Meta-Analysis Workshop (part 2)

Michael LaValley
December 12th 2014
Villanova University

BOSTON UNIVERSITY

Fixed Effect Combining

- Each study i provides an effect size estimate d_i of the population value δ
- For the inverse variance weighting method, the d_i should follow a normal distribution

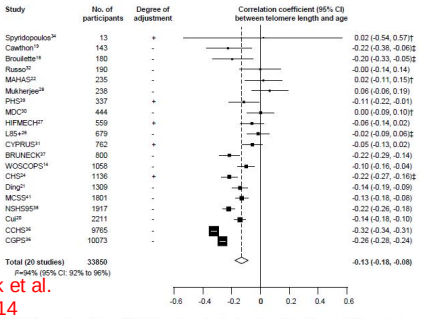
Why Do Meta-regression?

- Sometimes there are interesting questions beyond "*What is the average effect of a treatment or policy?*"
  - The combined effect sizes estimate an **average** effect across all studies in the meta-analysis
- The presence of heterogeneity threatens the validity of a single average effect
  - There may not be a single population effect that applies to all studies

Why Do Meta-regression?

- We have used random effects meta-analysis to address mild to moderate heterogeneity
- In meta-regression we use the study characteristics to try to explain excess variation in effect sizes
- This makes meta-regression a useful **complement** to the methods of combining effect sizes that we have discussed

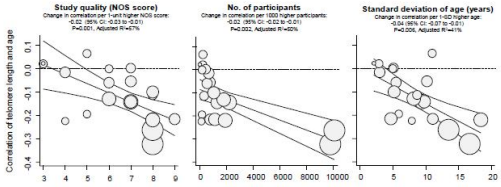
Supplementary Figure 2 | Reported correlation coefficients between telomere length and age in 20 studies*.



Haycock et al. *BMJ* 2014

*20/24 studies reported correlation coefficients between telomere length and age using participants from cohort studies or controls from case–control studies. Sizes of data markers are proportional to the inverse of the within-study variance. The overall correlation was calculated by random effects meta-analysis. CI, confidence interval. Degree of adjustment: - unadjusted, + adjusted for sex. †Obtained through correspondence; ‡estimated from the t-statistic of the linear association between telomere length and age.

Supplementary Figure 3 | Correlation between telomere length and age as a function of selected characteristics.



The plotted data show how the strength of the relationship between telomere length and age varies by the selected characteristic. The plotted solid lines indicate the predicted correlation and its 95% confidence interval. Circle size is proportional to the inverse of the within-study variance. CI – confidence interval, NOS, Newcastle-Ottawa Scale, SD – standard deviation. Study quality, the no. of participants, and standard deviation of age collectively explained 80% of the between-study heterogeneity in reported correlations between telomere length and age. The correlation between telomere length and age was: -0.22 (95% CI: -0.27 to -0.16) in studies with a NOS score ≥8; -0.20 (95% CI: -0.25 to -0.14) in studies with >1,000 participants; and -0.16 (95% CI: -0.22 to -0.10) in studies with an standard deviation of age ≥8 years.

Meta-Regression

- We have a study effect size estimate d_i
- The study-level predictor is x_i

- In the previous example,
  - d_i was the (Fisher z-transformation) of the correlation between telomere length and age
  - x_i was the study quality, number of study participants, standard deviation of age

Fixed Effect Meta-Regression

- d_i is assumed to be **normally distributed** around $\beta_0 + \beta x_i$ with variance due to within study variability

$$d_i \sim N(\beta_0 + \beta x_i, v_i)$$

- The ***within study variability*** v_i is the square of the standard error of the estimate d_i

