# MEDB 5502, Module09, Meta-analysis

# Topics to be covered

- What you will learn
  - What is meta-analysis
  - Heterogeneity
  - Study quality
  - Publication bias
  - Interpretability of results
  - Numeric summaries
  - Strengths and weaknesses
  -

# Meta-analysis

- Quantitative pooling of results from multiple studies
  - Multi-center study
    - Each center has a different protocol
    - Some centers do not share results
- Contrast to systematic overview
  - Careful review of multiple studies
  - May or may not include quantitative pooling
- Contrast to scoping review
  - “Researchers may conduct scoping reviews instead of systematic reviews where the purpose of the review is to identify knowledge gaps, scope a body of literature, clarify concepts or to investigate research conduct.” Munn 2018

Speaker notes

Meta-analysis is the quantitative pooling of results from multiple independently published research studies. I joke about how meta-analysis is a multi-center research study but with a couple of qualifications. First, each center gets to use a different protocol. Second, some centers do not share their results with you. This hints at a couple of important issues I will talk about in detail: heterogeneity and publication bias. When I describe it as a chaotic multi-center trial, it sounds terrible. Well, maybe, but we have still learned a lot from meta-analytic studies in a broad range of scientific and medical areas.

A more commonly used term is “systematic overview” which is a superset of meta-analysis. A systematic overview is systematic meaning that it uses a careful and transparently documented approach to identify all research studies associated with a particular issue. It may or may not include a quantitative analysis. Thus all meta-analyses are systematic reviews but not all systematic reviews are meta-analyses.

There’s a new term, scoping review which I am less familiar with. Here’s a nice quote from a paper by Zachary Munn et al in BMC Medical Research Methodology. It sounds almost like a scoping review is a systematic overview with the pre-specified intent to stop before any serious meta-analytic intents.

I am going to focus only on meta-analysis because I am a statistician and this is a statistics course, so we all love anything quantitative. But never forget that there is more to research than just its quantitative aspects.

# Case study: Declining sperm counts

Speaker notes

In 1992, the British Medical Journal published a controversial meta-analysis. This study (BMJ 1992: 305(6854); 609-13) 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.

Several researchers (Fertil Steril 1996: 65(5); 1044-6 and Fertil Steril 1995: 63(4); 887-93) 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 United States 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.

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 multi-center trial where each center gets to use its own protocol and where some of the centers are left out.

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.

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.

# Major issues in meta-analysis

- Heterogeneity
  - Were apples combined with oranges?
- Publication bias
  - Were some apples left on the tree?
- Study quality
  - Were all the apples rotten?
- Interpretability
  - Did the pile of apples amount to more than just a hill of beans?

Speaker notes

There are four major issues that you should be aware of: heterogeneity, publication bias, study quality, and interpretability. We will tackle each of these in some detail.

# Break #1

- What you have learned
  - What is meta-analysis
- What’s coming next
  - Heterogeneity

# Too much heterogeneity is bad

- Mixing apples and oranges
- Example: statins
  - Primary versus secondary prevention

Speaker notes

Meta-analyses should not have too broad an inclusion criteria. Including too many studies can lead to problems with “apples-to-oranges” comparisons. For example, when you are studying the effect of cholesterol lowering drugs, it makes no sense to combine a study of patients with recent heart attacks with another study of patients with high cholesterol but no previous heart attacks.

There is a lot of variability in how research is conducted. Even in carefully controlled randomized control trials, researchers have tremendous discretion (Am J Med 1987: 82(3); 498-510.). Sometimes this discretion creates heterogeneity among studies, making it difficult to combine the studies.

# Heterogeneity in patient populations

- Demographics of patients
- Exclusion critieria
- Baseline health

Speaker notes

You might see heterogeneity in the composition of the treatment and control groups. The patients may be older or younger. They may have different race and ethnicity proportions.

Researchers can differ in the inclusion and exclusion criteria.

Even if these criteria do not differ, there may still be differences in the baseline levels of health in the patients, due to geographical differences in the patient population.

# Heterogeneity in sampling

- Independent versus matched samples
- Type of control
- Type of treatment

Speaker notes

The controls could be selected independently, or they could be matched to the treatment group subjects.

The control subjects could be given no treatment, a placebo, or a standard treatment.

The treatment could differ, such as differences in dose or timing of a drug.

# Heterogeneity in study design

- Length of follow-up
- Drop-out rates

Speaker notes

Heterogeneity in the design of the study

The length of follow-up for the patients could differ.

The proportion of patients who drop out could differ as well as the proposed statistical treatment of these dropouts.

# Heterogeneity in patient management

- Treatment of comorbid conditions
- Handling of complications
- Physician discretion

Speaker notes

Heterogeneity in the management of the patients and in the outcome

How comorbid conditions are treated. Can patients take a variety of medicines unrelated to the study?

How complications are handled. Do patients with complications stay in the study and if so, what treatment options are allowed.

How much discretion the patient’s physician has in controlling patient care. Some studies allow a wide degree of discretion in treatment but others are very detailed in what the physician is allowed to do. If a physician in these studies sees a need for something that is not allowed, the patient has to be dropped from the study.

# Heterogenity in outcome measures

- Example: hypertension treatment
  - cardiovascular deaths
  - cardiovascular events
  - cerebrovascular deaths
  - cerebrovascular events
  - deaths due to any cause

Speaker notes

The outcome measure itself could differ. For example, Abramson (Public Health Rev 1990: 18(1); 1-47) discusses a meta-analysis of hypertension treatment in the elderly. Some of the studies examined cardiovascular deaths and others examined cardiovascular events. Other studies examined cerebrovascular deaths, cerebrovascular events, cardiac deaths, coronary heart disease deaths, and/or total deaths.

# Examples of heterogeneity

- Antiretroviral combination therapy
  - Effective in shorter trials
  - Ineffective in longer trials
- Dust mite control measures
  - Chemical or physical interventions
  - Crossover or parallel design
  - Variation in blinding
  - Age of patients
  - Duration of trial

Speaker notes

In a meta-analysis (BMJ 2002; 324(7340): 757) looking at antiretroviral combination therapy, 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).

In a meta-analysis (BMJ 1998: 317(7166); 1105-1110) looking at dust mite control measures to help asthmatic patients, the studies exhibited heterogeneity across several factors. Six studies examined chemical interventions, thirteen examined physical interventions, and four examined a combination approach. Nine of these trials were crossovers, and in the remaining fourteen, there was a parallel control group. Seven studies had no blinding, three studies had partial blinding, and the remaining thirteen studies used a double blind. In nine studies the average age of the patients was only 9 or 10 years, but nine other studies had an average age of 30 or more. Eleven studies lasted eight weeks or less and five studies lasted a full year. You can find a table summarizing these studies on the web.

# How to handle heterogeneity

- Some heterogenity is actually encouraged
- Subgroup analysis
- Meta-regression

Speaker notes

Some level of heterogeneity is acceptable. After all, the purpose of research is to generalize results to large groups of patients. Furthermore, demonstrating that a treatment shows consistent results across a variety of conditions strengthens our confidence in that treatment.

Nevertheless, you should be aware of the problems that excessive heterogeneity can cause. Mixing apples and oranges may not be so bad; you get a fruit salad this way. But when heterogeneity becomes too large, you might end up combining not apples and oranges but apples and onions.

