leading specialty journals
  - e.g. Eur Heart J, BJOG, Anesthesiolgy, Neurosurgery, etc.
- Statistical editors
  - are (or can be) part of the decision-making process.
  - have an aim to identify errors, improve clarity and ultimately the quality of the study.
  - **are eager for good research!**

Relevance: The research question

(make use of the PICO strategy)

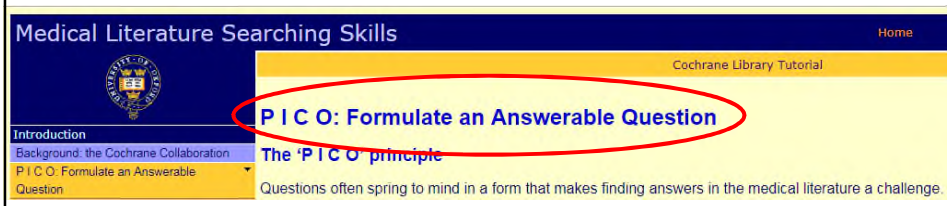


Table 2 – Description of the PICO strategy

Acronym	Definition	Description
P	Patient or problem	Can be only one patient, a group of patients with a particular condition or a health problem
I	Intervention	Represents the intervention of interest, which can be therapeutic (e.g. several kinds of dressings), preventive (e.g. vaccination), diagnostic (e.g. blood pressure measure), prognostic, administrative or related to economic issues
C	Control or comparison	Defined as a standard intervention, the most used intervention or no intervention
O	Outcome	Expected result

Time

Study

4

Good research

- Relevant (in terms of the research question)
- Accurate (in terms of the methods)
- Complete (about conduct and findings)
- Transparent (about limitations and pitfalls)
- Sensible/responsible (in terms of conclusions; i.e. inferences and/or usefulness)

Good research

- A good research question is **not enough**!
- Good research requires appropriate statistical methods and is reported clearly and in full, otherwise it will:
  - **annoy** the statistical reviewer/editor
  - **likely fail** getting through peer review
  - have a **lack of credibility** of the study findings

Statistical Analysis Plan

- Strongly recommended having one!
  - A structured stand-alone document detailing all intend analysis
  - A requirement for RCTs
- Consider publishing one (and then refer back to it)!
  - Journals are more receptive to them (e.g. Trials, BMC Methods, BMJ Open)
- At least include in a protocol paper (and then refer back to it)!

7

Statistical Analysis Plan

- Strongly recommended having one!
  - A structured stand-alone document detailing all intend analysis
  - A requirement for RCTs
- Consider publishing one (and then refer back to it)!
  - Journals are more receptive to them (e.g. Trials, BMC Methods, BMJ Open)
- At least include in a protocol paper (and then refer back to it)!

Most importantly, tell the whole story!!
(describe changes and give reasons)

8

Statistical Analysis [Methods]

"Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data […] to verify the reported results" [ICMJE]

At a minimum, enable other scientists to replicate the analysis with different data.

Statistical Analysis [Methods]

- All statistical analyses described in the methods section should have a corresponding set of results (and *vice versa*)
- Explain the purpose of your analysis
- If the analysis is using complex statistical methods, then a citation to appropriate papers should be given
  - pay special attention to how these are described
  - further (or technical) details can be presented in supplementary or online material (if necessary)

Statistical Analysis [Methods]

"t-tests were used for comparisons of continuous variables and Fisher's Exact test or Chi-squared test (where appropriate) were used for comparisons of binary variables"

Rather vague!

"The **primary outcome**, time to readiness for PACU discharge eligibility, was analysed using a two-sample **Wilcoxon rank-sum test**. The **secondary outcomes** of voiding and vasopressor/glycopyrrolate use were analysed using the **Chi square and Fisher's exact test, respectively**, and the **secondary outcome** of time to recovery of S2 sensation was analysed using a **two-sample Wilcoxon rank-sum test**. All analyses were carried out on a **per protocol basis** using SAS version 9.1 (SAS Institute, Cary, NC, USA)."

Much better!

Statistical Analysis [Methods]

- All statistical analyses described in the methods section should have a corresponding set of results (and *vice versa*)
- Explain the purpose of your analysis
- If the analysis is using complex statistical methods, then a citation to appropriate papers should be given
  - pay special attention to how these are described
  - further (or technical) details can be presented in supplementary or online material (if necessary)

Statistical Analysis [Methods]

- Multivariable analyses (i.e. regression) should be clearly explained (e.g. multiple [linear], logistic, Cox)
- Specify the outcome being analysed in the regression model
- Specify all variables included in the regression analysis
  - are all important key confounders (prognostic variables) being adjusted for?
- Specify whether and how variables were selected for inclusion in the model

Statistical Analysis [Methods]

- Using the wrong name for statistical methods or statistical terms will annoy the statistical reviewer, e.g.
  - multivariate instead of multivariable
  - variance analysis instead of analysis of variance
  - decile instead of tenth; quintile instead of fifth
    - Deciles, quintiles are cut-points to create equal sized groups

