Equipping Healthcare Providers: A Guide to Study Databases, Biostatistical Analyses, and Research Interpretation

René F. Najera, MPH, DrPH, FCPP

1

Disclosures

- No known conflicts of interest
- I am not that kind of doctor; not giving medical advice
- My views/opinions do not necessarily reflect those of my employers, affiliated institutions, friends, family, or the cats

2

Bad Journal Articles - What's the Harm?

- Most retractions are inconsequential
- Joachim Boldt has had 184 retracted articles, the most known to date (Feb 2024)
- Why did he do it?
- False statements and incomplete documentation
- Bad data
- Who are the gatekeepers?
- Did anyone get hurt?

bmj.com
News: Exclusion of suspect data raises question mark over safety of common plasma substitute (BMJ 2013;346:f1132)
News: Researcher didn't get ethical approval for 68 studies, investigators say (BMJ 2013;347:f4013)

Boldt: the great pretender

The withdrawal of almost 90 fraudulent studies by a German anaesthetist is one of the biggest medical research scandals of recent time. **Jacqui Wise** examines what happened and what lessons have been learnt

Fooling it: Joachim Boldt

Joachim Boldt was a prominent German anaesthetist with an international research reputation. He was respected as a leading specialist in intravenous fluid management and was an advocate for the use of colloids, particularly hydroxyethyl starch solutions, to boost blood volume during surgery.

However, a lengthy investigation has led to 88 out of the 102 studies that Boldt has published since 1999 being withdrawn from the medical literature. He has been found guilty of research misconduct, including failure to acquire ethical approval and fabrication of study data, and sacked from his position as professor at Klinikum Ludwigshafen, a large teaching hospital in Ludwigshafen, Germany, where he carried out his research. The retraction of such a large body of work has had far reaching effects on clinical practice, research oversight, and editorial policies.

Deception unmasked

The story starts in December 2009 when the journal *Anesthesia and Analgesia* published a study comparing the effect of two hypoosmolar priming solutions, albumin and hydroxyethyl starch colloidal solution, on markers of postoperative inflammation and organ function.[1] On 18 December, two weeks after publication, a reader sent an email to the journal's editor in chief, Steven Shafer, saying he was puzzled by the research. The email said: "The results are all very consistent, all very much statistically significant with very small standard deviation." He described the results as "extraordinary given the small number of included patients." And "the reported effect on coagulation tests and bleeding is particularly 'magic.'"

Shafer, who is also professor of anaesthesiology at Columbia University, told the *BMJ*: "Boldt was incredibly prolific. He submitted an average one manuscript a month to us and other anaesthesia journals. He tended to publish small studies that you would think were underpowered but often had an interesting finding. All his papers were multiauthored by people with a reputation in anesthesia. He appeared to have a very effective team."

A second email arrived a day later, also questioning the authenticity of the research. Shafer sent an email to Boldt asking him to discuss the

3

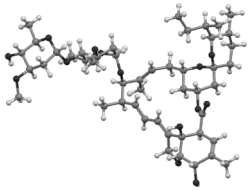
Paper Mills Full of Fake Data

- A BMJ study found 1182 retracted paper mill papers, with the first appearing in 2004 and retractions starting in 2016.
  - A significant portion originated from China, with hospital affiliations common among authors.
  - Most retracted papers were published in journals of the second highest impact factor quartile.
- What's the harm?
  - Undermines trust in medical/scientific expertise
  - Leads legitimate researchers down wasteful rabbit holes
- What's being done?
  - Pre-registries
  - Transparency in publishing, especially data
  - AI checking?

4

Remember Ivermectin?

- Antiparasitic with no known antiviral effects *in vivo*
  - *In vitro:* acts on importin α/β1 nuclear transport proteins, at concentrations that would be highly toxic or deadly to humans
- 2020 research in Australia found a reduction in NCoV-2 concentrations *in vitro*
- Proponents became vocal about it, including politicians
- No clinical evidence of benefit from ivermectin to treat or prevent COVID-19
- 246 retracted papers (as of February 2024, per *RetractionWatch.com*)



5

"The Death of Expertise"

- We think we know more than we do, because we read something online (or hear someone online) telling us what we want to hear.
- Opinion is confused with facts
- Distrust of experts for different reasons
- Spread of mis/disinformation and harmful beliefs/actions
- Experts have made mistakes, but expertise is self-correcting by peers and scientific process…

BUT WHAT IF THE PROCESS IS FULL OF BAD SCIENCE AND FAKE EVIDENCE?

