Data Collection and Analysis

Monish Maharaj

Neurosurgery Registrar

Why is this important?

- Knowing how to do your own statistics is a unique skill people need
- Understanding data collection allows you to work harder at automating this process
  - Makes your life easier
  - Makes the quality of your work easier
- Understanding data and methods allows you to actually appraise studies

What to take away from this talk?

- Know what resources you have access to
- Change your way of thinking / develop your critical appraisal skills
- Appreciate the art of the study design phase to make good quality work
- **There is no substitution for good quality data collection and management**

Pre-Collection Considerations

- Ethics
  - Name the variables you are collection but try to keep things vague
  - What are you study aims/hypotheses
- Detail
  - Find the balance between detailed data collection and simple categories/numbers
  - Details can be simplified in data cleaning but not the other way around
  - In the setting of low volume subjects consider simplyfying variables to binary/categorical cut-offs
- Study Power
  - Use an online power calculator to get an idea of how many subjects you'll need to hit significance in your study
  - Normal alpha 0.05, Power 80% (Beta 0.2)
  - https://clincalc.com/Stats/SampleSize.aspx

Dichotomous Endpoint, Two Independent Sample Study

Sample Size	
Group 1	81
Group 2	81
Total	**162**

Study Parameters	
Incidence, group 1	40%
Incidence, group 2	20%
Alpha	0.05
Beta	0.2
Power	0.8

View Power Calculations

What kind of data are we collecting?

- Demographics
- Patient questionnaires
- Qualitative vs quantitative data
- Outcome data for meta-analysing
- Potential for multivariate modelling

How to collect the data

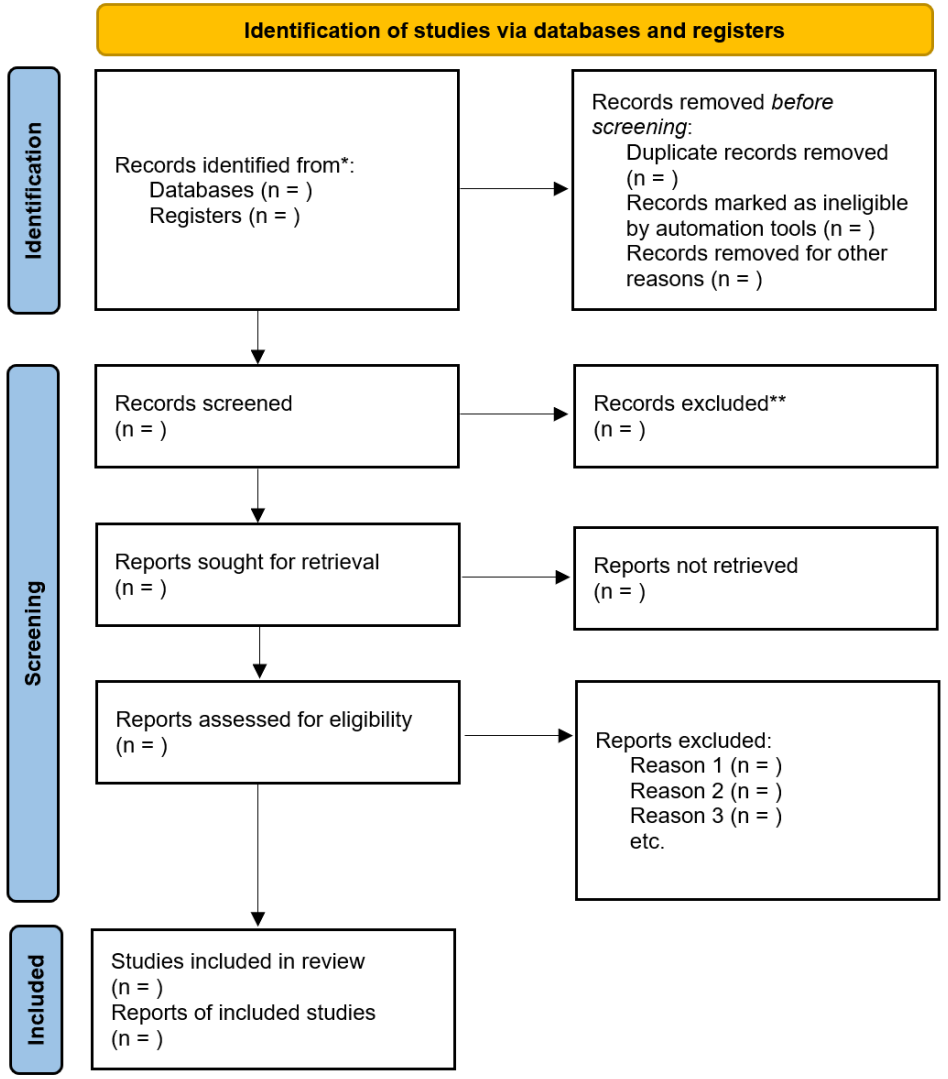
- Teamwork
  - Divide roles and make timelines/goals
- MEDICAL STUDENTS
- Automate as much as possible
  - RSS Feeds for literature reviews
  - Libraries will do lit searches for Research students
  - Surveymonkey premium, push notifications etc
  - Check check check – compliance in studies is difficult to achieve.