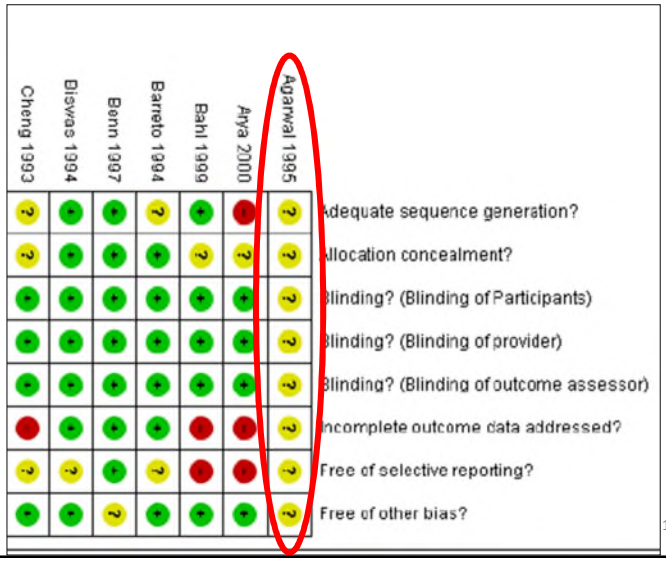
Sample Size

- Should be reported in sufficient detail to enable replication
  - stats reviewers/editors will often (I do) try to replicate the calculation based on what was described in the paper
  - report the type I error (alpha), type II error (power)
  - report all estimates based (ideally from previous research, pilot study) used in the sample size calculation
    - defines your clinically important difference

If no sample size calculation was done, then say so!

Risk of bias assessment

(good research poorly reported or imperfect research well reported?)



16

Missing data

- Information is rarely collected and available for all participants in your study
  - patients drop out, forms are not completed/returned, etc.
- Analysing only those with 'complete' data *can* lead to biased results
  - is there anything special about those who had missing data and therefore excluded?
  - do they have better outcomes compared to those with incomplete data?

Failing to mention missing data is yet another way to annoy your statistical reviewer!
(a flow diagram can help)

Missing data

- DESIGN:
  - Did you anticipate drop-out prior to starting your study and adjust your sample size accordingly?
  - If yes, then clearly report this in the sample size section
- METHODS & RESULTS:
  - When describing your data:
    - how many participants had missing data?
    - what was missing?
  - What did you do with the missing data?
    - Omit them from the analysis?
    - Impute them?

Steps of Initial Data Analysis

Data cleaning: aimed at identifying and correcting errors in the data.

Data screening: understanding the properties of the data that may affect future analysis and interpretation.

Data manipulation: transformation of variables, imputation of missing data, (re)categorisation of variables.

Notes:

I. IDA provides relevant insights obtained from data cleaning/screening (e.g.; publication of preliminary results, early reports of clinical trials);

II. Sometimes findings of IDA help refining and updating analysis plan;

III. Relevant findings and steps DO impact interpretations!

Results: Describing data

- Describe characteristics of your data
  - often a "Table 1" in an article
  - report means (standard deviations)
    - Normally distributed data
  - or report medians (interquartile ranges)
    - non-Normally distributed data
  - make sure tables add up!
    - e.g. columns for group A, group B and total
  - Report n (%) for binary or categorical data

In the Results section, report mean (SD), not mean ± SD

- If the primary analysis involves comparing groups describe the characteristics of each group

A picture is worth a thousand words!

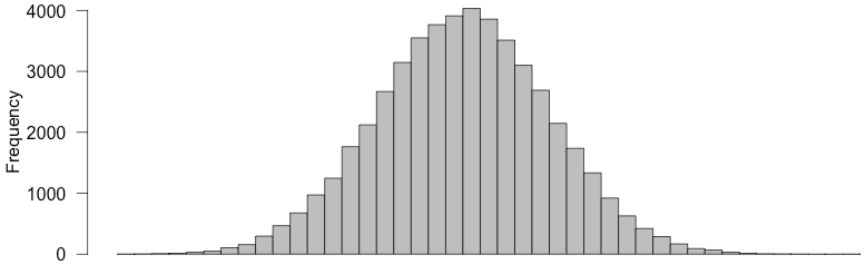
"...make **both** calculations **and** graphs. Both sorts of output should be studied; each will contribute to understanding."

F. J. Anscombe, 1973

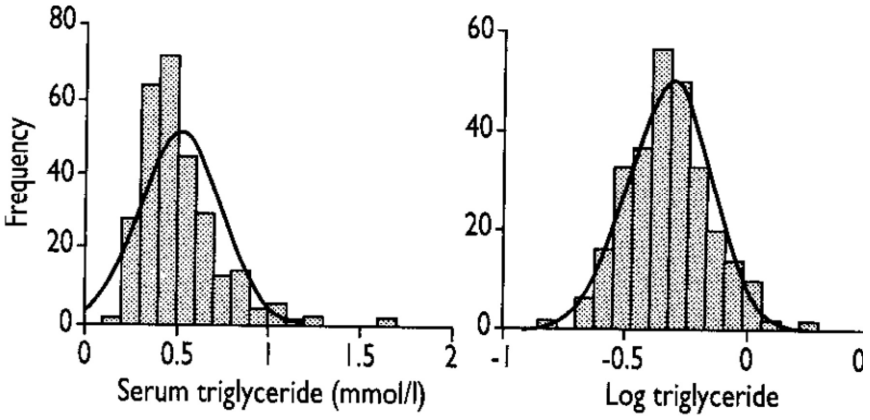
21

A hat or the normal distribution?