Random Effects Meta-Regression

- This is a **2-level** model
  - **Level 1**: study effect size d_i provides an estimate of $\delta_i + \beta x_i$

$$d_i \sim N(\delta_i + \beta x_i, v_i)$$

  - The random effect δ_i replaces the fixed intercept

Random Effects Meta-Regression

- **Level 2**: random effects δ_i have a normal distribution around the central value β_0 with variance τ^2

$$\delta_i \sim N(\beta_0, \tau^2)$$

- τ^2 is the **between-study** variance

Random Effects Meta-regression

- This is a *random intercept* model
  - Allows each study to have an different intercept (random effect)
  - But the slope with the predictor is the same for all studies (fixed effect)
- If the predictor explains a lot of between study variation, then the estimate of τ^2 can be smaller in random effects meta-regression than in random effects meta-analysis

BCG Vaccine and TB

- Bacille Calmette-Guerin (BCG) vaccine first used on humans in 1921 to prevent tuberculosis (TB)
- Disagreement over efficacy
  - Published studies range from -57% to 80%

BCG Vaccine and TB

- Historically, BCG vaccine was not used in US
  - Unclear efficacy
  - Low rate of TB in US
  - TB skin-test conversion to positive
- In 80's TB incidence increased in US
  - Association with HIV
  - Multi-drug resistant strains
- Should BCG be used in US?

BCG Vaccine and TB

- To address BCG efficacy a meta-analysis was commissioned
- Meta-analysis indicated
  - Protective effect of BCG (~50%)
  - Variance among study results linked to distance from study site to equator (latitude)

BCG Vaccine and TB

From Higgins and Thompson 2004

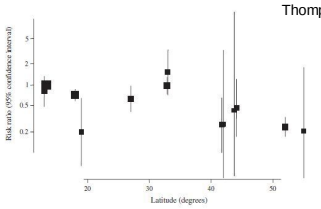
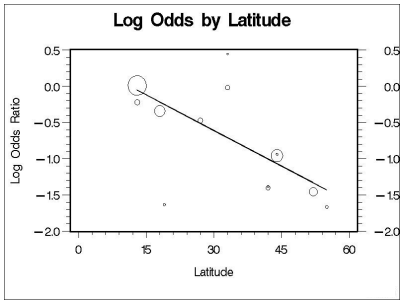


Figure 1. Results from 13 trials of BCG vaccines for preventing tuberculosis [14], plotted against absolute latitude in degrees.

The studies are ordered left to right according to increasing latitude

BCG Vaccine and TB

Meta-regression result



BCG Vaccine and TB

- The meta-regression analysis provided some clarity on the efficacy of BCG
- Widespread use of BCG was not adopted for the US
- Limited use recommended for
  - Infants and children exposed to untreated TB patients
  - Health care workers under specific conditions

Limitations of Meta-regression

- **Confounding** by related study characteristics is very possible
- For example, studies using higher doses of chemotherapy for breast cancer might also tend to have younger subjects than studies of lower dose chemotherapy
  - A trend in survival attributed to dose might really be due to age

Limitations of Meta-regression

- There may be a different association between the outcome and the covariate within studies than is seen between studies
- This is sometimes called the **ecologic fallacy**
  - The fallacy arises when a covariate is associated with the outcome at the study level (or at the level of different populations) but not at the individual level

Limitations of Meta-regression

- In general, we care about the individual level much more than the study level
  - Is someone at a higher latitude offered more or less protection from TB from the BCG vaccine?
- Unfortunately, if our data is summary measures from studies, we cannot investigate whether associations hold at the level of the individual
  - Big advantage for Individual Patient Data (IPD) meta-analysis

Within study trend with age is different that between study trend with age

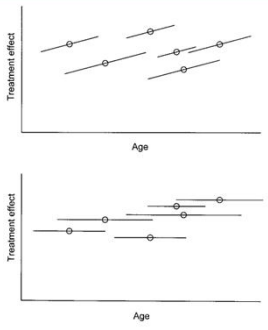


Figure 2. Hypothetical relationships between age and treatment effect both within trials (represented by the lines) and between trials (represented by the dots). See Section 3(ii) for explanation.