Subgroup analysis

When there is substantial heterogeneity, you can look and compare subgroups of the studies. In a meta-analysis (BMJ. 2000; 321(7273): 1371-6) studying atypical antipsychotics, the dose of the comparison drug (haloperidol or an equivalent) varied substantially. Among those studies where the dose of haloperidol was greater than 12 mg/day, atypical antipsychotics showed advantages in efficacy or tolerability. When the dose was less than or equal to 12 mg/day, the atypical antipsychotics showed no advantages in these areas.

Meta-regression

You can try to adjust for heterogeneity in a meta-analysis. This would work very similarly to the adjustment for covariates in a regression model. For example, Derry et al (BMJ 2000: 321(7270); 1183-7) used meta-analysis to see if long term aspirin therapy was associated with problems with gastrointestinal hemorrhage. They identified 24 studies that looked at aspirin as a preventive measure against heart attacks. In each of these studies, the rate of gastrointestinal hemorrhages were recorded for both the aspirin group and the placebo or no treatment group. There was substantial heterogeneity in the dosage of aspirin used in the studies, however, with some studies giving as little as 50 mg/day and some as much as 1500 mg/day.

This was actually good news in a way, because the researchers wanted to see if the risk of gastrointestinal hemorrhage was dependent on the dose of aspirin. A plot of the dose versus the risk showed that there was indeed an increased risk but that this risk seemed to be unrelated to the dosage.

# Inclusion of very old studies

- Varies by discipline
- Look for “game changing” events
  - Surfactants in neonatology
  - Anti-retrovirals for HIV patients
  - Fluouridation of water supplies

Speaker notes

The time frame depends a lot on the topic. Anything in the field of neonatology would have to have a very narrow time window because the field has changed so much so rapidly.

Other areas where the practice of medicine has been much more stable could have wider time windows. I've seen several reviews that have covered half a century of studies.

If you do select a wide time window be sure to see if your results are similar if you restrict yourself to just the most recent studies.

Ask yourself if there was a sudden change in technology that makes any comparisons before and after that technology an apples-to-oranges comparison. So, for example, a meta-analysis involving AIDS patients should restrict itself to the years following the use of AZT.

Also, ask yourself if researchers in your area tend to discount any research that is more than X years old. If so, then your meta-analysis would lose credibility among those researchers if it included studies older than X.

# Test of heterogeneity

- Cochran’s Q
  - Null hypothesis: homogeneity
  - Reject if chi-square statistic much larger than degrees of freedom
- I-squared
  - Descriptive measure
  - Always between 0% and 100%
  - Below 25% implies homogeneity

# Break #2

- What you have learned
  - Heterogeneity
- What’s coming next
  - Study quality

# Were all of the apples rotten?

- “You can’t make a silk purse out of a sow’s ear”
- Meta-analysis will amplify study flaws
- General (but not universal) trend
  - Lower quality leads to over-optimism

Speaker notes

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.

There is a general trend. It doesn’t occur all the time, but it is quite common. That is a trend that lower quality studies tend to lead to over-optimism. Compared to higher quality studies, they tend to overstate the effectiveness of a treatment.

# Observational studies in a meta-analysis

- Originally very controversial
  - An exercise in “mega-silliness”
- Analyze selective subgroups
  - Insights into size and direction of biases

Speaker notes

The use of meta-analysis on observational studies is very controversial. Some experts have argued that the biases inherent in observational studies make a meta-analysis an exercise in mega-silliness. But even those experts who do not take such an extreme viewpoint warn that the current statistical methods for summarizing the results of observational studies may grossly understate the amount of uncertainty in the final result (BMJ 1998: 316(7125); 140-4).

Sensitivity analysis may be a useful way of highlighting the uncertainties in a meta-analysis of observational studies. Restricting the meta-analysis to selective subgroups of the data can yield insight into the size and direction of biases in observational studies. For example, the researchers could contrast case-control designs with cohort designs, with the latter expected to show less bias, in general. Or the researchers could compare retrospective studies to prospective studies, where again, the latter is expected to show less bias in general. Another possibility for comparison involve comparing studies by the amount to which measurement error is expected to cause problems. In general, researchers should try to stratify the observational studies by known sources of bias.

Etminan et al (BMJ. 2003; 327(7407): 128) examined the risk of Alzheimer’s disease in users of non-steroidal anti-inflammatory drugs. They identified six cohort studies which showed a combined relative risk of 0.84 (95% CI 0.54 to 1.05) and three case-control studies which showed a much lower combined relative risk, 0.62 (95% CI 0.45 to 0.82).

# Meta-analyses of randomized trials

- Often a primary inclusion factor
- Other quality concerns
  - Lack of blinding
  - High dropout rates

Speaker notes

Some meta-analyses restrict their attention to randomized trials because these studies are less likely to have problems with bias. In other words, they wish to avoid mixing bad observational apples with good randomized trial apples. Sometimes further restrictions can be made on the basis of partial or full blinding of results or on the proper accounting of dropouts.

Concato et al (NEJM 2000: 342(25); 1887-1892) evaluated clinical topics where there were publications of both randomized controlled trials and observational studies. In this review, the observational studies produced results quite similar to the randomized studies.

Even for randomized trials, sensitivity analysis may help. Researchers can use “quality scores” to rate individual studies and then see what happens when studies are restricted to those of highest quality only.

For example, Lucassen et al (BMJ 1998; 316(7144): 1563-9) looked at interventions for infant colic. Although substituting soy milk for cows milk appeared to have an effect, this effect disappeared when only studies of high methodological quality were considered.

# Incomplete report of quality issues

- Fairly common
- Assumption: guilty until proven innocent
  - But possibly just an oversight?

Speaker notes

Many times, the reporting of a study will be inadequate, and this will make it impossible to assess the quality of a study. There is indeed empirical evidence that incomplete reporting is associated with poor quality (JAMA 1995: 273(5); 408-12). In such a case, a “guilty until proven innocent” approach may make sense (BMJ 2001: 323(7303); 42-6). For example, if the authors fail to mention whether their study was blinded, assume that it was not. You might expect that authors are quick to report strengths of their study, but may (perhaps unconsciously) forget to mention their weaknesses. On the other hand, Liberati (J Clin Oncol 1986: 4(6); 942-51) rated the quality of 63 randomized trials, and found that the quality scores increased by seven points on average on a 100 point scale after talking to the researchers over the telephone. So some small amount of ambiguity may relate to carelessness in reporting rather than quality problems.

# Meta-analysis of studies with small sample sizes

- Be very cautious
  - Publication bias is a much bigger threat

Speaker notes

Some experts advocate great caution in the assessment of meta-analyses where all of the trials consist of small sample size studies. The effect of publication bias can be far more pronounced here than in situations where some medium and large size trials are included.

# Intentional exclusion of studies

- Where to draw the line?
- Example: mammography for 40-50 year old women
  - Seven studies: positive result
  - Two best studies only: negative result

Speaker notes

In any meta-analysis, you have to draw a line somewhere. Studies that fail to meet your criteria will not be included in the results. But this can lead to serious controversy. In a Cochrane Review of mammography (Cochrane 2001: (4); CD001877), seven studies were identified, but only two were of sufficient quality to be used. The Cochrane Review of these two studies reached a negative conclusion, but would have reached an opposite conclusion if the other five studies were added back in (BMJ. 2001; 323(7319): 956).