Unfortunately, not all data look like this…



Data transformation

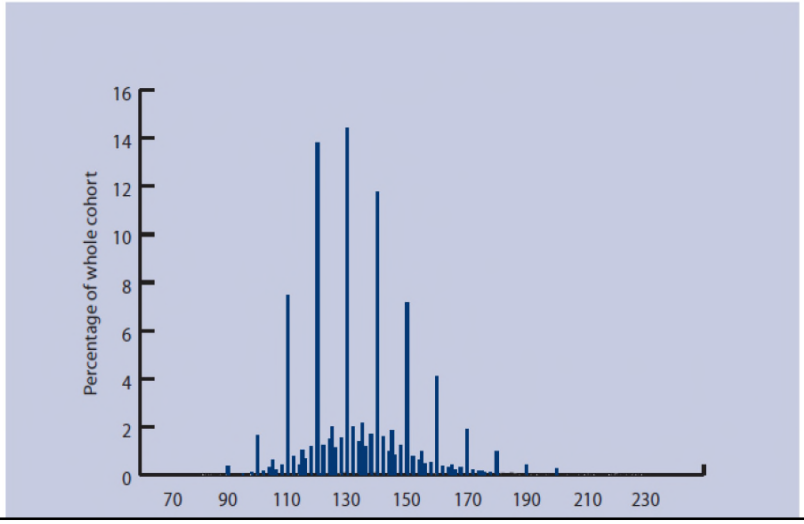


Bland & Altman. BMJ 1996

Show me the data!!! (1)

(data distribution)

Figure 1. Distribution of (a) systolic and (b) diastolic blood pressure in 23 676 patients.

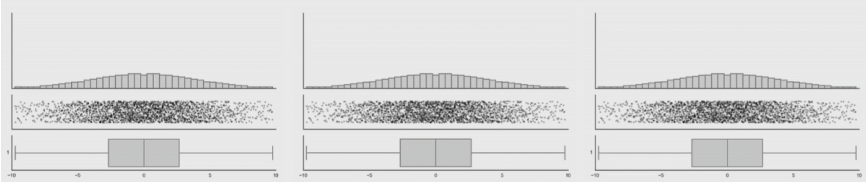


24

Show me the data!!! (1)
(data distribution)

Three ways of showing data distribution:

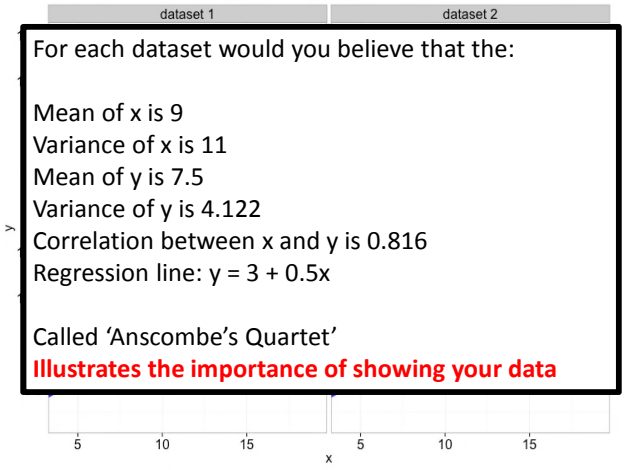
- Histograms
- Dot plots
- Box-plots



https://www.autodeskresearch.com/publications/samestats

25

Show me the data!!! (2)
(correlations)



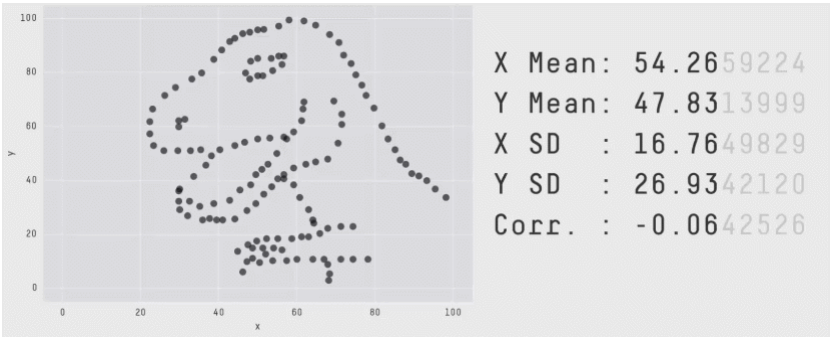
For each dataset would you believe that the:

Mean of x is 9
Variance of x is 11
Mean of y is 7.5
Variance of y is 4.122
Correlation between x and y is 0.816
Regression line: y = 3 + 0.5x

Called 'Anscombe's Quartet'
Illustrates the importance of showing your data

Show me the data!!! (2)
(correlations)

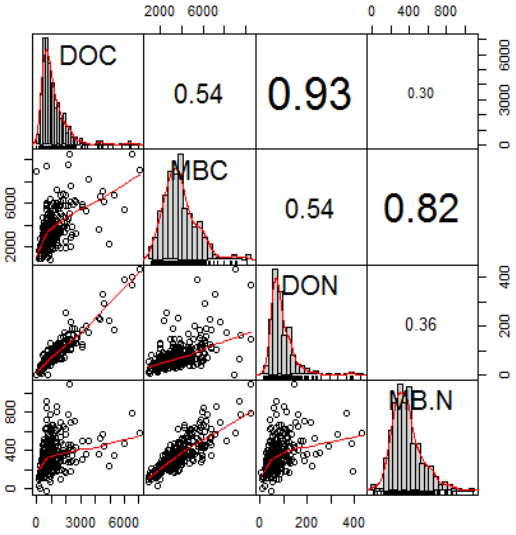
Never trust summary statistics alone!



https://www.autodeskresearch.com/publications/samestats

27

Show me the data!!!
(all your data)



Graphs
(good examples)

- Statisticians like to (ideally) see the raw data!
  - And you should too!!
- Use graphs to describe results, for example
  - Dot plots (more informative with boxplots)
    - good for data distribution and comparing groups
  - Scatter plots
    - good to accompany correlation analyses
  - Survival curves
    - time-to-event analyses
  - Line graphs
    - trends over time
  - Forest plots
    - meta-analyses