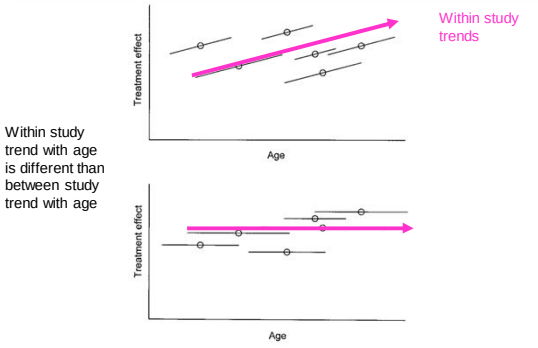


Figure 2. Hypothetical relationships between age and treatment effect both within trials (represented by the lines) and between trials (represented by the dots). See Section 3(ii) for explanation.

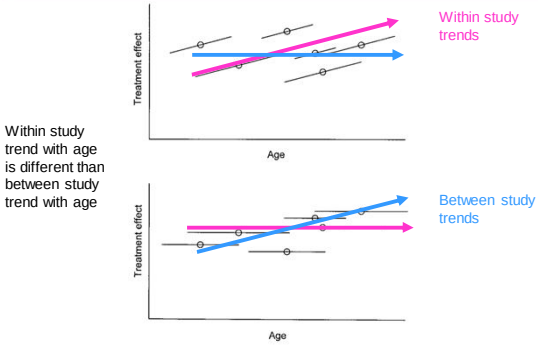


Figure 2. Hypothetical relationships between age and treatment effect both within trials (represented by the lines) and between trials (represented by the dots). See Section 3(ii) for explanation.

Limitations of Meta-regression

- Need to be careful of Data-dredging
  - Many characteristics differ between the studies in a meta-regression and usually there aren't many studies
  - So, the number of candidate predictors is large relative to the number of data points (studies)
  - This can lead to false positive results – that do not replicate on other data sets
- To minimize data-dredging, avoid post-hoc analyses and stick to a limited number of pre-specified covariates

Limitations of Meta-regression

- For these reasons, I think that the conclusions reached by meta-regression are even more tentative than those reached by meta-analysis
- On, the other hand it provides a useful tool to explore the available data and generate hypotheses

Threats to Validity of Meta-Analytic Results

- Heterogeneity
  - With heterogeneity there is extra variability in the study effect size estimates
- Bias
  - With bias there is a systematic mean difference between the combined effect size estimate and the true population effect size

Bias

- Various sources of bias
  - Selection of positive studies for publication
  - Choice of outcome toward a positive finding
- In meta-analysis biases push the combined effect **away** from the null
  - No worry about attenuation of effects
  - Not affected by increasing numbers of studies

Publication Bias

- The most noted form of bias is called **publication bias** and is due to the publication process
  - Studies reaching a statistically significant conclusion are more likely to be published than those that do not reach significance
  - Studies that reach significant conclusions are published more rapidly than studies that do not reach significance

Publication Bias

- Because studies that reach significance make up a greater proportion of published studies than they do of all studies
  - Published studies are more likely to be significant
  - Published studies tend to have larger effect sizes than unpublished studies
- Combining only published studies is likely to give a more impressive combined estimate than if all studies were used

Publication Bias

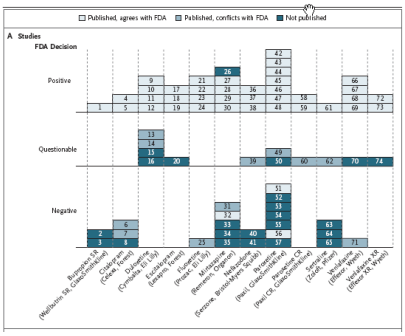
- Turner et al. looked at clinical trials submitted to the FDA in support of approval for marketing in the US
  - 12 Antidepressants
  - 74 clinical trials
- Pre-marketing approval documents available for FDA approved medications

Publication Bias

- Out of 74 trials, 23 (31%) were unpublished
- Publication status was strongly associated with the trial findings
  - Trials that the FDA viewed as positive had almost all been published
  - Half the trials the FDA viewed as questionable had been published, but presented as positive in publication
  - Two-thirds of the negative studies had not been published, and out of those published most were presented as positive

Publication Bias

Results from Turner et al.



