Evidence-Based Medicine and Statistics

Introductory Overview Lecture

JSM 2011

Christopher Schmid
Tufts University
1 August 2011
Evidence-based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients

Healthcare Spending and Quality

<table>
<thead>
<tr>
<th>Difference Between Actual And Expected Health Care Spending Per Capita And Actual And Expected Life Expectancy In Organization For Economic Cooperation And Development (OECD) Countries, 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference in actual and expected life expectancy (years)</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>3.0</td>
</tr>
<tr>
<td>2.0</td>
</tr>
<tr>
<td>2.0</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>-1.0</td>
</tr>
<tr>
<td>-2.0</td>
</tr>
<tr>
<td>-3.0</td>
</tr>
<tr>
<td>-4.0</td>
</tr>
<tr>
<td>-1,000</td>
</tr>
<tr>
<td>Difference in actual and expected health care spending per capita ($ PPP)</td>
</tr>
</tbody>
</table>


**NOTES:** Regression equation for expected health spending is $y = 0.1174x - 706.35$ with $R^2 = 0.79$, where $y$ is health care spending per capita ($$ purchasing power parity, or PPP) in 2005 and $x$ is gross domestic product (GDP) per capita ($$ PPP) in 2005. Regression equation for expected life expectancy is $y = 0.0002x - 72.503$ with $R^2 = 0.57$, where $y$ is life expectancy in years in 2005 and $x$ is GDP per capita ($$ PPP) in 2005. For details, see Notes 15, 16, and 18 in text. For Australia, Hungary, Japan, and the Netherlands, health spending data for 2004 are used. For Canada and the United States, life expectancy data for 2004 are used. Country abbreviations are spelled out in Exhibit 2. Luxembourg (LX) is omitted from this analysis.

- $2,197 per capita more than expected
- 3.1 life years less than expected
How health care varies by region
Percentage of Acute Otitis Media Patients Given Antibiotics

- U.S.: 97.9%
- Switzerland: 91.2%
- New Zealand: 98.2%
- Netherlands: 31.2%
- Israel: 91.4%
- Great Britain: 96.8%
- Belgium: 85.1%
- Australia: 98.2%

Figure taken from Froom J et al. Diagnosis and antibiotic treatment of acute otitis media: report from International Primary Care Network. BMJ 1990;300:582-6.
Tradition-based Medicine

• Emphasizes
  ─ primacy of knowledge
  ─ experience
  ─ intuition in exercising good clinical judgment

• Observational

• Susceptible to bias

• Individual experiences limited and problems heterogeneous

• Lack of conceptual framework for synthesizing evidence

• Lack of conceptual framework for clinical decision making
Evidence-Based Medicine

Stresses
• examination of evidence from clinical research
• systematic collection of evidence
• synthesis of evidence

De-emphasizes
• intuition
• unsystematic experience
• pathophysiological rationale (surrogates)
Broad View of Clinical Research

• Improve health outcomes of individual patients and society

• Translate (basic) science discoveries into clinical practice

• Optimize use and delivery of healthcare technologies in society

• Provide information to guide
  – Patient management
  – Individual decision making
  – Policy decision making
    • Public health
    • Reimbursement
  – Research agenda of funding agencies
Limitations of Current Best Evidence

Little evidence about which treatments work best for which patients

- Summary results
- Trial and study exclusions
- Poor comparators

Little evidence about whether the benefits of more expensive therapies warrant their additional costs

- Few RCTs include a cost study
- Poor data
- Skepticism about cost effectiveness analysis, simulation and other decision analysis methods that incorporate cost information
Hierarchy of Evidence

- Systematic Reviews
- Critically-Appraised Topics [Evidence Syntheses]
- Critically-Appraised Individual Articles [Article Synopses]
- Randomized Controlled Trials (RCTs)
- Cohort Studies
- Case-Controlled Studies, Case Series / Reports
- Background Information / Expert Opinion

TRIP Database searches these simultaneously

UNFILTERED INFORMATION

FILTERED INFORMATION
Institute of Medicine Definition

CER is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels.
### Evidence-Based Science