Dotplot with a boxplot

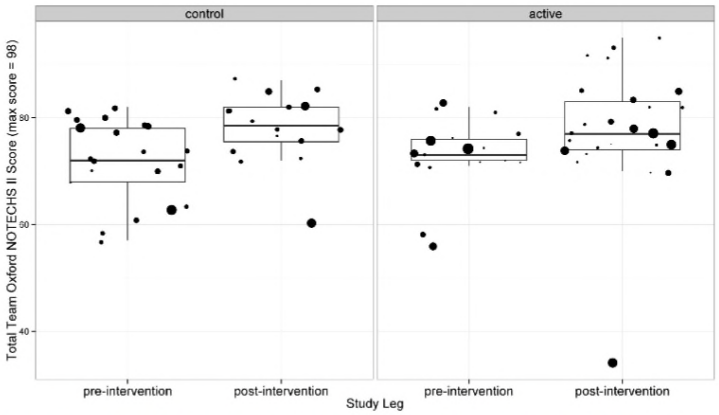
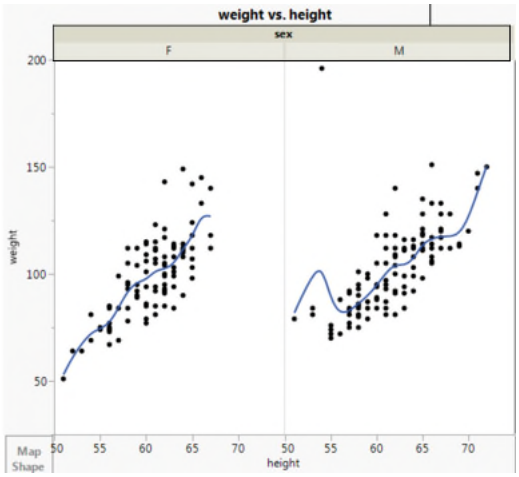


Fig 2. NOTECHS II Results. Each dot is the individual Total Oxford NOTECHS II score for an individual operation, and the size of the dot represents the duration of the operation.

Graphs
(good examples)

- Statisticians like to (ideally) see the raw data!
  - And you should too!!
- Use graphs to describe results, for example
  - Dot plots (more informative with boxplots)
    - good for data distribution and comparing groups
  - Scatter plots
    - good to accompany correlation analyses
  - Survival curves
    - time-to-event analyses
  - Line graphs
    - trends over time
  - Forest plots
    - meta-analyses

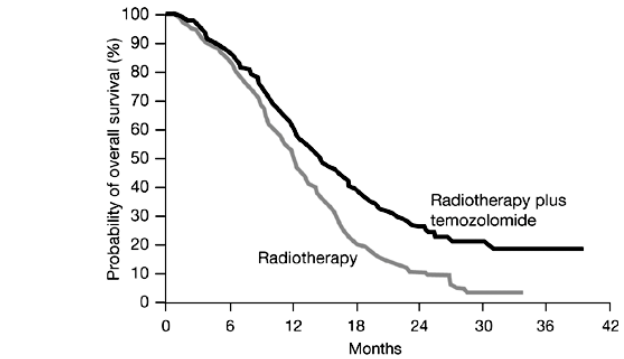
Scatter plot
(stratified by sex)



Graphs
(good examples)

- Statisticians like to (ideally) see the raw data!
  - And you should too!!
- Use graphs to describe results, for example
  - Dot plots (more informative with boxplots)
    - good for data distribution and comparing groups
  - Scatter plots
    - good to accompany correlation analyses
  - Survival curves
    - time-to-event analyses
  - Line graphs
    - trends over time
  - Forest plots
    - meta-analyses

Survival curve



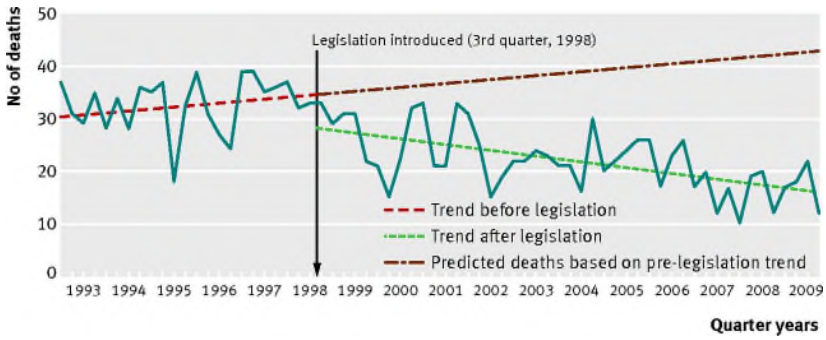
Number at risk							
Radiotherapy	286	240	144	59	23	2	0
Radiotherapy plus temozolomide	287	246	174	109	57	27	4

Graphs
(good examples)

- Statisticians like to (ideally) see the raw data!
  - And you should too!!
- Use graphs to describe results, for example
  - Dot plots (more informative with boxplots)
    - good for data distribution and comparing groups
  - Scatter plots
    - good to accompany correlation analyses
  - Survival curves
    - time-to-event analyses
  - Line graphs
    - trends over time
  - Forest plots
    - meta-analyses

Line plot
(interrupted time series)

Fig 1 Suicide and open verdict deaths involving paracetamol only, in people aged 10 years and over in England and Wales, 1993-2009, and best fit regression lines related to 1998 legislation.