6

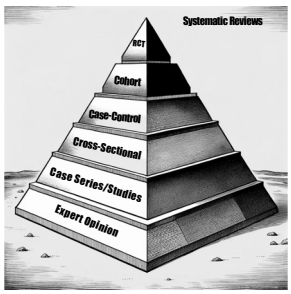
Who CAN You Trust?

- Replicated studies from multiple sources
  - Plausibility test
  - Hierarchy of evidence (more on that soon)
- Large, multi-year studies with heavy oversight
  - Framingham Heart Study
  - Baltimore Longitudinal Study of Aging
  - National Child Development Study
- Studies reviewed by expert panels who then give recommendations
  - US Preventive Services Task Force
  - Advisory Committee on Immunization Practices
- Studies from your peers, where they are open to discussion and in-person presentation of the evidence of their claims

7

Hierarchy of Evidence

8



The Hierarchy of Evidence

9

Expert Opinion

- "Opinion" is the operative word
- Based on facts, but which (or whose) facts?
- On what medium? What length?
  - Television spot
  - Op-ed piece
  - Editorial
  - Lecture
  - Online podcast/video
  - Tiktok

10

Case Reports/Series

- Good jump-off point for rare events/diseases
- Reporting on new treatments and therapies
- There are guidelines to writing them, for standardization
- Should include as many details from each case as possible
- Online tools are available to help write these reports/series. (e.g. https://care-writer.com/)

11

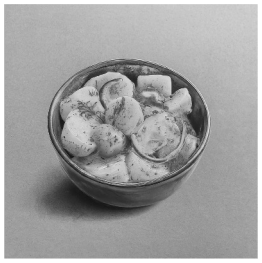
Cross-Sectional Studies (aka Surveys)

- Pros:
  - Quick and easy, cheap
  - Good for creating hypotheses for further studies
  - Can get information from a large group quickly
- Cons:
  - Cannot establish causality (no temporality)
  - Can give hints on associations, but open to misinterpretation
  - Not good for rare conditions

12

Case-Control Studies

- Most often used in outbreak investigations, or for rare exposures/outcomes
  (it's always the potato salad)
- Pros
  - Good for rare diseases
  - Quick-ish, and cheap-ish
  - Can assess risk based on multiple exposures
- Cons
  - Biases: Recall, Selection, Observer, Self-Selection
  - Confounding (dealt with good statistical analysis)
  - Association but not causality
  - One outcome at a time



13

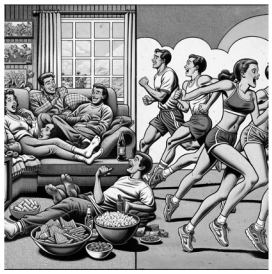
Cohort Studies

- Begin with a group of healthy and unexposed people and follow them through time, measuring exposures and outcomes
- Pros
  - Can establish causality
  - Good for rare exposures, multiple outcomes
- Cons
  - Prone to biases
  - No control over who is exposed and who isn't
  - Biases and confounders

14

Randomized Controlled Trials

- Take a group and randomize into different groups. Ensures group comparability.
- Pros:
  - Comparable groups reduce biases and confounders.
  - Good for causality.
  - Experimental design allows for control of variables
- Cons:
  - Expensive and time-consuming
  - Generalizability can be tricky, depending on who is in and who is out
  - Without blinding, subject to participants' behavior
  - Might miss rare outcomes



15

Systematic Reviews and Meta-Analyses

- Take studies (and their data) and combine to create a "synthesis" of the evidence
- Pros
  - Done well, can use highest-quality evidence
  - Quick way to present all the evidence
- Cons
  - Not all studies are created equal
  - Selection bias in what studies are included/excluded
  - Combining data can lead to errors (more on that later)

16

Existing Databases

17

PubMed

- Combination of MEDLINE, PubMed Central, and Bookshelf
- "More than 36 million citations and abstracts"
- Maintained by NIH
- Best practices:
  - Use MeSH terms (more on that later)
  - Combine terms using Boolean operators (AND, OR)
  - Filter by date, type, etc.
  - Always access the full text… Don't just go with the abstract.
  - There is a user guide