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;evidence-based&quot;</td>
<td>35,200,000</td>
</tr>
<tr>
<td>&quot;evidence-based medicine&quot;</td>
<td>1,880,000</td>
</tr>
<tr>
<td>&quot;evidence-based practice&quot;</td>
<td>1,390,000</td>
</tr>
<tr>
<td>&quot;evidence-based nursing&quot;</td>
<td>525,000</td>
</tr>
<tr>
<td>&quot;evidence-based healthcare&quot;</td>
<td>374,000</td>
</tr>
<tr>
<td>&quot;evidence-based mental health&quot;</td>
<td>168,000</td>
</tr>
<tr>
<td>&quot;evidence-based nutrition&quot;</td>
<td>467,000</td>
</tr>
<tr>
<td>&quot;evidence-based dentistry&quot;</td>
<td>156,000</td>
</tr>
<tr>
<td>&quot;evidence-based pediatrics&quot;</td>
<td>33,900</td>
</tr>
<tr>
<td>&quot;evidence-based surgery&quot;</td>
<td>33,700</td>
</tr>
<tr>
<td>&quot;evidence-based veterinary medicine&quot;</td>
<td>362,000</td>
</tr>
<tr>
<td>&quot;evidence-based management&quot;</td>
<td>4,280,000</td>
</tr>
<tr>
<td>&quot;evidence-based social&quot;</td>
<td>2,200,000</td>
</tr>
<tr>
<td>&quot;evidence-based education&quot;</td>
<td>66,800</td>
</tr>
<tr>
<td>&quot;evidence-based marketing&quot;</td>
<td>1,270,000</td>
</tr>
<tr>
<td>&quot;evidence-based politics&quot;</td>
<td>44,100</td>
</tr>
<tr>
<td>&quot;clinical practice guideline&quot;</td>
<td>867,000</td>
</tr>
<tr>
<td>&quot;systematic review&quot;</td>
<td>1,970,000</td>
</tr>
<tr>
<td>&quot;meta-analysis&quot;</td>
<td>3,880,000</td>
</tr>
</tbody>
</table>

Source: Google – December 12, 2010
Evidence-Based Medicine

1) Systematic Reviews and Meta-Analyses
2) Randomized Controlled Clinical Trials
3) Observational Studies
4) Case reports

Special case: decision modeling, including simulations and cost effectiveness analysis
Patient Centered Outcomes Research Institute (PCORI)

• Independent agency outside US Government

• Roles and responsibilities
  – Set research priorities
  – Determine project agenda and methods to be used
  – Award contracts with preference to NIH and AHRQ
  – Appoint expert advisory panels
  – Develop methods and methods standards
  – Conduct peer review
  – Disseminate research findings
PCORI Governing Board

- AHRQ Director
- NIH Director
- 19 Stakeholders – clinicians, patients, researchers, consumers
In the late 18th century, King Gustav III of Sweden decided that coffee was poison and ordered a clinical trial.

**Intervention**: Convicted murderer to drink coffee daily

**Control**: Another murderer to drink tea daily

**Outcome**: Death

**Outcome Assessment**: 2 physicians to determine outcome
Results

• Two doctors died first

• King was murdered

• Both convicts enjoyed long life until tea drinker died at age 83 (Age of coffee drinker not reported)
Discussion

• One should not rely on such a small sample size
• Perhaps the end point was too hard
• Outcome of trial had no effect on decision makers
• Coffee was forbidden in Sweden in 1794 and again in 1822
Conclusions

• None possible regarding the effect of coffee
• External events and other biases may have confounded result
• Kings shouldn’t mess with clinical trials
Randomized Clinical Trials (RCTs)

- Use random treatment assignment to determine efficacy of intervention under ideal circumstances

- Patients are randomly assigned to treatment or control groups with pre- and post-treatment measurement, double blinding and closely followed treatment protocols

- 1993 conference reviewing quality of publications reporting clinical trials found considerable variation in quality and issued new standard for measuring quality of RCT reports
CONSORT Statement
(Consolidated Standards of Reporting Trials)

• Checklist for reporting of 25 items:
  – Title and Abstract
  – Scientific background and rationale
  – Methods
  – Results
  – Discussion

• Flow diagram to describe patient flows through enrollment, intervention allocation, follow-up and data analysis
Advantages of RCTs

• A priori hypothesis
• Internal validity if randomized and controlled
• Near-certain test of efficacy of intervention vs. placebo

Well-designed clinical trials excel at testing an a priori causal hypotheses, typically comparing the effect of an intervention against placebo, for an ideal population, in a controlled setting.
Limitations of RCTs

- Limited external validity
- Uncertain effectiveness of the intervention
- Uncertain comparison to alternatives
- Difficult to apply summary results to individual patients
- RCTs are often very slow to produce results

Even well-designed trials may not be very good at determining the effects of an intervention, compared to existing alternatives under the usual conditions in which they are be applied.
Pragmatic Trials

Practical or pragmatic trial designed to determine effects of intervention under usual conditions in which it will be applied.
ClinicalTrials.Gov

- Website to register RCT protocols and results
- Required by many journals, US funding agencies and FDA
- May reduce problems of publication and reporting bias
Publication Bias

• Negative studies are more likely than positive studies to remain unpublished

• Negative” studies are likely to be small

• In general, not concerned about unpublished “positive” studies.