Systematic reviews and meta-analyses of RCTs*
Randomized controlled trials
Cohort studies
Case-control studies
Cross-sectional studies, surveys
Case reports, case studies
Mechanistic studies
Editorials, expert opinion
Higher
Quality of evidence
Lower
Lower
Risk of bias
Higher

Reviews and Meta-Analyses

- Search Strategy:
  - http://prisma-statement.org/
  - Quality assessment: PRISMA / MOOSE Checklists
- Design the perfect question
  - You want to have a question that is specific enough that it is meaningful but not so specific that there is only a small yield of studies
  - Rough number of results is 15-40.

Search Strategy



TITLE			
Title	1	Identify the report as a systematic review.	
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	

Meta-analyses – R - https://www.r-project.org/

- *Assumption is that the data is uniform across design/intervention/outcomes*
- Test for heterogeneity
  - Cochran's Q Stat / chi-square
    - Assumes same population = fixed effects model
      - If chi-square is high → data is heterogeneous = might not suitable for meta-analysis*
        - If the studies are still similar design can try a random-effects model.
  - I^2
    - 30% or less is good indicator of homogeneous data
- Make a forest plot
- Assess publication bias
- Consider sub-analyses if data permits

Table 3 Baseline characteristics

Baseline	RR or MD (95% CI)	I^2	P value for heterogeneity	P value overall
Age (years)	–0.78 (–1.64, 0.08)	54	0.004	0.08
BMI	–0.48 (–0.96, 0)	33	0.16	0.05
Males	0.96 (0.90, 1.03)	0	0.86	0.28
NDI	–0.23 (–0.82, 0.35)	0	0.59	0.43
VAS (neck)	–0.01 (–0.21, 0.19)	0	0.67	0.94
VAS (arm)	0.06 (–0.31, 0.43)	0	0.92	0.75
ROM F/E (sup)	0.18 (–0.25, 0.62)	39	0.11	0.41
ROM F/E (inf)	–0.34 (–2.03, 1.35)	91	<0.00001	0.69
ROM	–0.20 (–0.78, 0.37)	76	<0.00001	0.49

RR, relative risk; MD, mean difference; CI, confidence interval; VAS, visual analog scale; BMI, body mass index; NDI, neck disability index; Sup, superior segment; inf, inferior segment; ROM, range of motion.

Table 4 Operational outcomes

Operational outcome	MD (95% CI)	I^2	P value for heterogeneity	P value overall
Operation duration (min)	6.14 (–0.91, 13.19)	87	<0.00001	0.09
Blood loss (mL)	–3.22 (–14.03, 7.60)	88	<0.00001	0.56
Length of stay (days)	–0.05 (–0.17, 0.07)	47	0.11	0.45

MD, mean difference; CI, confidence interval.