Publication Bias

- As a consequence of
  - The lack of complete publication
  - Changes in analyses from the FDA results to the published results
- The summary effect size for the published studies was 32% greater than that of the FDA

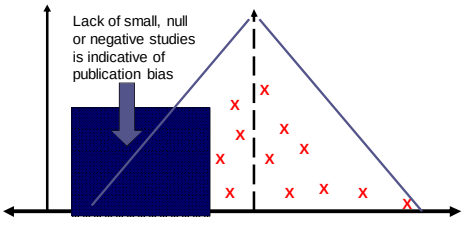
Outcome Selection Bias

- Often there is more than 1 possible outcome of interest for a study
- Bias is created when multiple outcomes are measured in a study and the outcome with the most positive result is presented as primary
- Often a secondary endpoint becomes the primary due to its statistical significance
  - Introduces an optimistic bias into the study results

Tools for Bias Assessment

- The main diagnostic tool for bias is to evaluate the **funnel plot** for ***asymmetry***
- Asymmetry in a funnel plot is usually attributed to **publication bias**, but other sources of bias or even heterogeneity can create asymmetry
- Asymmetry is suggests bias, but is not conclusive evidence

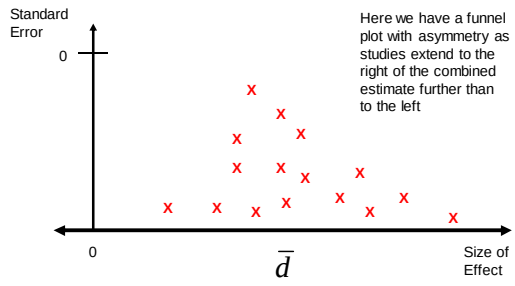
Funnel Plots



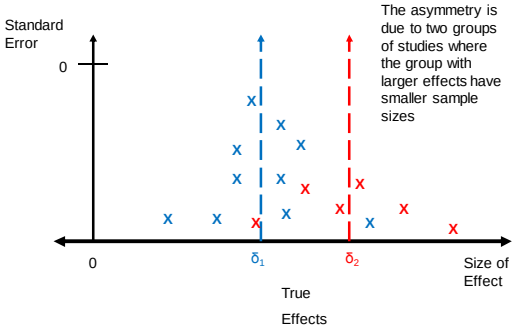
Tools for Bias Assessment

- There are specific issues with bias assessment for binary outcomes
  - Log odds ratio and log relative risk are associated with their standard errors
  - This introduces asymmetry even without bias
- Also bias assessment is made more challenging when there is heterogeneity
  - Heterogeneity can make the funnel plot asymmetric

Heterogeneity and Funnel Plots



Heterogeneity and Funnel Plots



Funnel Plots

- The funnel plot can be overlaid with contours corresponding to levels of statistical significance to see if publication bias could create the observed asymmetry
  - Studies absent in areas of non-significance would be expected with publication bias
  - If studies absent in areas of significance, that would suggest this is not due to publication bias

Contour Plots for Bias Evaluation

From Holt-Lundstad et al. PLoS Med 2010 7(7): e1000316

Looking at the effect of social relationships and mortality

Concluded not much effect of publication bias

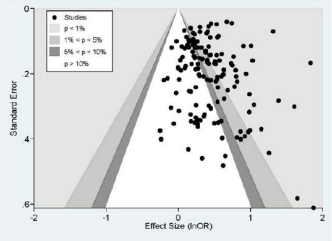


Figure 2. Contour enhanced funnel plot.
doi:10.1371/journal.pmed.1000316.g002

Testing Funnel Plot Asymmetry

- Tests have been developed to provide a p-value for the degree of asymmetry in the funnel plot
  - May be thrown off if heterogeneity is present
- 2 tests for symmetry of a funnel plot
  - The first uses *rank correlation* between the effect size and the study precision
  - The second is based on a *linear regression* of the effect size on the standard error