• Negative studies might invalidate meta-analysis results

• Publication bias is only a part of the bigger “missing data” problem in meta-analysis (and clinical research)

• Selective reporting bias may be a bigger problem
Case Reports (Case Series)

• Detailed report of diagnosis, treatment, and follow-up of individual patient

• Contain some demographic information about patient

Advantages
  – Helpful in medical education to describe unusual occurrences
  – Development of clinical judgement

Limitations
  – Anecdotal evidence
  – Limited (to no) generalizibility
Observational Studies

**Case-control and cohort designs** typically use existing population data, a hypothesis and statistical controls to evaluate a problem or identify associations between an “intervention” and an “outcome”

**Other non-randomized designs**
- Cross sectional studies
- Surveillance studies using registry data

**Advantages**
- For retrospective approaches, readily available data
- Faster results
- Hypothesis-generating

**Limitations**
- Confounding
- Limited causal inference
- Limited external validity (often, not always)
Major Impacts of Non-randomized Evidence

- Lind, 1747, 6 pairs of sailors with scurvy
- Jenner, smallpox, late 18th century
- Fleming, penicillin, 1928-1940s
Observational Study Findings Later Disproved

- Hormone therapy / cardio-protective effects of estrogen

- β carotene and α-tocopherol and cancer

- Fiber and colon cancer
Major Impacts of Randomized Evidence

- Streptomycin for tuberculosis
- Polio vaccine
- Treatments for acute myocardial infarction
- Estrogen Replacement Therapy
Observational vs. Randomized Evidence

• Treatment effects in RCTs and observational studies on same topic tend to be highly correlated
  – Discrepancies occur in about 1 out of 6 cases, even when accounting for between-study heterogeneity
  – Discrepant pairs tend to show more favorable results in observational studies

• Discrepancies in magnitude of effect very common
• Observational studies exhibit larger variability in treatment effects than RCTs
• Discrepancies more common with retrospective designs
Systematic Review

• Scientific discipline to combine information across studies using defined protocol to answer focused research question(s)

• Formulate well-focused study question

• Establish eligibility criteria (study, patient, and disease characteristics, intervention, comparator, outcomes)

• Review literature comprehensively

• Identify relevant studies

• Extract data

• Critically appraise study quality and conclusions
Meta-Analysis

- Quantitative analysis of data from systematic review
- Estimate effect size and uncertainty (treatment effect, association, test accuracy) by statistical methods
- Combine “under-powered” studies to give more definitive conclusion
- Explore heterogeneity / explain discrepancies
- Identify research gaps and need for future studies
Systematic Reviews and Meta-Analyses

Advantages

• Resolve inconsistent studies
• Guide clinical research w/ new hypotheses
• Identify effects earlier through cumulative analysis

Limitations

• Difficult to identify all relevant studies (limitations of electronic searches + publication bias)
• Difficult to judge the quality of all identified studies
• Difficult to apply summary results to individual patients
• Difficult to account for between-trial differences
### Comparing systematic reviews with narrative “non-systematic” reviews

<table>
<thead>
<tr>
<th><strong>Narrative Reviews</strong></th>
<th><strong>Systematic Reviews</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Give panoramic view, usually cover whole topic. Example: textbook chapters</td>
<td>Give telescopic view, usually address one question or a few questions</td>
</tr>
<tr>
<td>Emphasize “background” knowledge: What causes the disorder? What are the clinical manifestations? What treatment options are available?</td>
<td>Focus on “foreground” knowledge: For example, in treating patients with this disorder, which of the two available treatments is better at improving clinical outcomes safely?</td>
</tr>
<tr>
<td>Susceptible to bias in selecting, appraising and combining studies to answer questions</td>
<td>Use rigorous methods to minimize bias and help improve reliability and accuracy of conclusions</td>
</tr>
<tr>
<td></td>
<td>Can provide pooled estimates of treatment benefits and risks</td>
</tr>
</tbody>
</table>
Applying SR and MA in Healthcare

- **Interventions** (most common) estimate efficacies and harms of treatments
- **Epidemiologic** (many) to provide more reliable estimates of risks, associations
- **Diagnostic tests** (increasing) provide more reliable estimates of diagnostic accuracy of tests
- **Genomics** (rapidly increasing) estimate effects of microarray and GWAS studies
- **Health economics**
REPORT ON CERTAIN ENTERIC FEVER INOCULATION STATISTICS.

Provided by Lieutenant-Colonel R. J. S. Simpson, C.M.G., R.A.M.C.