18

Cochrane Library

- Collection of databases, mostly focused on systematic reviews and meta-analyses
- Best practices:
  - Define your question. What do you want answered, exactly?
  - Consider the quality of the evidence. (Remember the hierarchy?)
  - Cochrane Handbook for understanding

19

Excerpta Medica Database (EMBASE)

- Comprehensive biomedical and pharmacological database.
  - Contains over 40 million records from more than 8,500 journals.
  - Strong focus on drug research, pharmacology, toxicology, and medical devices.
  - Includes unique records not found in MEDLINE.
- Content and Coverage
  - Updated daily and weekly, adding around 2 million records annually.
  - Spans from 1974 to present with about 30 million records.
  - MEDLINE supplement with approximately 10 million records.
  - International journal collection from over 90 countries.
- Best Practices
  - Systematic Literature Reviews: Utilize the PICO method for advanced queries.
  - Clinical Trials and Systematic Reviews: Essential for biomedical evidence collection.
  - Special Filters and Syntax: Use Ovid filters and specific syntax for refined searches.
  - Training and Support: Leverage training resources and support for effective database navigation.

20

Web of Science

- Comprehensive research platform and citation indexing service covering multiple disciplines.
- Content: Over 21,000 scholarly journals, 205,000 conference proceedings, 104,000 books, and more
- Uses: Literature searches, citation tracking, research impact analysis, and discovering new research areas
- Best Practices:
  - Utilize Boolean operators and quotation marks for precise searches
  - Analyze publications by author, organization, or funding agency
  - Access bibliometric indices like impact factors for journal selection
  - Stay updated with alerts on relevant topics or funding
  - Engage with training resources for platform familiarity
  - Explore Open Access content for unrestricted articles
  - Leverage citation analysis tools for research impact assessment
  - Citation reports available for tracking citations and analyzing research impact

21

SCOPUS

- More soon…

22

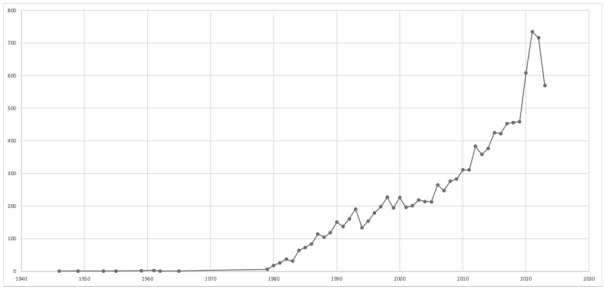
BioMed Central

- More soon…

23

Things to Consider

24



Number of Articles with “Ivermectin” in title in PubMed, by year (1946 to 2023)

25

Journal Impact Factor

- More soon…

26

MeSH Terms

- More soon

27

Biostatistics Crash Course

28

Biostatistics: Use of Statistics in Biological Sciences

- Determine the role of "random chance" in findings
- Uses parametric or non-parametric tests
  - Parametric: Analysis of numerical variables with known distributions
  - Non-parametric: Analysis of categorical or other variables with unknown or non-established distributions
- Distributions
  - Normal distribution: The means (averages) of repeated samples are normally (bell-curve) distributed around a mean
    - Mean, Standard Deviations, Variance, etc.
  - Other distributions: Poisson for counts, Chi-Square for distribution of a characteristic across categories…
- Measures of association (statistics)
  - Odds Ratio
  - Rate Ratio (Risk Ratio)
  - Hazard Ratio
  - Correlation Coefficient

29

The Central Limit Theorem

Throw dice and record the sum of the numbers that come up. Throw again, and record that sum… Over and over.

The average of that sum will be our center. Do this hundreds of times, and the rules will be:

1. 95% of your added up numbers on each throw will be within 1.96 standard deviations of the average
2. 99% of your added up numbers on each throw will be within 3 standard deviations

So, what does this mean? Three cool things:

1. **Predictability**: No matter how random the individual dice throws are, the averages will form a predictable pattern - that bell curve. This means we can start to make predictions about what the average dice throw is likely to be.
2. **The Average of Averages**: The center of this bell curve (the peak) will be pretty close to the average of the entire population (in this case, the average of all possible dice throws if we could do them an infinite number of times).
3. **Less Spread with More Samples**: The more dice you throw each time (increasing your sample size), the narrower your bell curve becomes. This means your averages will be closer together and more consistent.

