Putting it all together: Meta-analyses and systematic overviews

Steve Simon
P.Mean Consulting
www.pmean.com

2. Why do I offer this webinar for free?

I offer free statistics webinars partly for fun and partly to build up goodwill for my consulting business,

- www.pmean.com/consult.html.

I also provide a free newsletter about Statistics, The Monthly Mean. To sign up for the newsletter, go to

- www.pmean.com/news

3. Abstract

This class helps you assess the quality of a systematic overview or meta-analysis. In this class you will learn how to: recognize sources of heterogeneity in meta-analysis; identify and avoid problems with publication bias; and explain the ethical concerns with failure to publish and with duplicate publication.

This material is derived mainly from Chapter 5 of **Statistical Evidence in Medical Trials**.

4. Outline

1. Pop quiz
2. Introduction and motivating example
3. Were apples combined with oranges?
4. Were some apples left on the tree?
5. Were all of the apples rotten?
6. Repeat of pop quiz

Note: there are also issues involving practical significance (did the pile of apples amount to more than just a hill of beans?) but we will not have time to discuss that issue today.

5. Pop quiz #1

A funnel plot is useful for assessing

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don't know

6. Pop quiz #2

Cochran's Q and I^2 are measures of

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don't know

7. Pop quiz #3

The Jadad score is a measure of

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don't know

8. Introduction

- When there are multiple research studies evaluating a new intervention, you need to find a way to assess the cumulative evidence of these studies. You can do this informally, but medical researchers now use a formal process, known as meta-analysis. Meta-analysis, involves the quantitative pooling of data from two or more studies.

9. Introduction

- More recently, another term, systematic overview, has come into favor. A systematic overview involves the careful review and identification of all research studies associated with a topic, but it may or may not end up pooling the results of these studies. So meta-analysis represents a subset of all the systematic overviews.

10. Motivating example

- In 1992, the British Medical Journal published a controversial meta-analysis. This study (Carlsen 1992) reviewed 61 papers published from 1938 and 1991 and showed that there was a significant decrease in sperm count and in seminal volume over this period of time. For example, a linear regression model on the pooled data provided an estimated average count of 113 million per ml in 1940 and 66 million per ml in 1990.

11. Motivating example

- Several researchers (Olsen 1995; Fisch 1996) noted heterogeneity in this meta-analysis, a mixing of apples and oranges. Studies before 1970 were dominated by studies in the United States and particularly studies in New York. Studies after 1970 included many other locations including third world countries. Thus the early studies were US apples. The later studies were international oranges. There was also substantial variation in collection methods, especially in the extent to which the subjects adhered to a minimum abstinence period.

12. Motivating example

- The original meta-analysis and the criticisms of it highlight both the greatest weakness and the greatest strength of meta-analysis. Meta-analysis is the quantitative pooling of data from studies with sometimes small and sometimes large disparities. Think of it as a multicenter trial where each center gets to use its own protocol and where some of the centers are left out.

13. Motivating example

- On the other hand, a meta-analysis lays all the cards on the table. Sitting out in the open are all the methods for selecting studies, abstracting information, and combining the findings. Meta-analysis allows objective criticism of these overt methods and even allows replication of the research.

14. Motivating example

- Contrast this to an invited editorial or commentary that provides a subjective summary of a research area. Even when the subjective summary is done well, you cannot effectively replicate the findings. Since a subjective review is a black box, the only way, it seems, to repudiate a subjective summary is to attack the messenger.

15. Were apples combined with oranges?

- Meta-analyses should not have too broad an inclusion criteria. Including too broad a range of studies can lead to problems with heterogeneity (mixing apples and oranges).

16. First example of heterogeneity

- In a meta-analysis looking at antiretroviral combination therapy (Jordan 2002), both short-term and long-term outcomes were examined. A plot of duration of trial versus the log odds ratio showed that shorter duration trials of zidovudine had substantial evidence of effect (odds ratios much smaller than 1) but that the largest duration studies had little or no evidence of effect (odds ratios very close to 1).

17. Second example of heterogeneity

- Example: In a meta-analysis looking at dust mite control measures to help asthmatic patients (Gotzsche 1998), the studies exhibited heterogeneity across several factors.