Testing Funnel Plot Asymmetry

- Both tests look for a **correlation** between effect size and standard error of the effect
- If studies are removed from the bottom left section of the funnel plot, that creates negative correlation between effect size and precision
- But, a plot with a full funnel shape will have a correlation of 0

Rank Correlation Test

- The removal of studies from the lower left of the funnel might not create a **linear** trend between effect size and variance, so a rank based method was developed by Begg and Mazumdar
- Rank-based procedures are generally considered to have low power to reject the null hypothesis
- If the number of studies included in a meta-analysis is not large, lower power can be a problem for this test

Linear Regression Test

- An alternative test due to Egger uses a weighted least-squares regression to test for lack of symmetry in the funnel plot
- Evaluates if there is a linear trend between the standard error and the effect size
- Usually has greater power to detect an association than a rank-based test

Funnel Plot for Binary Data

- Statistical tests assessing funnel plot asymmetry assume that association between effect size and precision is due to bias
- Need to use an effect size without this association for funnel plot tests on binary outcome data
- Harbord's test modifies the linear regression test for use with binary outcomes

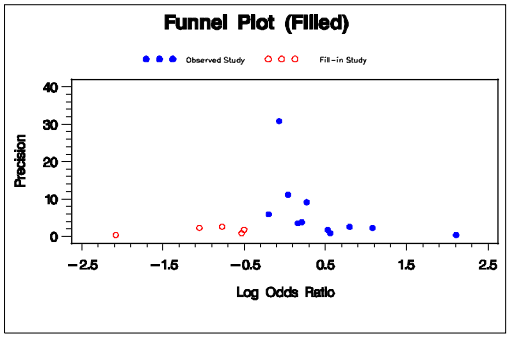
Level of Significance for Bias Tests

- For bias tests, using a significance level of 0.10 is suggested
- To accommodate low power to detect a departure from symmetry of funnel plot

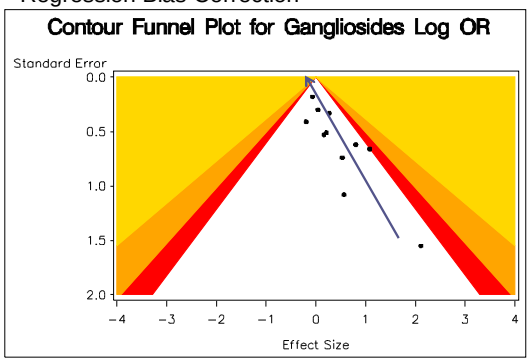
Bias Correction

- Recently methods have been developed to provide estimates of what the combined effect would be if there were no bias
- Trim-and-Fill takes the approach of imputing studies to make the funnel plot symmetric
- Regression methods are used to extrapolate to what the combined effect would be with a standard error of zero

Trim and Fill



Regression Bias Correction



Bias Correction

- I consider bias correction methods as sensitivity analyses and not primary analyses
- They provide a good sense of how big an impact bias may have on the combined estimate, but neither is ironclad
  - Trim and fill imputes studies for which there is no direct evidence
  - Regression bias correction extrapolates beyond the range of the data

Quality Scoring

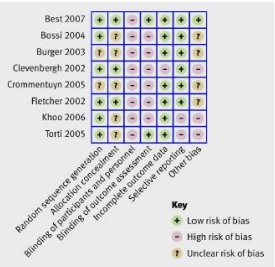
- The idea is that some primary studies are conducted in a manner that make the results more trustworthy than those of studies that are of similar size but performed in a less rigorous manner
- The quality score assess the **methods** used in the primary studies and should reflect how well the study was conducted