Keith Hawton et al. BMJ 2013;346:bmj.f403

Same data in a table

Table 1| Suicide, open verdict and accidental deaths in England and Wales for people aged 10 years and over, 1993-2009

		All drugs		Paracetamol		Paracetamol compound*	
Year	All causes (suicide, open)	Suicide, open	Suicide, open, accidental	Suicide, open	Suicide, open, accidental	Suicide, open	Suicide, open, accidental
1993	5182	1314	1897	132 (10.0)	181 (9.5)	17 (1.3)	22 (1.2)
1994	5090	1298	2003	126 (9.7)	163 (8.1)	15 (1.2)	20 (1.0)
1995	5127	1390	2140	122 (8.8)	155 (7.2)	24 (1.7)	27 (1.3)
1996	4910	1325	2103	121 (9.1)	158 (7.5)	24 (1.8)	26 (1.2)
1997	4830	1406	2252	149 (10.6)	204 (9.1)	21 (1.5)	27 (1.2)
1998	5347	1432	2246	135 (9.4)	183 (8.1)	16 (1.1)	20 (1.9)
1999	5241	1414	2294	113 (8.0)	150 (6.5)	29 (2.1)	33 (1.4)
2000	5081	1309	2143	90 (6.9)	123 (5.7)	29 (2.2)	31 (1.4)
2001	4904	1280	2176	108 (8.4)	142 (6.5)	18 (1.4)	27 (1.2)
2002	4762	1227	1983	90 (7.3)	124 (6.3)	22 (1.8)	28 (1.4)
2003	4811	1194	1843	91 (7.6)	120 (6.5)	22 (1.8)	28 (1.5)
2004	4883	1246	2008	88 (7.1)	127 (6.3)	31 (2.5)	40 (2.0)
2005	4718	1154	1926	92 (8.0)	126 (6.5)	33 (2.9)	37 (1.9)
2006	4513	979	1821	92 (9.4)	131 (7.2)	35 (3.6)	46 (2.5)
2007	4322	888	1852	66 (7.4)	90 (4.9)	23 (2.6)	36 (1.9)
2008	4603	884	2071	61 (6.9)	106 (5.1)	26 (2.9)	44 (2.1)
2009	4682	898	2185	69 (7.7)	125 (5.7)	26 (2.9)	39 (1.8)

Open=open verdict; accidental=accidental poisoning. Data are no or no (%) of deaths.
*Compounds include paracetamol and codeine, paracetamol and dihydrocodeine, paracetamol and ibuprofen, and paracetamol and aspirin.

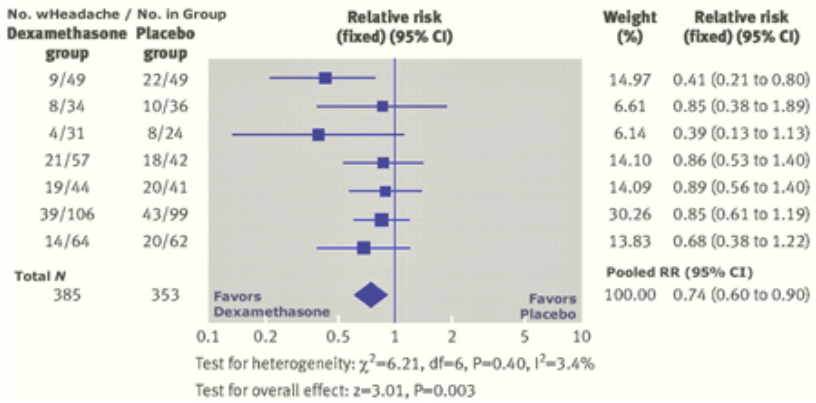
37

Graphs
(good examples)

- Statisticians like to (ideally) see the raw data!
  - And you should too!!
- Use graphs to describe results, for example
  - Dot plots (more informative with boxplots)
    - good for data distribution and comparing groups
  - Scatter plots
    - good to accompany correlation analyses
  - Survival curves
    - time-to-event analyses
  - Line graphs
    - trends over time
  - Forest plots
    - meta-analyses

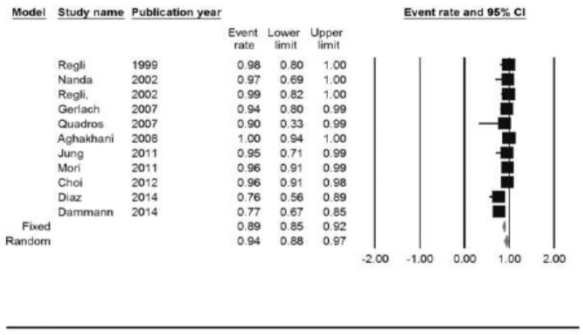
Forest plot

(point estimates, 95% CI, line of no effect, proportionally sized individual study estimates)

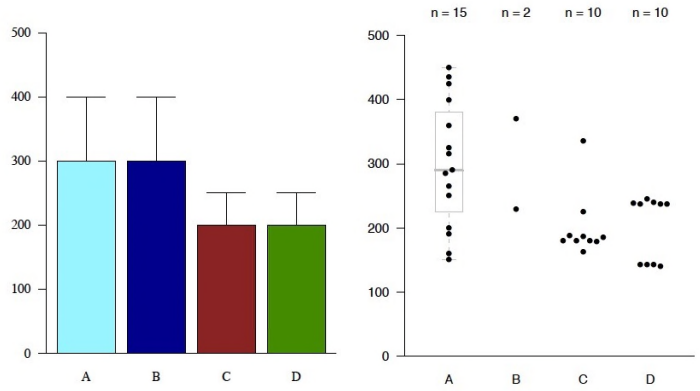


Forest plot

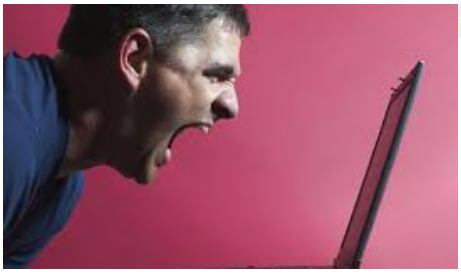
(Not very helpful)