18. Second example of heterogeneity

- Type of intervention:
  - six examined chemical interventions,
  - thirteen examined physical interventions,
  - four examined a combination approach.
- Research design:
  - nine of these trials were crossovers,
  - fourteen had a parallel control group.
- Blinding
  - seven studies had no blinding,
  - three studies had partial blinding,
  - thirteen studies used a double blind.

19. Second example of heterogeneity

- Age of patients
  - nine studies the average age of the patients was only 9 or 10 years,
  - nine other studies had an average age of 30 or more,
  - five studies had a greater mix of ages.
- Duration
  - eleven studies lasted eight weeks or less,
  - five studies lasted a full year,
  - seven studies had an intermediate duration

20. Possible sources of heterogeneity

- This list is adapted from Horwitz 1987
  - Inclusion/exclusion criteria
  - Geographical limitations
  - Independent versus matched controls
  - Dose/timing of drug administration
  - Length of follow-up
  - Drop-out rates
  - Allowable physician discretion
  - Outcome measure

21. Measuring heterogeneity

- Cochran's Q: A value close to the number of studies is good, but a value much larger is bad.
- I^2: ranges between 0% and 100%, larger values indicating greater heterogeneity.
- Many researchers recommend a qualitiative assessment of heterogeneity.

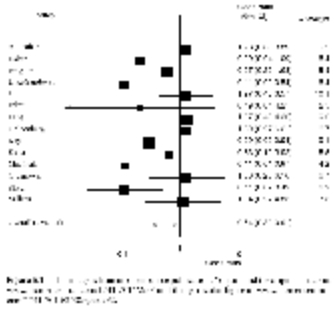
22. Forest plot

- The forest plot provides a graphical summary of the studies. This plot can be used to evaluate heterogeneity.
  - Location of square represents the point estimate,
  - Size of square represents weight associated with that estimate, and
  - Lines drawn to upper and lower confidence limits.

23. Forest plot

- Look for marked departures from a normal random scatter:
  - Most studies cluster together, but one or two outlying studies (but okay if outlying studies have small sample sizes).
  - Bimodal patterns (e.g., half the studies show a strong effect, half show little or no effect).

24. Forest plot example



25. L'Abbe plot

- This plot shows the degree of heterogeneity in the placebo response rate.
  - Horizontal axis: response rate in placebo group.
  - Vertical axis: response rate in treatment group.
  - Diameter of circles are proportional to the sample size of the individual studies.

26. L'Abbe plot

- A diagonal line separates the plot into two regions, the region lower and to the right represents studies where the percentage is higher in the placebo group. The region higher and to the left represents studies where the percentage is higher in the treatment group.

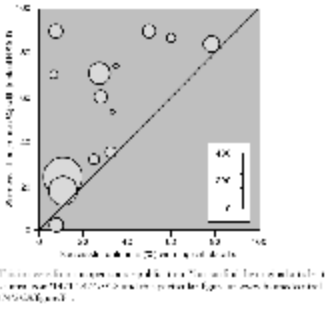
27. L'Abbe plot

- Studies with a high placebo response rate (those on the right half of the graph), may represent situations where the patients were not very ill to begin with, because even a placebo cures most of them.

28. L'Abbe plot

- Examine
  - Variations in the placebo response rate.
  - Whether the superiority of the treatment group is uniform across low and high placebo response rates.

29. L'Abbe plot example



30. Handling heterogeneity

- There are several common approaches for coping with heterogeneity
  - Strict inclusion/exclusion criteria
  - Sensitivity/subgroup analysis
  - Meta-regression
  - "Just say no"

31. Example of strict inclusion/exclusion criteria

- A meta-analysis of topical NSAIDs for musculoskelatal pain (Mason 2004) identified 60 target papers, but for 12 of the papers, there was no data that could be extracted for a meta-analysis. An additional 23 studies were removed based on the following exclusion criteria:
  - no studies for mouth or eye diseases;
  - no studies where fewer than 10 patients were randomized to the treatment;
  - no studies where treatment occurred less frequently than daily;
  - no observational studies; and
  - no unblinded studies.

32. First example of strict sensitivity/subgroup analysis

- In a study of extra corporeal shock wave therapy for plantar heel pain (Thomson 2005), six studies met the researchers inclusion criteria, but one study did not report a standard deviation for the outcome measure. The authors were forced to estimate what the standard deviation should be for this study. As a quality check, they also ran a meta-analysis without this study and found that a modest effect in favor of the therapy was no longer statistically significant.