# Break #3

- What you have learned
  - Study quality
- What’s coming next
  - Publication bias

# Were some apples left on the tree?

- Publication bias
  - Some research studies never get published
  - These studies more likely to be negative
- Registration of clinical trials has helped

Speaker notes

Many important studies are never published; these studies are more likely to be negative (Dickersin 1990). This is known as publication bias. The inclusion of unpublished studies, however, is controversial (Cook 1993).

Publication bias is 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. Much of what has been learned about publication bias comes from the social sciences, less from the field of medicine. In medicine, three studies have provided direct evidence for this bias. Prevention of publication bias is important both from the scientific perspective (complete dissemination of knowledge) and from the perspective of those who combine results from a number of similar studies (meta-analysis). If treatment decisions are based on the published literature, then the literature must include all available data that is of acceptable quality. Currently, obtaining information regarding all studies undertaken in a given field is difficult, even impossible. Registration of clinical trials, and perhaps other types of studies, is the direction in which the scientific community should move.

Another aspect of publication bias is that the delay in publication of negative results is likely to be longer than that for positive studies. For example, Stern and Simes 1997 showed that among 130 clinical trials, the median time to publication was 4.7 years among the positive studies and 8.0 years among the negative studies. So a meta-analysis restricted to a certain time window may be more likely to exclude published research that is negative.

Many experts are advocating the registration of trials as a way of avoiding publication bias. If trials are registered prospectively (i.e., prior to data collection and analysis) then they can be included in any appropriate meta-analysis without worry about publication bias.

# Duplicate publication

- Some studies pubished twice
- Multiple publications more likely to be positive - Bias from double counting
- Hard to track duplicate publications

Speaker notes

Duplicate publication is the flip side of the publication bias coin. Studies which are positive are more likely to appear more than once in publication. This is especially problematic for multi-center trials where an individual centers may publish results specific to their site. Tramer et al (1997) found 84 studies of the effect of ondansetron on postoperative emesis. Unfortunately, 14 of these studies (17%) were second or even third time publications of the same data set. 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 corss-reference the other studies. Four pairs of identical trials were published by completely different authors without any common authorship

# The limitations of a Pubmed search

- Pubmed only covers 3,000 of the 13,000 medical journals
  - Studies not in Pubmed tend to be negative

Speaker notes

While a Medline search is the most convenient way to identify published research, it should not be the only source of publications for a meta-analysis. Medline searches cover only 3,000 of some 13,000 medical journals (Halvorsen 1992). The studies missed by Medline and other databases are more likely to be negative studies.

Furthermore, these databases may fail to index major journals in the third world that can provide important trials. Egger (1997) cites an interesting example of how Medline excludes most Indian journals, even though these journals are published in English and India produces a significant amount of medical research.

# Foreign language publications

- Convenient to restrict to English publications
- Native language publications more likely to be negative

Speaker notes

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. Interestingly, some studies have shown that the quality of studies published in other languages is comparable to the quality of studies published in English.

# Picking the low hanging fruit

- Articles with full free text
- Articles with abstracts

Speaker notes

In an informal meta-analyis, you should also worry about the tendency for people to preferentially choose articles that are convenient. For example, there is a natural tendency to rely on articles where the full text is available on the Internet or where the abstract is available for review (Wentz 2002).

# How to avoid bias from exclusion of publications

- Multiple bibliographic databases
- Registries
- Examination of bibliographies
- Call for “gray literature”

Speaker notes

Search for studies should involve several bibliographic databases, registries for clinical trials, examination of bibliographies of all articles found, the so-called gray literature (presentation abstracts, dissertations, theses, etc.) and a letter calling for unpublished papers to be sent out to key researchers.

Consider the search strategy adopted in Evers et al 2001.

Relevant trials were identified in the Cochrane Menstrual Disorders and Subfertility Group’s specialised register of controlled trials. A MEDLINE search, using the group’s search strategy, was performed for the period 1966-2000. Also, hand searching was performed of 22 specialist journals in the field from their first issue till 2000. Cross references and references from review articles were checked.

# Subjectivity in article selection

- Blinding
  - No author list
  - No university affiliation
  - Methods section only
- Duplicate extraction of information
  - Look for 90% or better agreement
- Detailed protocol

Speaker notes

“Blinding,” a common tool in other research areas should also be used in meta-analyses. Blinding prevents the differential application of inclusion/exclusion criteria. The people deciding whether a paper meets the inclusion/exclusion criteria should be unaware of the authors of that paper and the journal. They should also include or exclude the paper on the basis of the methods section only; they should not see the results section until later.

There is empirical evidence, however, that blinding does not affect the conclusions of a meta-analysis (Jadad et al 1996, Berlin et al 1997). Furthermore, blinding takes substantial time and energy.

Data should be extracted from papers by multiple sources and their level of agreement should be assessed. Researchers have found disagreements even on such fundamental concepts such as whether a study was positive or negative (Glass 1981).

Like any other research project, an overview or meta-analysis needs a protocol. Unfortunately, many published meta-analyses do not state whether a protocol was used (Sacks 1992). The protocol should specify: the inclusion/exclusion criteria for studies; a detailed description of the process used to identify studies; and the statistical methods used to combine results. Without a protocol, the meta-analysis research is not reproducible.

Authors have been shown to be biased in the articles that they cite in the bibliographies of their research papers (Gotsche 1987; Ravnskov 1992). This same bias could potentially affect the selection of articles in a meta-analysis.

If the authors do not present objective criteria for the selection of articles in their overview or meta-analysis, then you should be concerned about possible conscious or sub-conscious bias in the selection process.

Researchers should also list all of the articles found in the original search, not just the articles used. This allows others to examine whether the inclusion/exclusion criteria were applied appropriately.

# Detecting and correcting for publication bias

- Compare to the unpublished studies you did find
- Stratify by sample sizes
- Funnel plot

Speaker notes

Sensitivity analysis is also useful here. If the results from published studies are comparable to the results from unpublished studies, for example, then publication bias is less of a concern. Along the same lines, the authors can estimate the number of undiscovered negative studies that would be required to overturn the results of this meta-analysis.

Publication bias is also more likely to occur for studies with small sample sizes. If the results of a meta-analysis are stratified by the sample sizes in the studies, a shift away from the null hypothesis in the smaller studies would be a warning flag about the possibility of publication bias. Statistical and graphical methods have been proposed to examine this further but you should be cautious, however, because sometimes there are other explanations. For example, smaller studies may tend to use less rigorous designs and these designs may be associated with exaggerated effects (Sterne et al 2001).

McManus et al (1998) highlight the importance of consulting experts in the area. They we trying to identify all publications associated with near patient testing, tests where the results are available without sending materials to a lab. The authors used a search of electronic databases, a survey of experts in the area, and hand searching of specific journals. The electronic databases yielded the most number of publications, 50, but still missed 52 publications found by the other two methods.

Copas and Shi (2000) present a re-analysis of a meta-analysis on lung cancer that adjusts for publication bias, but this adjustment is controversial (Johnson et al 2000).

# Break #4

- What you have learned
  - Publication bias
- What’s coming next
  - Interpretability of results

# Did the pile of apples amount to more than just a hill of beans?

- A significant meta-analysis is not enough
- Effect sizes difficult to interpret

Speaker notes

It’s not enough to know that the overall effect of a therapy is positive. You have to balance the magnitude of the effect versus the added cost and/or the side effects of the new therapy. Unfortunately, most meta-analyses use an effect size (the improvement due to the therapy divided by the standard deviation). The effect size is unitless, allowing the combination of results from studies where slightly different outcomes with slightly different measurement units might have been used.