Quality of Randomized Trials

- The most up-to-date tool for evaluation of quality of randomized controlled trials is the Cochrane Risk of Bias tool
- This is focused on the internal validity of the treatment comparison and only considers items related to bias
- Specifically choose not to develop a numeric scaling of the quality and use a graphical approach

Quality of Randomized Trials

Presentation of risk of bias of studies in Cochrane review of therapeutic monitoring of anti-retroviral drugs in persons with HIV

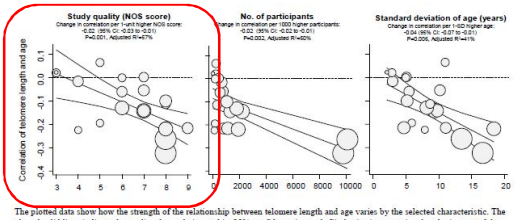


From Higgins et al., 2011

Quality of Observational Studies

- For observational studies the Newcastle-Ottawa scoring tool is the most up-to-date tool
- Is set up to provide a numeric scale
- Evaluates bias control in the observational results

Supplementary Figure 3 | Correlation between telomere length and age as a function of selected characteristics.



The plotted data show how the strength of the relationship between telomere length and age varies by the selected characteristic. The plotted solid lines indicate the predicted correlation and its 95% confidence interval. Circle size is proportional to the inverse of the within-study variance. CI – confidence interval, NOS, Newcastle-Ottawa Scale, SD – standard deviation. Study quality, the no. of participants, and standard deviation of age collectively explained 30% of the between-study heterogeneity in reported correlations between telomere length and age. The correlation between telomere length and age was: -0.22 (95% CI: -0.27 to -0.16) in studies with a NOS score ≥8; -0.20 (95% CI: -0.25 to -0.14) in studies with >1,000 participants; and -0.16 (95% CI: -0.22 to -0.10) in studies with an standard deviation of age ≥8 years.

Quality Scoring

- To account for study quality in meta-analysis, several proposals have been made
  - Remove studies below a quality threshold from the meta-analysis
  - Stratify the analysis by quality score
  - Perform stratified analyses on selected quality components
  - Meta-regression

Validity of Meta-Analysis

- When people discuss the validity of meta-analysis they are considering its validity as a research design
  - Does the method support the conclusions that we try to reach using meta-analysis?
- Several objections have been raised over the years on the validity of meta-analysis as a research method

Validity of Meta-Analysis

- Sharpe lists 3 main objections to the validity of meta-analysis
  - Apples and oranges – meta-analysis combines the results from studies that measure different things
  - File drawer – due to selective publication of positive results, published research is a biased subset of the population of research
  - Garbage in, garbage out – combining studies with weak methods in with studies using better methods lead to a distorted picture of the effectiveness of a therapy

Apples and Oranges

- There are at least three counter arguments to the apples and oranges objection
  - Carefully considering in advance the treatments or exposures, methods, and outcomes that will be included
  - Use of heterogeneity analysis to see if there is between study variability
  - With enough studies, meta-regression can be done to identify subgroups of studies

File Drawer

- There are techniques for the assessment of publication and other biases
  - Visual inspection of funnel plots
  - Egger test
  - Rank-correlation method
  - Trim-and-fill adjustment for bias
  - Regression adjustment for bias
- These techniques are not perfect but allow us to probe the set of study results for indications of bias

Garbage In, Garbage Out

- The garbage-in, garbage out objection can be addressed to some degree by
  - Giving details on the studies used in the analysis and allowing the reader to judge their worth
  - Cochrane risk of bias tool

Comparison to Large Clinical Trials

- Another validity argument against meta-analysis comes from a widely-cited paper by LeLorier et al.
- This article made a big impact in 1997 when it was published in the NEJM along with an editorial by statistician John Bailar III stating

> *"Other observers, including policy makers, also have reservations about meta-analyses, and there is some general concern about the credibility of the findings of meta-analysis. I know of no instance in medicine in which a meta-analysis led to a major change in policy before the time when a careful, conventional review of the literature led to the same change."*