33. Second example of strict sensitivity/subgroup analysis

- In a study of topical NSAIDs for osteoarthritis and tendinitis (Mason 2004), researchers identified 25 trials relating to efficacy or harm, including 14 placebo-controlled trials. These studies varied substantially in
  - quality scores,
  - number of patients studied,
  - type of outcome measure (physician determined versus self report) and
  - condition being treated (osteoarthritis versus other musculoskeletal conditions).
- But when the results were tabulated separately for low and high quality scores, small and large studies, etc., there were no statistically significant differences.

34. Meta-regression

- You can use meta-regression to try to adjust for heterogeneity in a metaanalysis. In meta-regression, each study becomes a data point, and various study characteristics, such as the severity of illness at baseline, the dose of the medication being given, etc. become independent variables. This is an approach that would work very similarly to the adjustment for covariates in a regression model. The result, meta-regression, is an area of active research and looks to be a promising way to handle heterogeneity in a more rigorous fashion.

35. Example of meta-regression

- In a study of diagnostic tests for endometrial hyperplasia (Clark 2004), researchers identified 27 studies using miniature endometrial biopsy devices or ultrasonography. In some of the studies, verification of the diagnosis was delayed by more than 24 hours. Although the ability to discriminate between diseased and healthy patients was present in most studies, the discriminatory power, as measured by the diagnostic odds ratio was four times weaker among studies with delayed verification than studies with no delay.

36. “Just say no”

- If the degree of heterogeneity is too extreme, you should just say no and refuse to run a meta-analysis. You can still discuss the studies in a qualitative fashion, but do not try to compute an overall estimate of effect because that estimate would be meaningless.

37. Example of “Just say no”

- In a systematic review of beta-2 agonists for treating chronic obstructive pulmonary disease (Husereau 2004), researchers identified 12 studies. But the authors could not pool the results because they
  - “found that even commonly measured outcomes, such as FEV1, could not be combined by meta-analysis because of differences in how they were reported. For example, in the six trials comparing salmeterol with placebo, FEV1 was reported as a mean change in percent predicted, a mean change overall, a mean difference between trial arms, no difference (without data), baseline and overall FEV1 (after 24 hrs without medication) and as an 0 to 12 hour area-under-the-curve (FEV1-AUC) function. We were not successful in obtaining more data from study authors. We also had concerns about the meta-analysis of data from trials of parallel and crossover design and differences in spirometry protocols including allowable medications. Therefore, we decided on a best evidence synthesis approach instead.”

38. Were some apples left on the tree?

- Publication bias: the tendency on the parts of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings. There is solid empirical evidence (e.g., Dickersin 1990) that negative studies are less likely to be published.

39. Ethical concerns with failure to publish

- Researchers who fail to publish their research, however, are behaving unethically (Chalmers 1990). These research studies almost always use human volunteers. These volunteers might be participating because they need the money or perhaps they are curious about the scientific process. But many of them volunteer because they want to help others who have the same disease or condition. These volunteers submit themselves willingly to some level of inconvenience, and possibly additional pain and risk. If you ask these volunteers to make this sacrifice, but you do not publish, you have abused their good will.

40. Should unpublished studies be included?

- The inclusion of unpublished studies, however, is controversial. At least one reference (Cook 1993), has argued that unpublished studies have failed to meet a basic quality screen, the peer review process. Including studies that have not been peer reviewed will lower the overall quality of the meta-analysis. This opinion, however, is in the minority, and most experts in meta-analysis suggest that you include unpublished studies if you can find them. Failure to include unpublished studies can lead to serious bias.

41. Duplicate publication

- Duplicate publication is the flip side of the same coin. The data from some studies may appear twice (or even three times) in the peer-reviewed literature, without appropriate attribution. If you double count these studies accidentally, you will produce a biased result because duplicate publications are more likely to be positive.

42. Ethical concerns with duplicate publication

- Duplicate publication raises serious ethical issues:
  - Violation of copyright
  - Padding of resumes
  - Abuse of volunteer services of referees/editors
  - Taking page space away from other deserving publications.
- There are reasonable justifications for duplicate publication, such as translating a publication into English to insure a wider dissemination of the research findings. These exceptions, however, would always have an obvious citation of the original source.