# Vote counting

- Categorize individual studies
  - Are some/most positive?
- Negative, but trending studies

Speaker notes

Avoid “vote counting” or the tallying of positive versus negative studies. Vote counts ignore the possibility that some studies are negative solely because of their sample size. Abramson (1990) notes, for example, a meta-analysis of parenteral nutrition in cancer patients undergoing chemotherapy. Although each of the seven randomized control trials in the meta-analysis failed to achieve statistical significance, the pooled results were highly significant.

# Unitless measures

- Convert back to original units
- Example smoking cessation efforts during pregnancy
  - Statistically significant
  - Only a 28 gram difference in birthweight

Speaker notes

When you are examining a continuous outcome measure, you should be sure that the results are presented in interpretable units. A measure of effect size does not help you much because it is unitless and impossible to interpret. Consider a store that is offering a sale and announces boldly

“All prices reduced by 0.8 standard deviations!”

One meta-analysis shows how important it is to express measurements in interpretable units. Lumley et al (2001) studied the effect of smoking cessation programs on the health of the fetus and infant. One of the outcome measures was birth weight, and the study showed that the typical program can improve birth weight by a statistically significant amount. The researchers then quantified the amount: 28g (95% confidence interval 9 to 49).

Keep in mind that this is measuring the effectiveness of the smoking cessation program, and not the effect of smoking cessation directly. Typically, you would have to send about 12 to 16 women to these programs in order to get one extra woman to quit smoking. So the effect seen here reflects, in part, how difficult it is to get people to change their behavior.

Still the small size of the effect is important. If you want to assess the costs and benefits of smoking cessation programs, it helps to know that the impact of the typical smoking cessation program on birth weight is quite small. This provides a useful yardstick for comparison to other prenatal interventions.

# Break #5

- What you have learned
  - Interpretability of results
- What’s coming next
  - Numeric summaries

# Summary statistics for continuous outcomes

- Standardized mean difference (SMD)
  - $\frac{\bar X_2-\bar X_1}{Estimated\ standard\ deviation}$
  - Be consistent with direction
- Choices for estimated standard deviation
  - Pooled standard deviation (Cohen's d)
  - Bias adjustment (Hedges' g)
  - Heteroscedascity?
    - Use control variance
    - Average the two variances

Speaker notes

You need a single number that summarizes what each study found. For continuous outcomes, this take the form of something akin to an effect size. It is not quite the same but close. It is a unitless quantity, which in theory allows you to combine results from a study where on study measures obesity as body mass index and another measures it as waist to hip ratio. You may not be comfortable with this, of course.

The standardized mean difference is the difference in the means divided by some measure of standard deviation. Be consistent in how you define differences. Define it so that if half of the studies use an outcome where a large value is good and half of the studies use an outcome where a small value is good, you flip the difference (group 1 minus group 2) for half of the studies. It doesn't matter which you flip around, just be consistent so a positive value always the same thing across all studies .

There is some debate about what to put for the estimated standard deviation. The simplest choice, the pooled standard deviation is called Cohen's d and is equal to the effect size. There is a bias adjustment that some researchers use called Hedges' g. It is only important for small sample sizes.

If there is substantial differences between the variances of the control and treatment group, you have to choose between using the variance of the control group and the average of the control and treatment variances. Note that this is not quite the same as averaging of the control and treatment standard deviations.

# Summary statistics for categorical outcomes

- Odds ratio
  - Always on a log scale
- Relative risk
  - Always on a log scale
- Absolute difference

Speaker notes

There is no consensus on how best to summarize results from a single study. Recall the debate over the odds ratio and the relative risk that I mentioned earlier.

Add to that the concern that the relative risk is more likely to have problems with heterogeneity. A relative risk for a fairly common event has a sharp upper bound. You can’t expect to see a tripling of risk if the event rate in the control group is 40%.

Nevertheless, many researchers prefer to use the relative risk because it is more interpretable.

Both the odds ratio and the relative risk should be analyzed on the log scale.

The absolute difference involves subtraction rather than division. This measure also has its proponents.

Because of the lack of consensus, seek some guidance first. Talk to your boss and your colleagues. See what is commonly done in your medical discipline.

# Description of art-malpresentations data, 1 of 3

```
data_dictionary: art-malpresentations.csv
description: |
  Assisted reproductive technology (ART) has helped many infertile couples, but there are concerns about the risks during pregnancy and child birth. This paper found 11 studies that examined whether the risk of malpresentation (e.g., breech births) was higher in ART pregnancies compared to natural conception (NC).
source: |
  Konstantinos Stavridis, Maria Pisimisi, Olga Triantafyllidou, Theodoros Kalampokas, Nikolaos Vlahos & Stavroula L. Kastora (2024) The association of assisted reproductive technology with fetal malpresentation: a systematic review and meta-analysis, The Journal of Maternal-Fetal & Neonatal Medicine, 37:1, DOI: 10.1080/14767058.2024.2313143
```

# Description of art-malpresentations data, 2 of 3

```
author:
  label: first author of study
study_type:
  values:
  - casecontrol
  - cohort
```

# Description of art-malpresentations data, 3 of 3

```
art_events:
  label: number of malpresentation in ART groups
art_total:
  label: number of patients in ART groups
nc_events:
  label: number of malpresentations in NC groups
nc_total:
  label: number of patients in NC groups
```

# ART results, 1 of 6

**Effect Size Estimates for Subgroup Analysis**

| | Effect Size | Std. Error | Z | Sig. (2-tailed) | 95% Confidence Interval | | Exp. Effect Size | Exp. 95% Confidence Interval | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Lower | Upper | | Lower | Upper |
| case-control | .320 | .2477 | 1.292 | .196 | -.166 | .805 | 1.377 | .847 | 2.238 |
| cohort | .441 | .0837 | 5.275 | <.001 | .277 | .606 | 1.555 | 1.320 | 1.832 |
| Overall | .410 | .0942 | 4.353 | <.001 | .226 | .595 | 1.507 | 1.253 | 1.813 |

# ART results, 2 of 6

**Effect Size Estimates for Individual Studies**

| | ID | Effect Size | Std. Error | Z | Sig. (2-tailed) | 95% Confidence Interval | | Exp. Effect Size | Exp. 95% Confidence Interval | | Weight | Weight (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper | | Lower | Upper | | |
| case-control | Frydman | 1.170 | .2899 | 4.036 | <.001 | .602 | 1.738 | 3.222 | 1.825 | 5.686 | 6.455 | 5.7 |
| | Isaksson | .204 | .3170 | .642 | .521 | -.418 | .825 | 1.226 | .659 | 2.282 | 5.834 | 5.2 |
| | Reubinoff | .302 | .3076 | .983 | .326 | -.301 | .905 | 1.353 | .740 | 2.472 | 6.042 | 5.4 |
| | Stojnic | .357 | .1649 | 2.163 | .031 | .033 | .680 | 1.429 | 1.034 | 1.974 | 10.193 | 9.0 |
| | Tan | -.399 | .2528 | -1.578 | .115 | -.894 | .097 | .671 | .409 | 1.101 | 7.418 | 6.6 |
| cohort | Chen | .280 | .0363 | 7.719 | <.001 | .209 | .351 | 1.323 | 1.232 | 1.421 | 13.846 | 12.3 |
| | Noli | .568 | .0414 | 13.725 | <.001 | .486 | .649 | 1.764 | 1.627 | 1.913 | 13.771 | 12.2 |
| | Romundstad | .415 | .0478 | 8.670 | <.001 | .321 | .509 | 1.514 | 1.379 | 1.663 | 13.662 | 12.1 |
| | Slavov | .915 | .2376 | 3.854 | <.001 | .450 | 1.381 | 2.498 | 1.568 | 3.979 | 7.853 | 7.0 |
| | Stern | .181 | .0337 | 5.372 | <.001 | .115 | .247 | 1.198 | 1.122 | 1.280 | 13.881 | 12.3 |
| | Zsirai | .577 | .0457 | 12.623 | <.001 | .487 | .667 | 1.781 | 1.628 | 1.947 | 13.700 | 12.2 |