30

p-value: What is the probability that the results you're seeing are just by chance?

31

95% Confidence Interval: We are 95% confident that the true value of the statistic we calculated is in the range of the interval in the population.

32

A Fictional Example

"In the randomized controlled trial assessing the efficacy of Remdesivir in the treatment of COVID-19, the analysis revealed that the proportion of patients who experienced clinical improvement was approximately 7% higher in the Remdesivir group compared to the placebo group. However, the difference was not statistically significant, with a p-value of 0.328, indicating insufficient evidence to reject the null hypothesis of no difference between the two groups.

Furthermore, the 95% confidence interval for the difference in the proportion of clinical improvement between the groups was calculated to be (-0.067, 0.217). This interval includes zero, which is consistent with the lack of statistically significant evidence for a difference in clinical improvement between the Remdesivir and placebo groups."

33

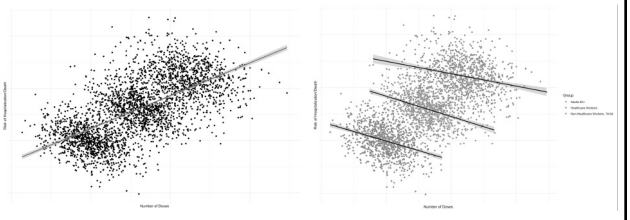
Quick Rules of Thumb

- If the p-value is greater than or equal to 0.05, there is a 5% probability (or higher) of the results being just by chance
- If the 95% confidence interval of a ratio includes 1 (one), then there is at least a 95% probability that the true ratio is 1, meaning the two values are the same
- If the 95% confidence interval of a difference includes 0 (zero), then there is at least a 95% probability that the true difference is 0, meaning there is no difference in values
- The higher the number of units analyzed (participants) in a study, the shorter and more precise the 95% confidence interval is
  - Could go both ways: tighter above or below "no difference"

34

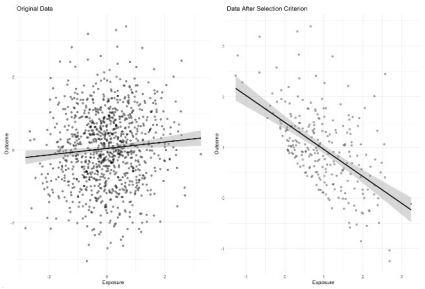
Simpson's and Berkson's Bias

35



Simpson's Bias (Confounding)

36

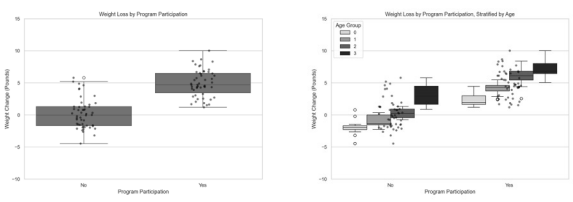


Berkson's Bias (Selection Bias)

37

Confounding

38



Confounding example: Left, unadjusted change in weight. Right, age-adjusted change in weight.

39

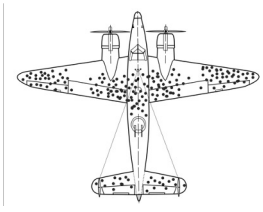
Loss to Follow-Up & Generalizability

40

What Happens When They Leave?

- Example: Clinical trial for weight loss
- People in treatment group stop taking the treatment due to side-effects
- People in treatment group start eating more calories, thinking the medicine will take care of it
- People in placebo group don't see changes and figure out they're in placebo group
- People in placebo group go on a diet and the gym

41



Survivorship Bias

42

Put It All Together

43

How to assess the quality of a paper?