By Karl Pearson, F.R.S.,
Professor of Applied Mathematics, University College, London.

The statistics in question were of two classes: (a) Incidence. (b) Mortality Statistics. Under each of these headings the data belonged to two groups: (i) Indian experience; (ii) South African War experience. These two experiences were of a somewhat different character. That for India covered apparently the European army, of whatever branch and wherever distributed; that for South Africa was given partly by locality, partly by column, and partly by special hospital. Thus the Indian and South African experiences seem hardly comparable. Many of the groups in the South African experience are far too small to allow of any definite opinion being formed at all, having regard to the size of the probable error involved. Accordingly, it was needful to group them into larger series. Even thus the material appears to be so heterogeneous, and the results so irregular, that it must be doubtful how much weight is to be attributed to the different results.
Systematic Review Products

- Journal publications
- Evidence reports
- Comparative effectiveness reviews (CER)
- Technology assessments
- Horizon scans
- Future research needs documents
- Feeders into clinical practice guidelines, coverage, and policy decision making
About 20 studies with usable primary data for pediatric population

450 reports on complication of sinusitis

233 narrative reviews
The Cochrane Collaboration

- International collaboration to promote research synthesis
- National centers (one in USA)
- Collaborative review groups organized by clinical area
- Over 2000 meta-analyses published
- Also has register of randomized controlled trials
PRISMA Statement

Checklist of 27 topics to present in Systematic Reviews

1) Background and Methods
2) Data Collection
3) Analysis Plan
4) Results
5) Summary
6) Synthesis
7) Conclusions
Institute of Medicine (IOM) Standards for Systematic Reviews

For more information about the report go to www.iom.edu/srstandards or www.nap.edu
STEPS OF PERFORMING A META-ANALYSIS

1. Formulate Study Question
2. Establish Protocol
3. Literature Search / Retrieval
4. Paper Selection per Protocol
5. Data Extraction
6. Critical Appraisal
7. Analysis and Interpretation
Formulating Answerable SR Questions

- Who is SR for and how will results be interpreted and used?
- Narrow versus broad question (e.g., for individual or population)
- Clinically meaningful and useful (based on sound biological and epidemiological principles)
- Very broadly defined questions may be criticized for mixing apples and oranges
- Very narrowly focused questions may have no data or have limited generalizability and sometimes may lead to misinterpretations
- Include stakeholders, clinicians, methodologists
PICO(TS) Formulation

• Population
• Interventions
• Comparators
• Outcomes
• Timing
• Study design

• Eligibility criteria
Example: The Well - Formulated Question

The Cochrane Collaboration "How to Conduct a Cochrane Systematic Review" 1996

Does drug therapy decrease long-term morbidity and mortality in older persons with mild to moderate hypertension?

**Intervention**
- ACE inhibitors
- Angiotensin Receptor Antagonists
- Combined Alpha and Beta Blockers
- Calcium-Channel Blockers
- Diuretics
- Alpha Adrenergic Blockers
- Central Sympatholytics
- Direct Vasodilators
- Peripheral Adrenergic Antagonist

**Outcomes**
- Fatal and non-fatal strokes
- Fatal and non-fatal Coronary Heart Disease (MI, sudden death)
- Cardiovascular events (above plus aneurysm, congestive heart failure, transient ischemic attacks)
- Total Mortality

**Population Setting**
- > 1 year
- > 60 yrs old outpatients
- Systolic 140-179
- Diastolic 90-109

**Condition of interest**
- > 1 year
- > 60 yrs old outpatients
- Systolic 140-179
- Diastolic 90-109
Identifying the Literature

• Guided by key questions and eligibility criteria

• Comprehensive but practical
  – Search multiple databases
  – Balance between feasibility, resources, and needs

• Minimize selection bias
  – Language: English only?
  – Include unpublished studies?
  – Multiple (overlapping) publications of same data

• Minimize errors

• Often iterative process with question formulation
18,000 citations were screened for the cancer pain evidence report
Principles of Data Extraction

• Extract data needed to survey literature
• Extract data needed to critically appraise study
• Extract data needed to conduct meta-analyses
• Take steps to minimize data extraction errors
  – Data extraction requires methods and domain knowledge
  – Create and test data collection form
  – Train and calibrate data extractors
  – Perform double independent data extraction or extract by one and verify by another
Some Data Extraction Problems

• Data reporting errors
• Non-uniform outcomes (different measurements in different studies)
• Incomplete data (frequent problem: no standard error or confidence interval)
• Discrepant data (different parts of same report gave different numbers)
• Confusing data (can’t figure out what authors reported)
• Non-numeric format (reported as graphs)
• Missing data (only conclusion reported)
• Multiple (overlapping) publications of same study
Example of Data Reporting Problem