# ART results, 3 of 6

**Test of Homogeneity**

| | Chi-square (Q statistic) | df | Sig. |
|---|---|---|---|
| case-control | 16.857 | 4 | .002 |
| cohort | 87.145 | 5 | <.001 |
| Overall | 104.362 | 10 | <.001 |

**Test of Subgroup Homogeneity**

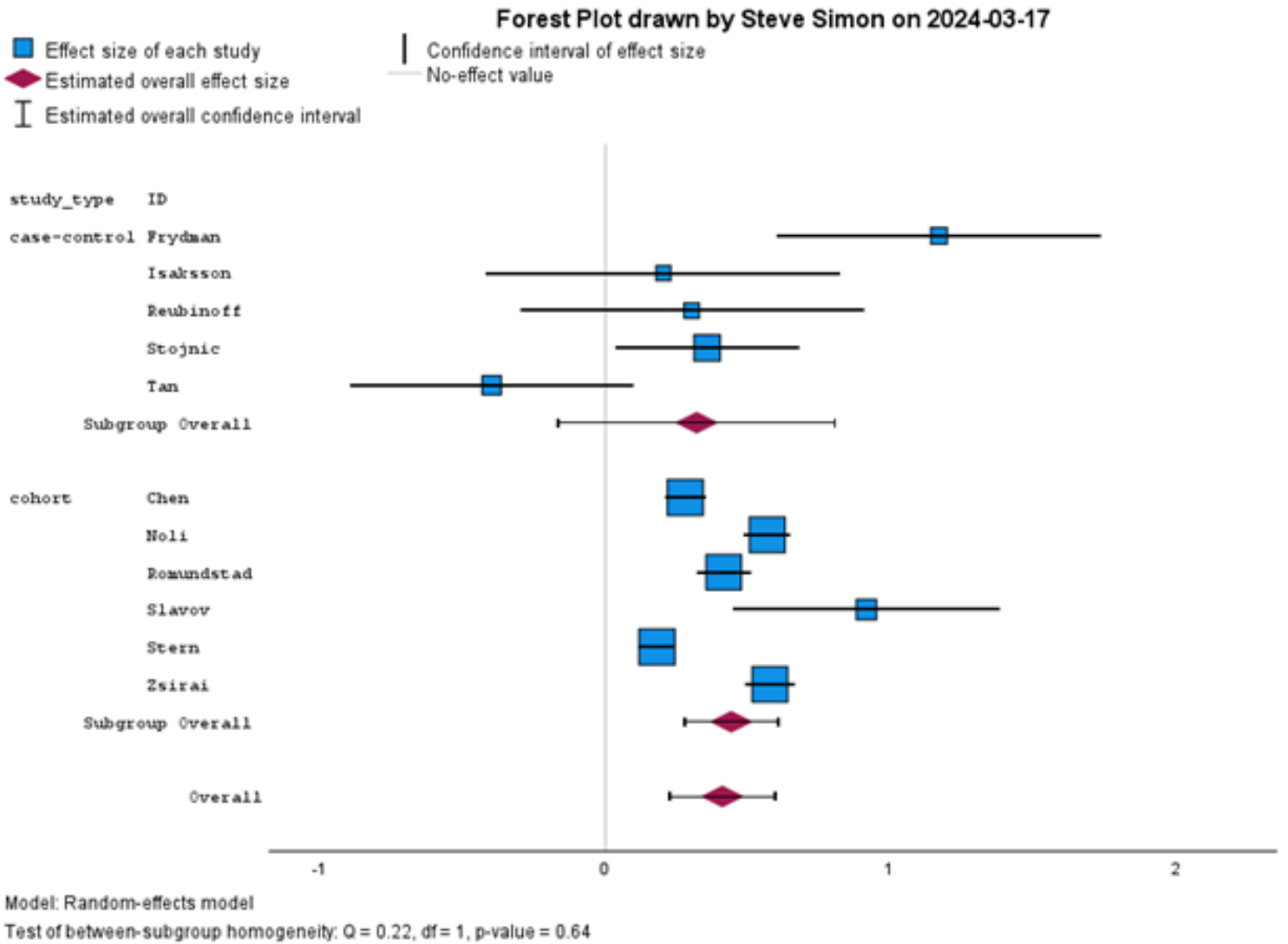
| | Chi-square (Q statistic) | df | Sig. |
|---|---|---|---|
| study_type | .216 | 1 | .642 |