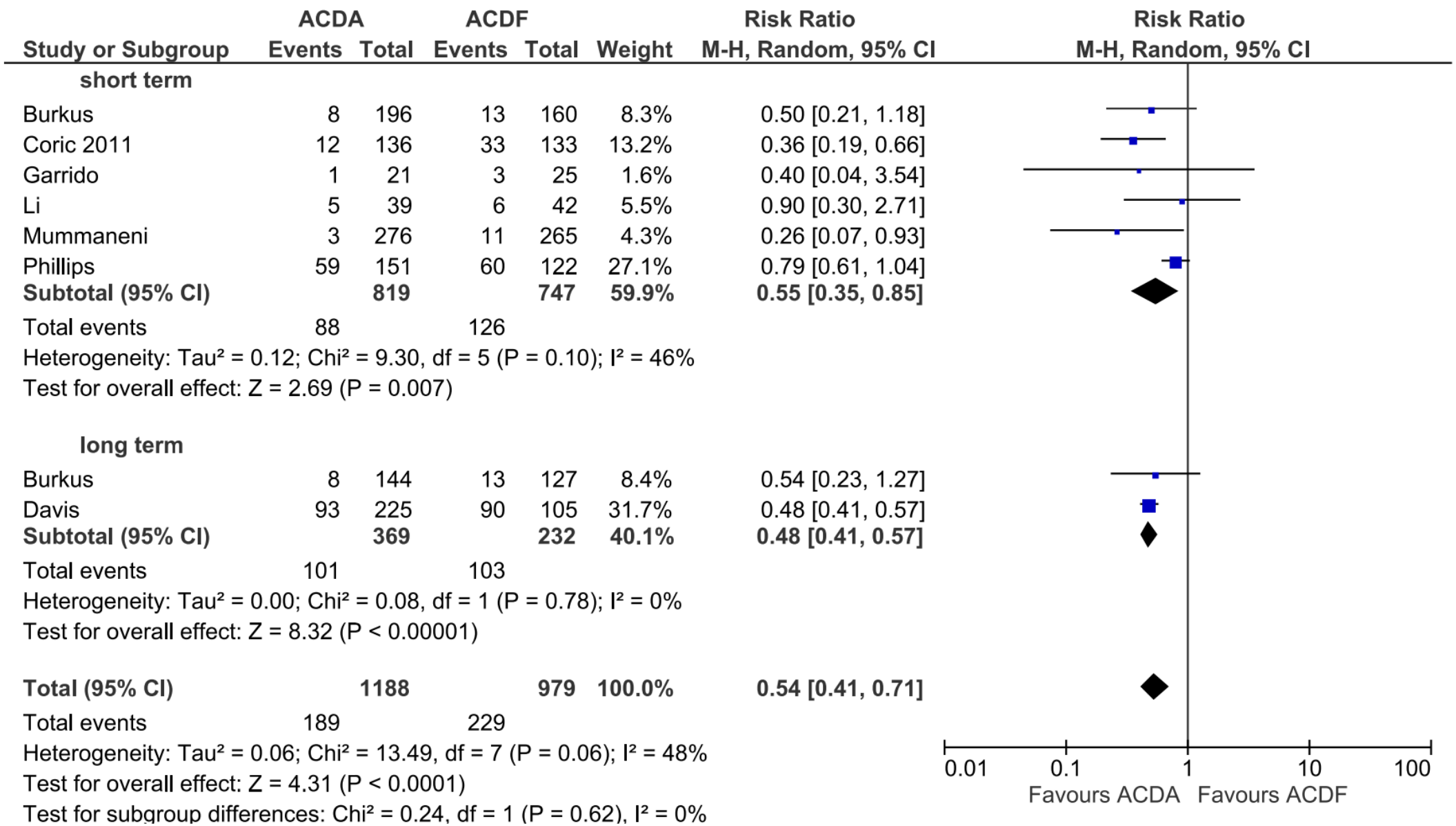
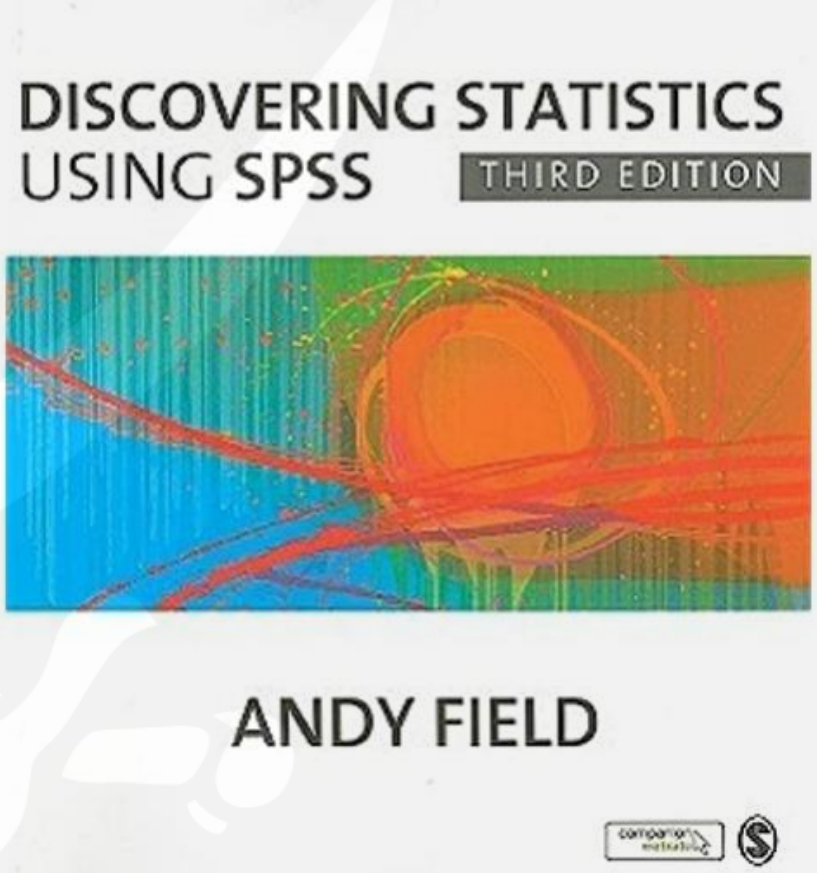