43. Example of duplicate publication

- In 84 studies of the effect of ondansetron on postoperative emesis, 14 (17%) were second or even third time publications of the same data-set (Tramer 1997). The duplicate studies had much larger effects and adding the duplicates to the originals produced an overestimation of treatment efficacy of 23%. Tracking down the duplicate publications was quite difficult. More than 90% of the duplicate publications did not crossreference the other studies. Four pairs of identical trials were published by completely different authors without any common authorship.

44. Don't rely exclusively on Medline

- While a Medline search is a very effective way to identify published research, it should not be the only source of publications for a meta-analysis. There are many important journals which are not included in Medline. It is hard to get an accurate count of how many journals do NOT appear in Medline, but the numbers appear to be substantial. You might suspect that journals indexed by Medline are more prestigious and more likely to publish positive findings than other journals, but I am unaware of any data to substantiate this. Still, a search that included only Medline articles would be considered grossly inadequate in most situations.

45. Don't rely English-language only publications

- Some meta-analyses restrict their attention to English language publications only. While this may seem like a convenience, in some situations, researchers might tend to publish in an English language journal for those trials which are positive, and publish in a (presumably less prestigious) native language journal for those trials which are negative (Gregoire 1995). Restrictions to English language only publications is especially troublesome for complementary and alternative medicine, since so much of this research appears in non-English language journals.

46. Using a funnel plot to detect publication bias

- The most common approach to evaluate publication bias is to use a funnel plot. The funnel plot displays
  - the results of the individual studies (e.g. the log odds ratio) on the horizontal axis,
  - the size of the study (or sometimes the standard error of the study) on the vertical axis.
- Often a reference line is drawn at the value that represents no effect.

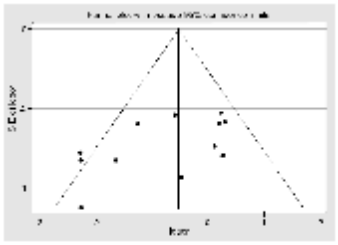
47. Using a funnel plot to detect publication bias

- The rationale behind this plot
  - big studies get published no matter what the result
  - smaller studies are subject to publication bias
- If there is no publication bias, then the funnel plot should show symmetry for both small sample sizes and large sample sizes, though you should expect to see less variation as the sample size increases. This leads to a funnel shape.

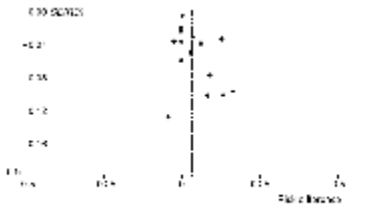
48. Example of a funnel plot

- The rationale behind this plot
  - big studies get published no matter what
  - smaller studies are subject to publication bias
- If there is no publication bias, then the funnel plot should show symmetry for both small sample sizes and large sample sizes, though you should expect to see less variation as the sample size increases. This leads to a funnel shape.
- Although funnel plots are commonly used, there is some suggestion that they are not effective.

49. Funnel plot example showing symmetry



50. Funnel plot example showing possible publication bias



51. How to avoid or minimize problems with publication bias

1. Use several bibliographic databases, not just Medline.
2. Search through registries of clinical trials.
3. Hand search through specialized journals
4. Examine bibliographies of articles found on first pass through.
5. Examine “gray literature” (presentations, dissertations, etc.)
6. Send out letter to prominent leaders in the area asking for help.

52. Were all of the apples rotten?

- A homogeneous set of studies is not good if the studies are homogeneously bad. The quality of a meta-analysis is constrained by the quality of articles that are used in a meta-analysis. Meta-analysis cannot correct or compensate for methodologically flawed studies. In fact, meta-analysis may reinforce or amplify the flaws of the original studies.

53. Meta-analysis of observational studies

- The use of meta-analysis on observational studies is very controversial. Many meta-analyses start off with randomization as part of the inclusion criteria, but others allow nonrandomized studies (observational studies) to participate as well. For some areas, observational studies may be the only studies available.

54. Meta-analysis of observational studies

- Is it acceptable to include observational studies in a meta-analysis? A collaborative effort known as MOOSE (meta-analysis of observational studies in epidemiology) provided reporting guidelines to improve the quality of these types of overviews (Stroup 2000). Some experts have argued, however, against including observational studies in a meta-analysis.