# ART results, 4 of 6

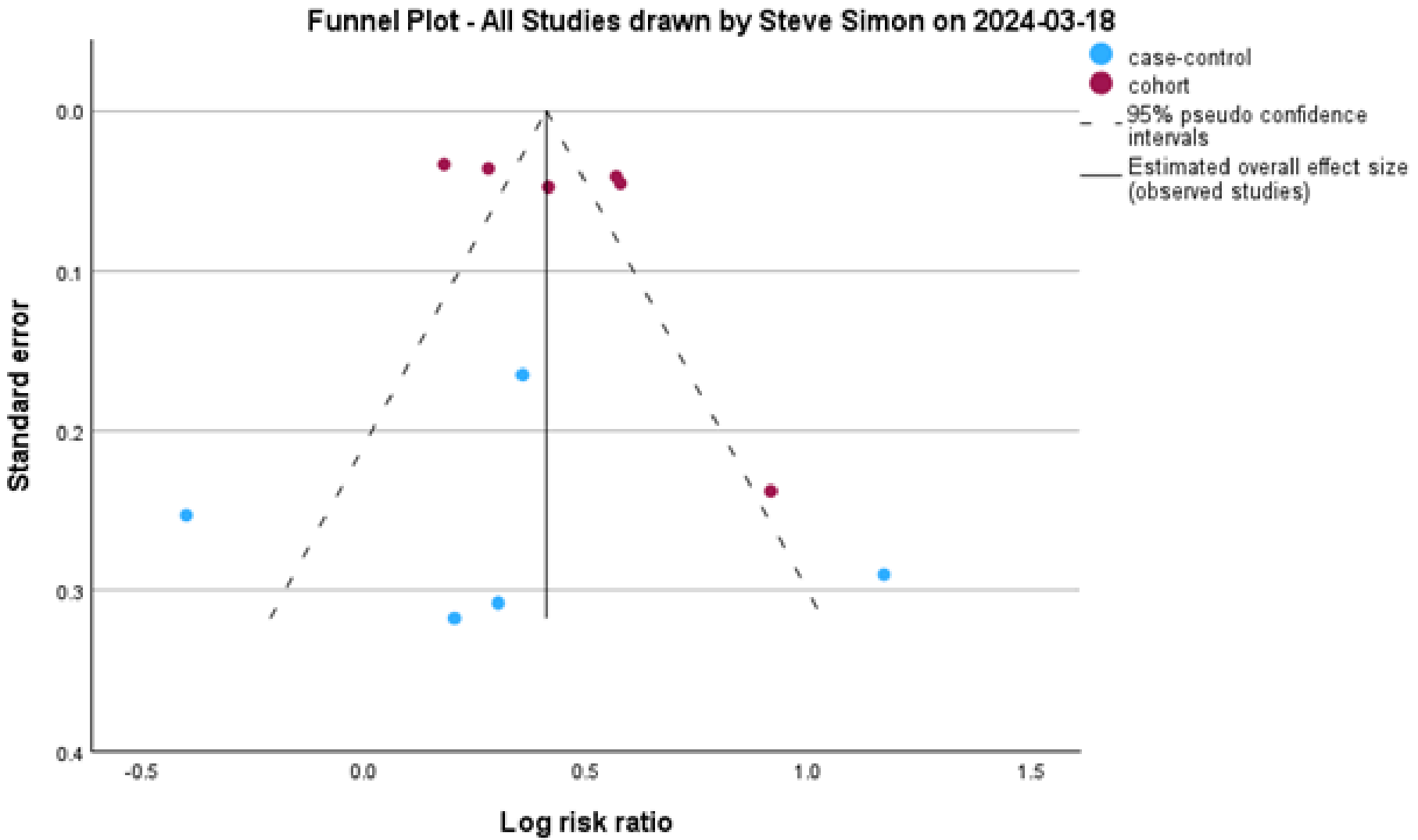
**Heterogeneity Measures**

| | | |
|---|---|---|
| case-control | Tau-squared | .235 |
| | H-squared | 4.637 |
| | I-squared (%) | 78.4 |
| cohort | Tau-squared | .036 |
| | H-squared | 19.023 |
| | I-squared (%) | 94.7 |
| Overall | Tau-squared | .071 |
| | H-squared | 19.377 |
| | I-squared (%) | 94.8 |

# ART results, 5 of 6



# ART results, 6 of 6



# Description of vaccine-willingness data, 1 of 3

```
data_dictionary: vaccine-willingness.yaml

description: |
  This paper identified seven studies that examined vaccine literacy (VL). Patients in these studies also stated their  vaccine prefernce and were categorized  as willing or unwilling.

source: |
  Isonne C, Iera J, Sciurti A, Renzi E, De Blasiis MR, Marzuillo C, Villari P, Baccolini V. How well does vaccine literacy predict intention to vaccinate and vaccination status? A systematic review and meta-analysis. Hum Vaccin Immunother. 2024 Dec 31;20(1):2300848. doi: 10.1080/21645515.2023.2300848. PMID: 38174706; PMCID: PMC10773666.
```

# Description of vaccine-willingness data, 2 of 3

```
author:
  label: first author of publication
vaccine_type:
  label: Type of vaccination
  values:
  - booster
  - primary
```

# Description of vaccine-willingness data, 3 of 3

```
willing_mean:
  label: Mean VL in willing patients
willing_sd:
  label: Standard deviation of VL in willing patients
willing_n:
  label: Number of willing patients
unwilling_mean:
  label: Mean VL in unwilling patients
unwilling_sd:
  label: Standard deviation of VL in unwilling patients
unwilling_n:
  label: Number of unwilling patients
```

# Vaccine results, 1 of 6

**Effect Size Estimates for Subgroup Analysis**

| | Effect Size | Std. Error | Z | Sig. (2-tailed) | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower | Upper |
| booster | .631 | .1046 | 6.032 | <.001 | .426 | .836 |
| primary | .095 | .0421 | 2.245 | .025 | .012 | .177 |
| Overall | .341 | .1146 | 2.973 | .003 | .116 | .566 |

# Vaccine results, 2 of 6

**Effect Size Estimates for Individual Studies**

| | ID | Effect Size | Std. Error | Z | Sig. (2-tailed) | 95% Confidence Interval Lower | 95% Confidence Interval Upper | Weight | Weight (%) |
|---|---|---|---|---|---|---|---|---|---|
| booster | Achrekar | .434 | .0777 | 5.586 | <.001 | .282 | .587 | 11.108 | 14.6 |
| | Batra | .814 | .0942 | 8.638 | <.001 | .629 | .998 | 10.769 | 14.2 |
| | Yadete | .657 | .0456 | 14.389 | <.001 | .567 | .746 | 11.619 | 15.3 |
| primary | Biasio | .100 | .1254 | .797 | .425 | -.146 | .346 | 10.029 | 13.2 |
| | Correa-Rodriguez | .133 | .1121 | 1.190 | .234 | -.086 | .353 | 10.358 | 13.6 |
| | Gendler | .148 | .0962 | 1.542 | .123 | -.040 | .337 | 10.726 | 14.1 |
| | Gutierrez | .065 | .0566 | 1.148 | .251 | -.046 | .176 | 11.469 | 15.1 |

# Vaccine results, 3 of 6

**Test of Homogeneity**

| | Chi-square (Q statistic) | df | Sig. |
|---|---|---|---|
| booster | 10.485 | 2 | .005 |
| primary | .705 | 3 | .872 |
| Overall | 104.335 | 6 | <.001 |

**Test of Subgroup Homogeneity**

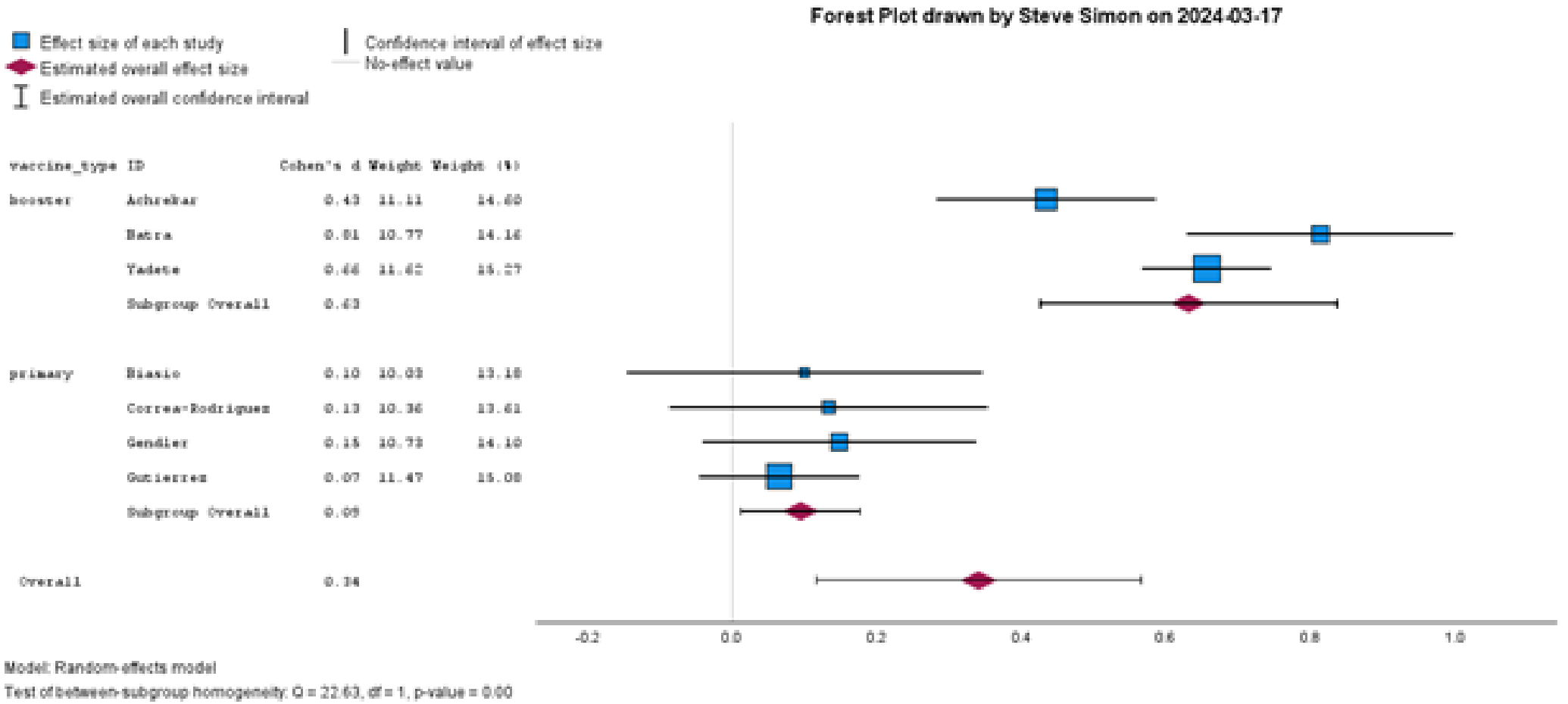
| | Chi-square (Q statistic) | df | Sig. |
|---|---|---|---|
| vaccine_type | 22.630 | 1 | <.001 |