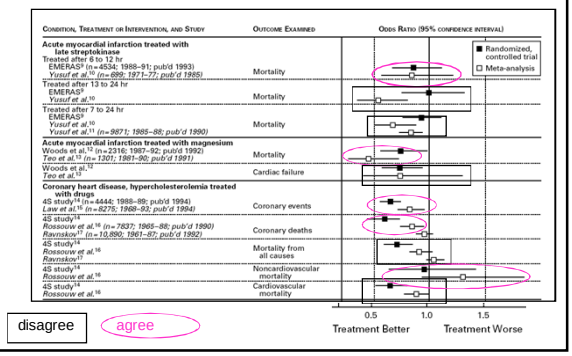
Comparison to Large Clinical Trials

- The LeLorier paper compared the results of meta-analyses to those of subsequent large clinical trials
  - They considered large clinical trials (at least 1000 subjects) as the gold standard for clinical investigation
  - They evaluated agreement in results between the published meta-analyses and subsequent large clinical trials

Comparison to Large Clinical Trials

- When did they consider the meta-analysis and the following large trial to be in agreement?
  - Both positive
    - Treatment better and $p < 0.05$
  - Both negative
    - Treatment same as control or worse than control

Comparison to Large Clinical Trials



Comparison to Large Clinical Trials

TABLE 1. AGREEMENT OR DISAGREEMENT BETWEEN RANDOMIZED, CONTROLLED TRIALS AND META-ANALYSES IN 40 CASES IN WHICH THE TWO WERE COMPATIBLE.*

RESULTS OF META-ANALYSIS	RESULTS OF RANDOMIZED, CONTROLLED TRIAL		
	POSITIVE	NEGATIVE	TOTAL
Positive	13	6	19
Negative	7	14	21
Total	20	20	40

*Positive indicates that the outcome of treatment was significantly better ($P<0.05$) than the outcome of no treatment, and negative indicates that the outcome of treatment was worse or the same.

13 Mismatches out of 40 pairs (35%)

Comparison to Large Clinical Trials

- LeLorier et al. concluded
  "The outcomes of the 12 large randomized, controlled trials that we studied were not predicted accurately 35 percent of the time by meta-analyses published previously on the same topics."
- So are the results of meta-analysis not valid?

Comparison to Large Clinical Trials

- There are some statistical limitations to the LeLorier study
  - They treated the 40 comparisons as statistically independent
  - Since they were based on 12 trials and 19 meta-analyses they weren't independent

Comparison to Large Clinical Trials

- Probably the biggest problem is the way they looked at agreement
  - Incorporating significance as a criterion for agreement is just a bad idea
  - Casual readers may use $p<0.05$ as a shorthand for "success of treatment", but that is a very limited view
  - Using the confidence interval would be more a more reasonable approach

Comparison to Large Clinical Trials

- Follow-up articles have examined the LeLorier agreement criteria
  - Furukawa et al. J Clin Epi (2000) found 33% mismatch between large trials and other large trials on the same topic
  - So, are large trials invalid too?

Comparison to Large Clinical Trials

- This way of looking at agreement between clinical studies is flawed
- Large clinical trials are the best way we have to assess treatment efficacy, but they are not an infallible *gold standard*
  - They don't always agree with each other
  - Clinical inquiry is never foolproof
  - One shouldn't assume that the conclusions of large well-conducted clinical trials are always true

Conclusions

- Meta-analysis is still a fairly young area of statistics
  - Most of these methods only date back to the 1980s
  - New approaches and methods are currently being developed at a fast rate
- It has real limitations due to it being an observational and retrospective study design
  - I get nervous when people present meta-analytic results as 'the truth'

Conclusions

- But, I think that a well done meta-analysis provides a meaningful summary of the research on a topic
- And it allows us to probe the limits and quality of the published literature on the topic
  - Evaluate heterogeneity
  - Evaluate bias
- It is a unique tool for this purpose