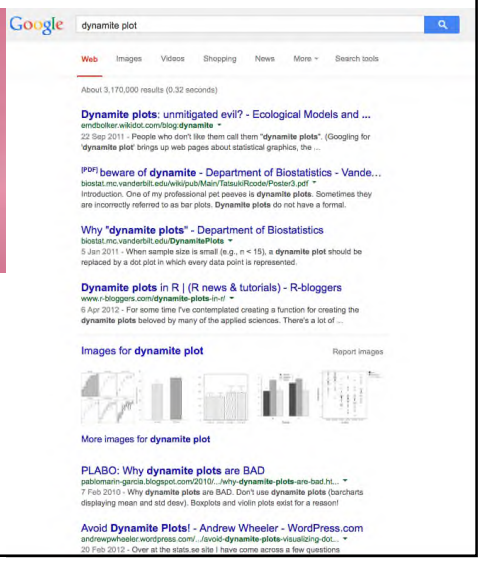


Graph to avoid!
(Dynamite plots)



Dynamite plots

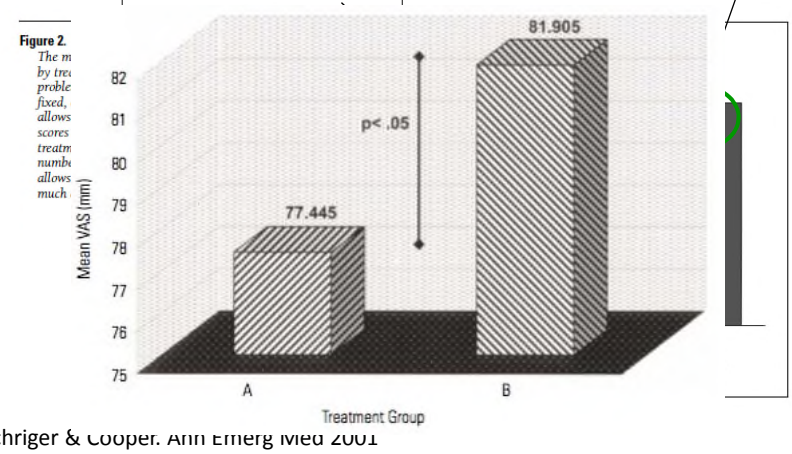




Dynamite plots

Mean=77.445 mm

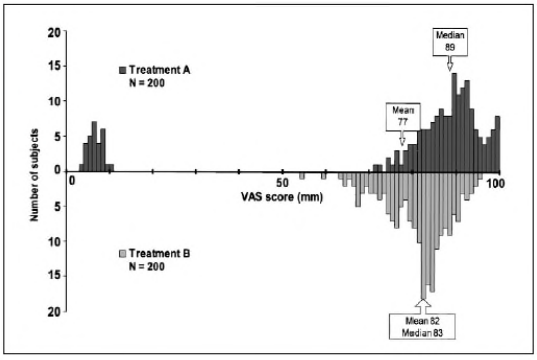
Mean=81.905 mm



Schriger & Cooper. Ann Emerg Med 2001

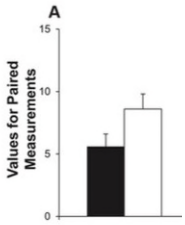
The same data presented differently

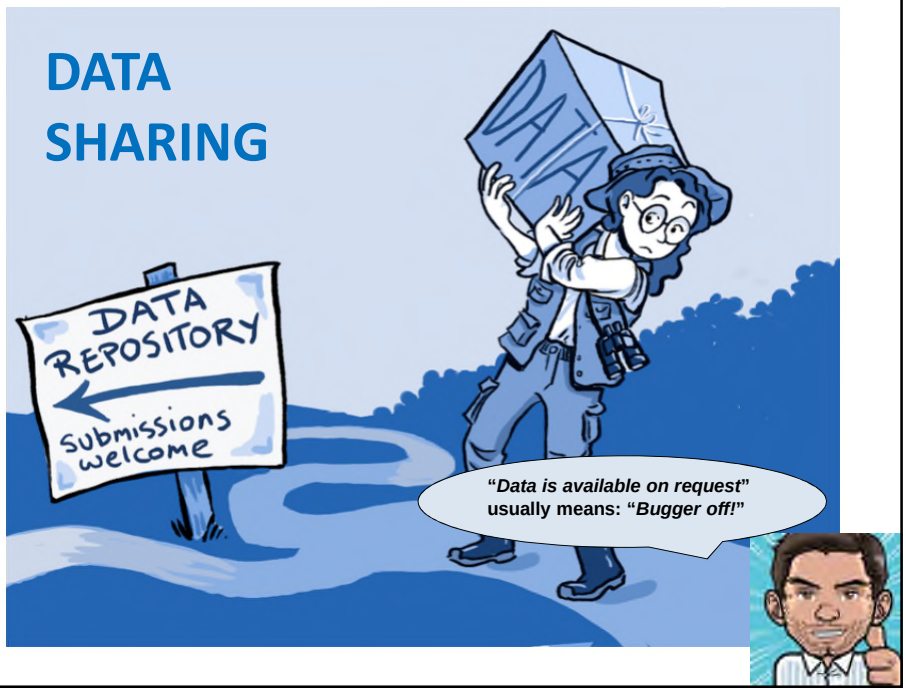
Figure 3.
VAS pain score by treatment group. This graphic depicts 2 adjacent histograms, 1 for each group. The number of subjects that reported each VAS value is shown. The x-axis is the VAS score, which ranges from 0 mm to 100 mm. The DDI of 1.5 cm^{-2} quantitatively demonstrates that Figure 3 provides far more information than Figures 1 and 2.