# Vaccine results, 4 of 6

**Heterogeneity Measures**

| | | |
|---|---|---|
| booster | Tau-squared | .027 |
| | H-squared | 6.505 |
| | I-squared (%) | 84.6 |
| primary | Tau-squared | .000 |
| | H-squared | 1.000 |
| | I-squared (%) | .0 |
| Overall | Tau-squared | .084 |
| | H-squared | 15.370 |
| | I-squared (%) | 93.5 |

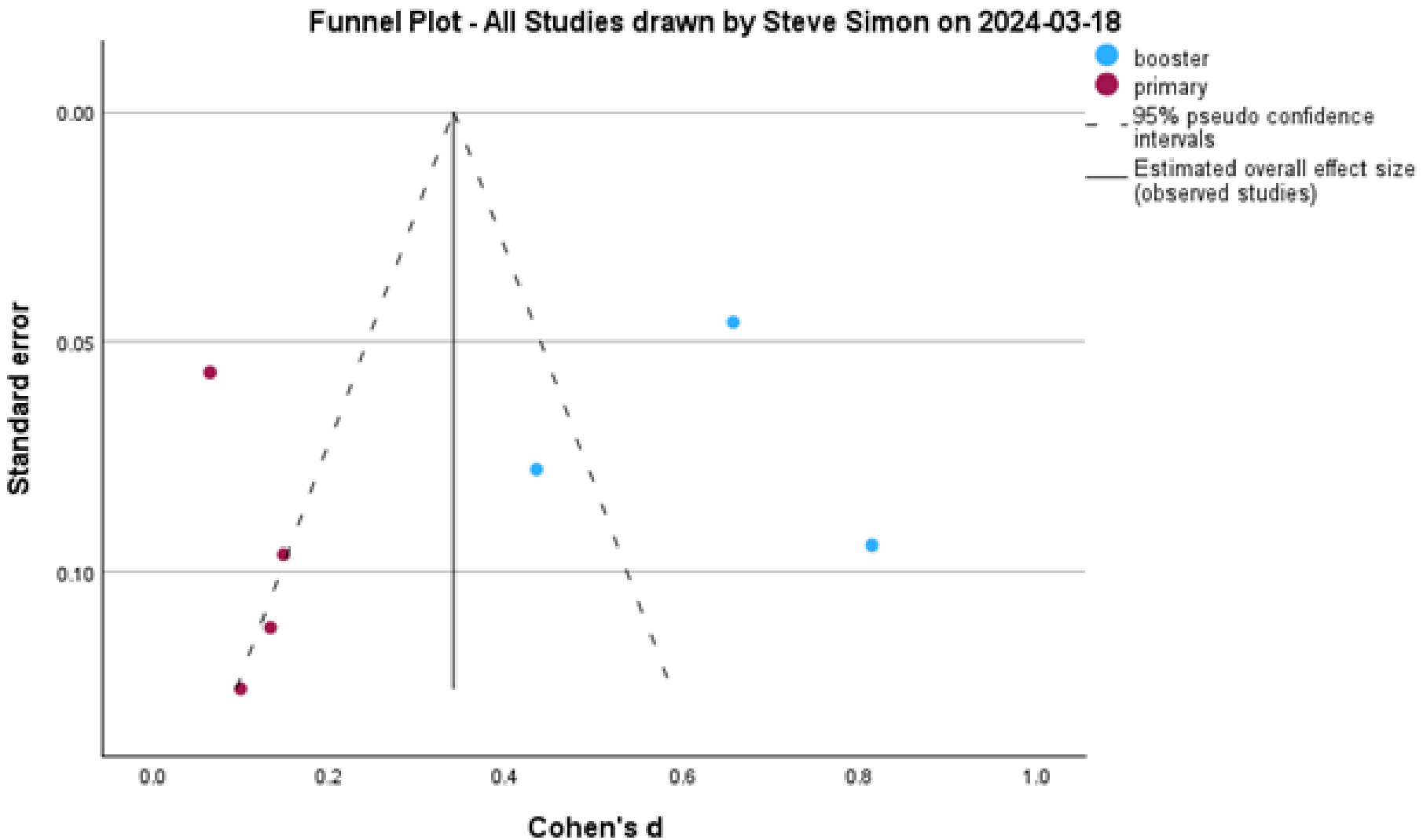
# Vaccine results, 5 of 6




Model: Random-effects model

Test of between-subgroup homogeneity: Q = 22.63, df = 1, p-value = 0.00

# Vaccine results, 6 of 6



# Live demo, Numeric summaries

# Break #6

- What you have learned
  - Numeric summaries
- What’s coming next
  - Strengths and weaknesses

# Where does meta-analysis sit on the hierarchy of evidence?

- Above randomized clinical trials?
  - Heterogeneity improves generalizability
  - Avoids accusations of “cherry-picking”
- Below randomized clinical trials?
  - Too much variation in conduct of trials
  - Uncertain sampling framework

# Unstated benefits of meta-analysis

- Characterizing sources of bias
- Identifying research gaps
- Insight into when to stop researching

# Summary

- What you have learned
  - What is meta-analysis
  - Heterogeneity
  - Study quality
  - Publication bias
  - Interpretability of results
  - Numeric summaries
  - Strengths and weaknesses

# Additional references

- See speaker notes for additional references

Speaker notes

Meta-analysis possesses certain flaws and limitations that preclude its use as a broad-based methodologic approach for formulating definitive therapeutic recommendations. – Boden 1992.

Meta-Analysis: A Review of Pros and Cons. Abramson J. Public Health Reviews 1990 18(1): 1-47.