Figure 5 Forest plot of adjacent segment disease for ACDA *vs.* ACDF, stratified into short-term and long-term outcomes. ACDA, anterior cervical disc arthroplasty; ACDF, anterior cervical discectomy and fusion.

Statistical resources

- https://statistics.laerd.com/spss-tutorials

- zLibrary
- SciHub



Principles for statistical analyses

- Clean your data
- Check your demographics for baseline differences
- Examine your outcome variables
- Have a play with the data
- Consider multivariate modelling if sample size permits

Stastistics Software


How do I stats though?

Baseline Characteristics – Table 1

- 1. Check for normality
  - Kolmogorov-Smirnov Test and the Shapiro-Wilk Test
    - Want it to be not significant = normal data
      - If not normally distributed should report descriptives as Median and IQR
      - If not normal other tests need to be non-parametric
- 2. Check for differences between your cohorts

What test do I use? Univariate Cohorts

Variable Type	Parametric	Non-Parametric
Binary	Chi-Square / Fisher's Exact	-
Continuous	t-Test (independent / paired)	Wilcoxon Rank Sum Mann-Whitney U
Categorical	ANOVA	Non-Parametric ANOVA

Table 5
Postoperative midline outcomes

Parameter	Preoperative	Postoperative	% Change	p
ADH	8.3 (3.1)	15.7 (1.9)	90	.00
PDH	4.8 (2.2)	8.5 (1.9)	77	.00
LDA	5.9 (3.9)	12.3 (3.9)	108	.00
LL	41.8 (11.3)	44.1 (13.0)	6	.02

ADH, anterior disc height; PDH, posterior disc height; LDA, local disc angle; LL, lumbar lordosis.

Presenting the data

- Consider your significant figures
- Tables vs Charts
  - If you have skewed data it may be useful to include boxplots for variables
- Simple and clear data presentation is more useful than complex flashy figures that are difficult to understand.

For simple single variable tests consider just putting a sentence in the manuscript rather than tabulating everything

What next?

- If you’ve collected and cleaned data well now is the chance to explore trends you may not have considered
- Multivariate modelling
  - Assumptions: factor collinearity, linear relationships between independent / dependent variables, adequate sample size etc
  - There are tests used in the dialogs specific for your model that provide insight onto model power and accuracy.
  - Model types: Binary logistic regression, One way MANOVA, Count Models (not in SPSS)

Final Tips

- Get the data and have a go
- Do as much as you can, then ask someone who you know understands the stats to check over your work.
- Read other papers methods to see what they've done and learn how to interpret ALL the numbers they publish.
- Review for journals and consider all the lessons from today!