- Does the study design fit the question being asked?
- What biases and confounders could there be, and were they accounted for?
- Does it follow reporting guidelines (CONSORT or STROBE)?
- Is the sample size big enough? Are the groups big enough and comparable to each other?
- Where the proper statistical analyses done?
- Are there other similar studies with similar results?
- If this study is "groundbreaking," pause and look at plausibility.
- Peer review and journal quality are good
- Reported conflicts of interest
- Study seems ethical and well-reasoned
- Data is presented or made available to reviewers

44

Example #1

45

Journal of the Neurological Sciences
Volume 271, Issues 1–2, 15 August 2008, Pages 110-118

Thimerosal exposure in infants and neurodevelopmental disorders: An assessment of computerized medical records in the Vaccine Safety Datalink

Heather A. Young [a], David A. Geier [b], Mark R. Geier [c]

46

Findings

"It was observed that there were **significantly increased rate ratios for the neurodevelopmental disorders of autism**, autism spectrum disorder (ASD), hyperkinetic syndrome of childhood (attention deficit disorder/attention deficit hyperactivity disorder), developmental disorder/learning disorder—not otherwise specified, disturbance of emotions specific to childhood and adolescence, and tics following additional Hg exposure from Thimerosal-containing childhood vaccines. For example, in the birth to 7 month period, **the rate of tics was approximately 3.4 times higher given a 100 microgram increase in Hg exposure in TCVs**. The increased rate ratios ranged from a low of 1.73 (developmental disorder/learning disorder-not otherwise specified) for a 100 μg increase in Hg exposure in the birth to 7 month period to a high of 4.51 (hyperkinetic syndrome of childhood) for a 100 μg increase in Hg exposure in the birth to 13 month period. By contrast, no significantly increased rate ratios for the control disorders of pneumonia, congenital anomalies, and failure to thrive were observed with increasing Hg exposure from Thimerosal-containing vaccines."

Table 3. Rate ratios (95% confidence intervals) for a 100 μg **difference in mercury exposure** for each diagnosis from Thimerosal-containing vaccines administered from birth to 7 months and from birth to 13 months

Diagnosis	100 μg Hg difference	100 μg Hg difference
	Birth to 7 months	Birth to 13 months
	Rate ratio (95% CI)	Rate ratio (95% CI)
Neurodevelopmental disorders		
Autism[1]	2.87 (1.39–6.94)	2.62 (1.35–6.01)
Autism spectrum disorders[2]	2.44 (1.18–5.18)	2.23 (1.19–4.48)
Hyperkinetic syndrome of childhood (ADD/ADHD)[3]	3.51 (2.38–4.57)	4.51 (3.48–5.84)
Developmental disorder/Learning disorder—not otherwise specified[4]	1.73 (1.08–2.75)	1.81 (1.17–2.80)
Disturbance of emotions specific to childhood and adolescence[5]	2.27 (1.26–3.80)	2.91 (1.83–4.68)
Tics[6]	3.39 (1.64–8.79)	4.11 (2.12–7.94)
Control disorders		
Pneumonia	0.88 (0.86–1.11)	0.82 (0.82–1.04)
Congenital anomalies	0.62 (0.34–1.14)	0.57 (0.33–1.00)
Failure to thrive	1.05 (0.74–1.47)	0.82 (0.67–1.27)

ADD/ADHD=Attention Deficit Disorder/Attention Deficit Hyperactivity Disorder.

[1] $p<0.05$.

[2] $p<0.01$.

[3] $p<0.001$.

47

The Problems - Making Up Cases

"Because of concern that the cohorts from 1995-1996 had only 4-6 years of follow-up, frequency distributions of age at diagnosis were examined for all years. This revealed that for some of the disorders a sizable proportion of children were diagnosed after 4.5 years. **Adjustments were made for counts of cases as needed for birth cohorts depending upon the disorder examined to correct for under ascertainment that occurred due to shorter follow-up times**. These adjustments were made for all disorders **including the control disorders** as appropriate based on the age distribution…"

"For example, 37% of autism cases in the study were diagnosed after 5 years old with about 50% diagnosed after 4.5 years old. This is a conservative estimate since it includes the 2 years (1995-1996) that had shorter follow-up times. Examination of the distribution of age of diagnosis by birth year for autism revealed that only about 15% of cases were diagnosed after 5 years of age in the 1995 birth cohort while the 1996 birth cohort had no cases diagnosed after 5 years of age and only 3.5% of cases diagnosed between 4.5 and 5 years of age. **Based on the average age at diagnosis for all cohorts the 1995 count of autism cases was increased by 45 cases with the assumption that all of these would have been added in the 5 year+ age group (bringing this percentage close to the overall average of of 37% diagnosed after 5 years of age.)** The same was done for 1996, but the number of cases was **augmented by 80** because it was assumed that these would be diagnosed in the 4.5 to 5 and 5+ groups essentially bringing the percentage after age 4.5 close to the overall average of 50% diagnosed after 4.5 years of age. The new augmented frequency counts of cases in 1995 and 1996 birth cohorts were then use as new case counts in the analysis."