Does Blinding of Readers Affect the Results of Meta-Analyses? Jesse A Berlin, on behalf of University of Pennsylvania Meta-analysis Blinding Study Group. Lancet 1997; 350: 185-186.

Evidence for Decreasing Quality of Semen During Past 50 Years. Carlsen E, Giwercman A, Keiding N, Skakkebaek NE. Bmj 1992; 305(6854): 609-13.

The Existence of Publication Bias and Risk Factors for its Occurrence. Dickersin, K. (1990). Jama 263(10): 1385-9.

Egger (1997)

Surgery or Embolisation for Varicocele in Subfertile Men (Cochrane Review). Evers JL, Collins JA, Vandekerckhove P. Cochrane Database Syst Rev 2001; 1: CD000479.

Should Unpublished Data Be Included in Meta-Analyses. Cook DJ, Guyatt GH, Ryan E, Clifton J, Buckingham L, Willan A, WcIlroy W, Oxman AD. Journal of the American Medical Association, 269: 2749-2753 (1993).

Geographic Variations in Sperm Counts: A Potential Cause of Bias in Studies of Semen Quality. Fisch H; Goluboff ET. Fertil Steril (United States), May 1996, 65(5) p1044-6.

Meta-analysis in Social Research. Glass GV, McGaw B, Smith ML. pp.18-20. Newbury Park CA: Sage (1981).

Comparison of Intrauterine and Intracervical Insemination with Frozen Donor Sperm: A Meta-Analysis. Goldberg JM, Mascha E, Falcone T, Attaran M. Fertil Steril 1999 Nov; 72(5): 792-5.

Reference Bias in Reports of Drug Trials. Gotzsche PC. Bmj 1992 295(6599): 654-6.

Combining Results from Independent Investigations: Meta-Analysis in Clinical Research. Halvorsen KT, Burdick E, Colditz GA, Frazier HS, Mosteller F. pp. 413-426, in Medical Uses of Statistics: 2nd Edition, Bailar JC and Mosteller F (editors), Boston MA: NEJM Books (1992).

Systematic Reviews in Health Care: Assessing the Quality of Controlled Clinical Trials. Peter Jüni, Douglas G Altman, and Matthias Egger. BMJ 2001; 323: 42-46. [Full text]

A Quality Assessment of Randomized Control Trials of Primary Treatment for Breast Cancer. Liberati A, Himel HN, Chalmers TC. J Clin Oncol 1986; 4: 942-951.

Interventions for Promoting Smoking Cessation During Pregnancy (Cochrane Review). Lumley J, Oliver S, Waters E. In: The Cochrane Library, 4, 2001. Oxford: Update Software. www.update-software.com/abstracts/ab001055.htm

Review of the Usefulness of Contacting Other Experts When Conducting A Literature Search for Systematic Reviews R J McManus, S Wilson, B C Delaney, D A Fitzmaurice, C J Hyde, R S Tobias, S Jowett, and F D R Hobbs BMJ 1998; 317: 1562-1563. [Full text]

Sperm Function Assays and Their Predictive Value for Fertilization Outcome in IVF Therapy: A Meta-Analysis. Oehninger S, Franken DR, Sayed E, Barroso G, Kolm P. Hum Reprod Update 2000 Mar-Apr; 6(2): 160-8.

Have Sperm Counts Been Reduced 50 Percent in 50 Years? A Statistical Model Revisited. Olsen GW; Bodner KM; Ramlow JM; Ross CE; Lipshultz LI . Fertil Steril (United States), Apr 1995, 63(4) p887-93

Frequency of Citation and Outcome of Cholesterol Lowering Trials. Ravnskov, U. BMJ 1992 305(6855): 717.

Meta-Analyses of Randomized Control Trials: An Update of the Quality and Methodology. Sacks HS, Berrier J, Reitman D, PAgano D, Chalmers TC. pp. 427-442, in Medical Uses of Statistics: 2nd Edition, Bailar JC and Mosteller F (editors), Boston MA: NEJM Books (1992).

Schulz et al 1995 JAMA

Publication Bias: Evidence of Delayed Publication in a Cohort Study of Clinical Research Projects Jerome M Stern and R John Simes BMJ 1997; 315: 640-645. [Abstract] [Full text]

Systematic Reviews in Health Care: Investigating and Dealing with Publication and Other Biases in Meta-Analysis Jonathan A C Sterne, Matthias Egger, and George Davey Smith BMJ 2001; 323: 101-105. [Full text]

Meta-Analysis of Observational Studies in Epidemiology: A Proposal for Reporting. Donna F. Stroup, PhD, MSc; Jesse A. Berlin, ScD; Sally C. Morton, PhD; Ingram Olkin, PhD; G. David Williamson, PhD; Drummond Rennie, MD; David Moher, MSc; Betsy J. Becker, PhD; Theresa Ann Sipe, PhD; Stephen B. Thacker, MD, MSc; for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group April 19, 2000. JAMA. 2000;283:2008-2012. Also available at www.consort-statement.org/MOOSE.pdf

Impact of Covert Duplicate Publication on Meta-Analysis: A Case Sudy. Martin R Tramèr, D John M Reynolds, R Andrew Moore, and Henry J McQuay. BMJ 1997; 315: 635-640. [Abstract] [Full text]

Cigarette Smoking and Sperm Density: A Meta-Analysis. Vine MF, Margolin BH, Morrison HI, Hulka BS. Fertil Steril 1994 Jan; 61(1): 35-43.

Visibility of Research: FUTON Bias. Wentz R. Lancet 2002 (October 19): 360 (9341); 1256.

The Cochrane Library. www.update-software.com/cochrane/cochrane-frame.html

“The Cochrane Library is an electronic publication designed to supply high quality evidence to inform people providing and receiving care, and those responsible for research, teaching, funding and administration at all levels.”

Meta-Analysis in Clinical Trials Reporting: Has a Tool Become a Weapon? [Editorial]. Boden, W. E. (1992). Am J Cardiol 69(6): 681-6.

A New System for Grading Recommendations in Evidence -Based Guidelines Robin Harbour and Juliet Miller BMJ 2001; 323: 334-336. [Full text]

Rating the Quality of Evidence for Clinical Practice Guidelines. Hadorn DC, Baker D, Hodges JS, Hicks N. J Clin Epidemiol 1996 Jul;49(7):749-54.

This article describes the system for rating the quality of medical evidence developed and used during creation of the Agency for Health Care Policy and Research-sponsored heart failure guideline. Previous approaches to rating evidence were not designed for use in the setting of clinical practice guidelines. The present system is based on the tenet that flaws in research design are serious to the extent they threaten the validity of the results of studies. A taxonomy of major and minor flaws based on that tenet was developed for randomized controlled trials and for cohort and medical registry studies. The use of the system is described in the context of two difficult clinical issues considered by the Panel: the role of coronary artery revascularization and the use of metoprolol.

PMID: 8691224 [PubMed - indexed for MEDLINE]

“Is Meta-Analysis a Valid Approach to the Evaluation of Small Effects in Observational Studies?” Shapiro S. Journal of Clinical Epidemiology. 50(3): 223-229 (1997).

Assessment Criteria www.jr2.ox.ac.uk/bandolier/band6/b6-5.html

Evidence-Based Everything www.jr2.ox.ac.uk/bandolier/band12/b12-1.html

Ionnidis et al 1998. [comparing meta-analyses to large trials]