Schriger & Cooper. Ann Emerg Med 2001

Dynamite plots
(changes over time)





Results: assessing effect

- Don't just report a P-value
- Report the exact P-value
  - very small P-values can be reported as P<0.001
  - but don't report P<0.05, P<0.01
  - AND avoid *, **, ***
  - AND avoid NS or >0.05 to denote not statistically significant
  - AND avoid P=0.000
- Report the measure of effect
  - treatment effect, correlation, differences, odds / hazard ratio
  - report absolute differences (ideally with frequencies)
- Provide a measure of uncertainty around the estimate (e.g. a 95% confidence interval [CI])
  - if different CIs are reported (e.g. 90% or 99%), then make this clear

Results: p-values

- Overreliance on P-values
- P-values provide no indication of the direction or magnitude of the effect (which was the focus of the study)
  - "the effect of the drug was statistically significant"
- Misinterpreted
  - non-statistically significant P-values (i.e. >0.05) are often misleadingly described *suggesting* significance
  - e.g. "trend towards significance", "approaching significance", "narrowly missed significance' and many many more => all nonsense
    - if you come back tomorrow and repeat the analysis, the P-value won't be any closer – they don't move!!!
    - oddly enough, very few P-values 'trend away'…

Still not significant!

(the association between personality and p-values)

Trendy
- a favourable trend (p=0.09)
- a nonsignificant trend toward significance (p=0.1)
- a strong tendency towards statistical significance (p=0.051)
- trend in a significant direction (p=0.09)

Creative
- medium level of significance (p=0.051)
- just shy of significance (p=0.053)
- not quite within the conventional bounds of statistical significance (p=0.12)
- very closely brushed the limit of statistical significance (p=0.051)

Insecure
- may not be significant (p=0.06)
- not quite borderline significance (p>=0.089)
- not very statistically significant (p=0.10)
- somewhat statistically significant (p=0.092)

Believer
- weakly non-significant (p=0.07)
- tantalisingly close to significance (p=0.104)
- barely escaped statistical significance (p=0.07)
- (barely) not statistically significant (p=0.052)

Matthew Hankins
https://mchankins.wordpress.com/ 49

Results: p-values

- Overreliance on P-values
- P-values provide no indication of the direction or magnitude of the effect (which was the focus of the study)
  - "the effect of the drug was statistically significant"
- Misinterpreted
  - non-statistically significant P-values (i.e. >0.05) are often misleadingly described *suggesting* significance
  - e.g. "trend towards significance", "approaching significance", "narrowly missed significance' and many many more => all nonsense
    - if you come back tomorrow and repeat the analysis, the P-value won't be any closer – they don't move!!!
    - oddly enough, very few P-values 'trend away'…

Results: assessing effect

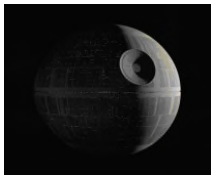
- Don't just report a P-value
- Report the exact P-value
  - very small P-values can be reported as P<0.001
  - but don't report P<0.05, P<0.01
  - AND avoid *, **, ***
  - AND avoid NS or >0.05 to denote not statistically significant
  - AND avoid P=0.000
- Report the measure of effect
  - treatment effect, correlation, differences, odds / hazard ratio
  - report absolute differences (ideally with frequencies)
- Provide a measure of uncertainty around the estimate (e.g. a 95% confidence interval [CI])
  - if different CIs are reported (e.g. 90% or 99%), then make this clear

Death star(s)

TABLE 4 Cross-Product Regression Analyses Evaluating Moderating Effects of Ethnicity on the Relationship Between the SPPA PA Subscale and YSR/CBCL

Subscale	*Global Self-Worth β*	*Physical Appearance β*	*Ethnicity β*	*Physical Appearance × Ethnicity β*	*F (df)*	R^2
YSR						
Social competence	.47*****	−.90***	−.05	.82**	(4, 37) = 2.59****	.22
Internalizing	−.72******	.10	−.18	.16	(4, 43) = 5.15******	.32
CBCL						
Social competence	.59******	−.73*	.02	.61	(4, 41) = 3.93******	.28
Internalizing	−.77******	.11	−.12	.37	(4, 43) = 4.33******	.29

* $p = .10$; ** $p = .08$; *** $p = .07$; **** $p = .05$; ***** $p < .05$; ****** $p < .01$.