48

The Problems - A huge conflict of interest

- Mark and David Geier (co-authors) had litigation pending in vaccine court
- They asked the court to pay for the study, which was denied
- They used this study to prove their claims, which would have compensated them if their claims were accepted (they were not)
- From court decision: ***"Thus, Dr. Geier does not appear to have had any formal academic training or degrees or medical faculty experience in epidemiology, and his medical experience has been chiefly in genetics rather than epidemiology. Thus, it is unclear why he was named a "Fellow" of the American College of Epidemiology, and it is doubtful whether he should be considered an expert in epidemiology. I conclude that the petitioners have failed to shoulder their burden of demonstrating that Dr. Geier should be considered an expert in epidemiology."***

49

Example #2

50

A Prospective Study of Postconcussive Outcomes After Return to Play in Australian Football

Michael Makdissi,*† MBBS, PhD, Paul McCrory,† MBBS, PhD,
Antony Ugoni,† BSc(Hons), MSc, David Darby,‡ MBBS, PhD, and Peter Brukner,† MBBS
From the †Centre for Health, Exercise and Sports Medicine, University of Melbourne, Parkville, Victoria, Australia, and ‡CogState Ltd, Melbourne, Victoria, Australia

51

Findings

"A total of **199 concussive injuries were observed in 158 players**. Sixty-one concussive injuries were excluded from analysis because of incomplete data (45 players) or presence of concurrent injury (16 players). Of the 138 concussive injuries assessed, 127 players returned to play without missing a game (92%). The remainder of concussed players returned to play after missing a single game (8%). **Overall, there was no significant decline in disposal rates in concussed players on return to competition.** Furthermore, there were no significant differences in injury rates between concussed and team, position, and game matched controls. In the subset of players who had completed screening cognitive tests, all had returned to their individual baseline performance before being returned to play."

TABLE 1
Summary of Performance Statistics[a]

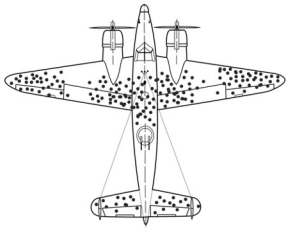
	3 Games	Single Game
$Ratio_{Concussed}$	1.04 (0.99-1.10)	1.08 (0.99-1.11)
$Ratio_{Control}$	1.08 (1.02-1.14)	1.23 (1.09-1.43)
$Ratio_{Concussed}$ to $Ratio_{Control}$	1.04 (0.97-1.12)	1.20 (1.03-1.38)

[a]All results are expressed as ratios of disposal rates per hour match time (95% confidence intervals [CI]); $Ratio_{Concussed}$, ratio of disposal rates pre- and postconcussion; $Ratio_{Control}$, ratio of disposal rates in control players; $Ratio_{Concussed}$ to $Ratio_{Control}$, ratio of concussed to control groups; statistical significance is defined as the exclusion of the value 1.00 within the 95% CI (α level 0.05, 2-tailed test).

52

The Problem - Berkson's Bias or Survivorship Bias?

- Nearly 40% of relevant cases, including high-profile retirees like Dean Kemp and Chad Rintoul, were not included.
- Critics argue the study's methodology and conclusions are flawed due to selective reporting and ignoring long-term concussion effects.
- Associate Professor Alan Pearce's AFL-related health study of retired players faced recruitment and scope limitations imposed by the AFL.
- Initially contracted study was restricted to transcranial magnetic stimulation (TMS) tests only.
- Highlights difficulties in conducting independent research and the need for AFL transparency and cooperation.



53

Conclusion

54

Some Words of Advice

- Trust, but verify… Has it been replicated?
- Do the authors discuss possible biases, confounding, conflicts of interest?
- Are the data from large datasets with oversight, or drawn from a quick survey or convenience sample?
- How generalizable are the results? Was the sample biased in some way?
- Was the measurement flawed in some way?
- Most of all… Is it plausible?

55

Questions/Comments/Concerns?

56