55. Meta-analysis of observational studies

- The theory behind these criticism notes first that observational studies have systematic biases, and there is no easy way to correct for systematic biases in a meta-analysis. Uncertainties associated with small sample sizes cause random variations in either direction, and these cancel out when you combine multiple studies. But uncertainties or biases associated with weak research designs tend to point in the same direction, and these biases are preserved in the meta-analysis. So the relative importance of bias may be moderate in a single small observational study, but it rises to a position of great prominence in a meta-analysis.

56. Meta-analysis of studies with uniformly small sample sizes

- You should be very careful in the assessment of meta-analyses where all of the trials have small sample sizes. The effect of publication bias can be far more pronounced here than in situations where some medium and large size trials are included. In addition, smaller studies tend to have greater problems with the methods of randomizing and blinding patients (Kjaergard 2001).

57. Quality problems with Chinese studies of alternative medicine

- Research published in Chinese journals have shown a substantial deficits in quality that should make you cautious about any meta-analysis using these studies. For example, a review of Chinese medicinal herbs in the treatment of hepatitis B (Liu 2002) showed inadequate documentation of the randomization method and failure of most studies to conceal the allocation list. Further, a small fraction of these studies showed a degree of imbalance between the treatment and control that was well beyond what you would expect by chance.

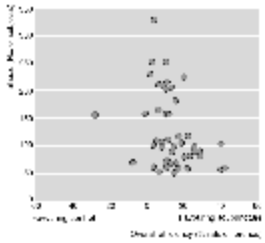
58. Quality problems with Chinese studies of alternative medicine

- A review of 2,938 publications in Chinese journals (Tang 1999) also noted many problems:
  - "Although methodological quality has been improving over the years, many problems remain. The method of randomisation was often inappropriately described. Blinding was used in only 15% of trials. Only a few studies had sample sizes of 300 subjects or more. Many trials used as a control another Chinese medicine treatment whose effectiveness had often not been evaluated by randomised controlled trials. Most trials focused on short term or intermediate rather than long term outcomes. Most trials did not report data on compliance and completeness of follow up. Effectiveness was rarely quantitatively expressed and reported. Intention to treat analysis was never mentioned."

59. Publication bias in Chinese studies of alternative medicine

- A review article on acupuncture (Vickers 1998) evaluated articles published in various countries. In China, 100% of the acupuncture studies showed a positive result. In areas other than acupuncture, the results were similar. In Chinese journals, 99% of the nonacupuncture studies were positive. To form a basis of comparison, only 75% of the studies published in England were positive. Another revealing statistic was that Chinese journals never published a finding to show that the new therapy was less effective than the control group. There were similar problems with publications from Japan, Taiwan, and Russia.

60. Publication bias in Chinese studies of alternative medicine



61. Quality scores

- Any effort to control for the quality of research studies needs a way, either quantitative or subjective, to assess that quality. The most common quantitiative measure of publication quality used in meta-analysis is the Jadad score (Jadad 1996). The Jadad score rates three things:
  1. Randomization (2 points if method is described well and is appropriate, 1 point if method has no description);
  2. Blinding (2 points if double blind with good description, 1 point if study is blinded but with no description);
  3. Withdawals/dropouts (1 point for description of the number of withdrawals and reasons).

62. How quality scores are used in meta-analysis

- You can use quality scores in several different ways:
  1. Use the quality score as one of your inclusion or exclusion criteria.
  2. Perform a subgroup analysis on the studies with quality scores above/below a certain threshold.
  3. Give greater weight to those studies with higher quality.
  4. Use quality scores in a meta-regression model.

63. Example of the use of quality scores

- In a meta-analysis of topical NSAIDs for chronic musculoskeletal pain (Mason 2004), all studies were rated on the Jadad quality scale. To be included in the meta-analysis, the study had to score at least two points on the Jadad scale. Later in the paper, studies scoring only two points were compared with studies scoring three or more points on the Jadad scale. When the low scoring studies were excluded, the pooled estimate of effect did not show a sizable change.

64. Repeat of pop quiz #1

A funnel plot is useful for assessing

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don't know

65. Repeat of pop quiz #2

Cochran's Q and I^2 are measures of

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don't know

66. Repeat of pop quiz #3

The Jadad score is a measure of

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don't know