Results: assessing effect

- Don't just report a P-value
- Report the exact P-value
  - very small P-values can be reported as P<0.001
  - but don't report P<0.05, P<0.01
  - AND avoid *, **, ***
  - AND avoid NS or >0.05 to denote not statistically significant
  - AND avoid P=0.000
- Report the measure of effect
  - treatment effect, correlation, differences, odds / hazard ratio
  - report absolute differences (ideally with frequencies)
- Provide a measure of uncertainty around the estimate (e.g. a 95% confidence interval [CI])
  - if different CIs are reported (e.g. 90% or 99%), then make this clear

Randomized Clinical Trials
(2 arms)

- No statistical tests on baseline descriptive data!
- The main comparison should be a direct comparison (i.e. a difference) of the two randomized interventions
  - and this should also be reported in the abstract
- Reporting differences (or P-values) within each arm (i.e. baseline and follow-up) provides no information on the difference between treatments
  - but may be reported in addition to reporting the difference between the two arms (the main result)

Justifying 'negative findings'

- **NEVER** report post-hoc sample size calculations
  - they do not justify non-statistical findings
  - they **do** however annoy the statistical reviewer
- Don't report non-statistically significant results as negative findings
  - you just haven't been able to find an effect

Summary

- Ensure methods are clearly written
  - Statisticians don't know every single statistical method or test
- Make sure you have described each analysis you have conducted
  - Avoid bland sentences that could cover a number of analyses
  - Make sure all results have a corresponding description in the methods

Don't be the cause of indigestion!

van der Zee *et al. BM*
DOI 10.1186/s40795-

Nutrition

RESEARCH

Statisti
four pi
and Bra

Tim van der Zee

errors and inconsistencies. Our attention was first drawn to this series of articles when the senior author wrote a blog post about the context in which the articles came to be written [10]. When we followed the references to the articles cited in that blog post we immediately noticed some apparent inconsistencies[1]. We therefore decided to perform a detailed reanalysis of the four articles that seemed to be closely related to each other, to see whether any other problems might emerge. A detailed list of approximately 150 individual inconsistencies and other problems is given in the Appendix; within the text of this article we discuss some of the overarching issues with the four target articles and the implications of what we found.

pen Access

ood

ab based on

Abstract

Background: \
data collected

Method: We calculated whether the means, standard deviations, and test statistics were compatible with the sample size. Test statistics and *p* values were recalculated. We also applied deductive logic to see whether the claims made in each article were compatible with the claims made in the others. We have so far been unable to obtain the data from the authors of the four articles.

57

Further treats (Easter eggs)

- Referring back to study registry/publications
- Referencing methods and justifying decisions
- Making raw data & codes/programs available
- Tell the software (with version) used to analyse the data
- Providing online supplemental material
- Following reporting guidelines (not only ticking boxes)
- Make figures and tables stand alone pieces of info

58

Recommended Reading

Greenwod DC, Freeman JV. How to spot a statistical problem: advice for a non-statistical review. *BMC Medicine* 2015; 13:270.

Lang TA. Twenty statistical errors even you can find in biomedical research articles. *Croat Med J* 2004; 45: 361-370.

Lang TA, Altman DG. Basic Statistical Reporting for Articles Published in Biomedical Journals: The "Statistical Analyses and Methods in the Published Literature" or The SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A (eds). *Science Editors' Handbook,* 2013.

Lee S. Avoiding negative reviewer comments: common statistical errors in anesthesia journals. *Korean Journal of Anesthesiology* 2016; 69: 219-226.

Vickers AV, Sjoberg DD. Guidelines for Reporting of Statistics in *European Urology*. *Eur Urol* 2015; 67: 181-187.

http://annals.org/SS/AuthorInformationStatisticsOnly.aspx [Instructions to authors]

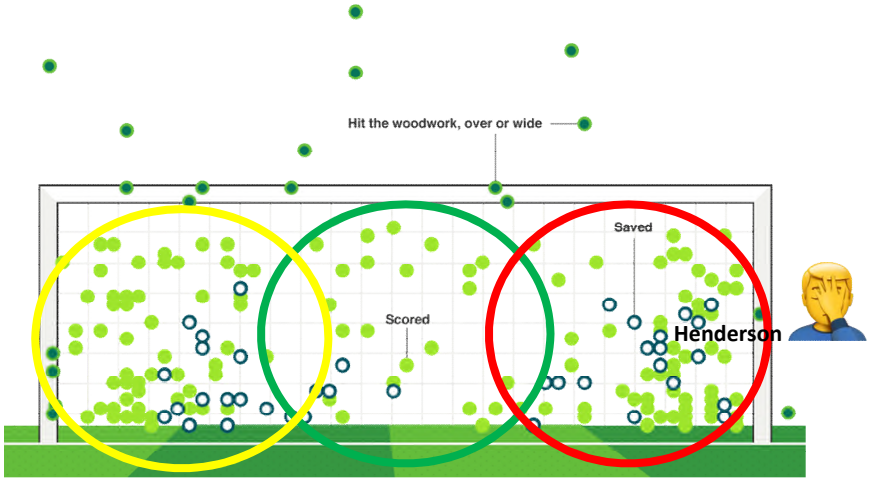
Not to worry?

61

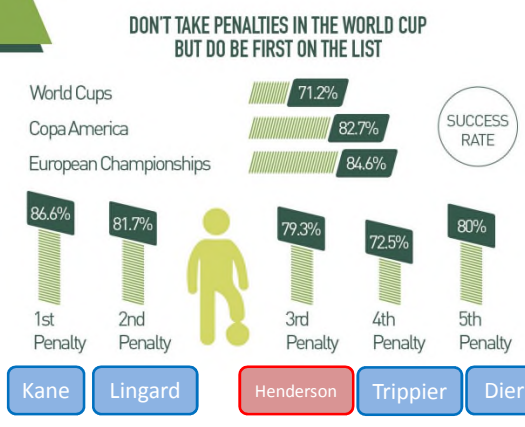
(I mean, really?)

62

Every penalty kick taken in a FIFA World Cup shootout (by the BBC)



63



64

Just another dot in the plot?

Good luck, lads!

Many thanks! @m_schlussel

65