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Demographic Characteristics of Clinically Evaluable Patients and Overall Description of Pathologies Treated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment Group</td>
</tr>
<tr>
<td></td>
<td>Roxithromycin</td>
</tr>
<tr>
<td></td>
<td>Clarithromycin</td>
</tr>
<tr>
<td>No. of patients</td>
<td>100</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>43</td>
</tr>
<tr>
<td>Male</td>
<td>57</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>39.30</td>
</tr>
<tr>
<td>Range</td>
<td>47–92</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>66.42</td>
</tr>
<tr>
<td>Range</td>
<td>47–92</td>
</tr>
</tbody>
</table>
Another Example of Data Reporting Problem

Data for the 40 patients who were given all four doses of medications were considered evaluable for efficacy and safety. The overall study population consisted of ten (44%) men and 24 (56%) women, with a racial composition of 38 (88%) whites and five (12%) blacks.
Rationale for Quality Appraisal

- Assess risk of bias and potential effect on conclusions
- Set threshold for inclusion and exclusion of studies in review
  - Use in sensitivity analysis (test robustness)
- Potentially explain differences in results between trials
- Weight statistical analysis of results
  - Quality scores not recommended
- Establish strength of recommendation in guidelines
- But poor reporting may be mistaken for poor quality
Commonly Assessed Quality Features

- Allocation concealment
- Blinding
- Description of intervention
- Withdrawals
- Statistical analysis
- Accuracy of reporting
Types of Data to Combine

- Dichotomous (events, e.g. deaths)
- Measures (odds ratios, correlations)
- Continuous data (mmHg, pain scores)
- Effect size
- Survival curves
- Diagnostic test (sensitivity, specificity)
- Individual patient data
Effect Size

- Dimensionless metric

- Basic idea is to combine standard deviations of diverse types of related effects

- However, availability and selection of reported effects may be biased, variable importance of different effects

- Frequently used in education, social science literature

- Infrequently used in medicine, difficulty in interpreting results
# What is the average difference in DBP?

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Size</th>
<th>Δ mmHg</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANBP</td>
<td>554</td>
<td>-6.2</td>
<td>-6.9 to -5.5</td>
</tr>
<tr>
<td>EWPHE</td>
<td>304</td>
<td>-7.7</td>
<td>-10.2 to -5.2</td>
</tr>
<tr>
<td>Kuramoto</td>
<td>39</td>
<td>-0.1</td>
<td>-6.5 to 6.3</td>
</tr>
</tbody>
</table>
Simple Average

\[
\frac{0.1-(+7.7-(+6.2)}{3} = -4.7 \text{ mmHg}
\]

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Size</th>
<th>Δ mmHg</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANBP</td>
<td>554</td>
<td>-6.2</td>
<td>-6.9 to -5.5</td>
</tr>
<tr>
<td>EWPHE</td>
<td>304</td>
<td>-7.7</td>
<td>-10.2 to -5.2</td>
</tr>
<tr>
<td>Kuramoto</td>
<td>39</td>
<td>-0.1</td>
<td>-6.5 to 6.3</td>
</tr>
</tbody>
</table>
Average Weighted by Sample Size

\[(554 \times -6.2) + (304 \times -7.7) + (39 \times -0.1)\]
\[\frac{554 + 304 + 39}{554 + 304 + 39} = -6.4 \text{ mmHg}\]

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Size</th>
<th>Δ mmHg</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANBP</td>
<td>554</td>
<td>-6.2</td>
<td>-6.9 to -5.5</td>
</tr>
<tr>
<td>EWPHE</td>
<td>304</td>
<td>-7.7</td>
<td>-10.2 to -5.2</td>
</tr>
<tr>
<td>Kuramoto</td>
<td>39</td>
<td>-0.1</td>
<td>-6.5 to 6.3</td>
</tr>
</tbody>
</table>
Forest Plot: Influenza Vaccine Efficacy
Heterogeneity (diversity)

• Is it reasonable (are studies and effects sufficiently similar) to estimate an average effect?

• Types of heterogeneity
  – **Conceptual** (clinical) heterogeneity: Are studies of similar treatments, populations, settings, design, etc., such that an average effect would be clinically meaningful?

  – **Statistical** heterogeneity: Is observed variability of effects greater than that expected by chance alone?
FIXED EFFECTS MODEL

TREATMENT EFFECTS (RD, OR, RR)

SINGLE TRUE TREATMENT EFFECT
TREATMENT EFFECTS (RD, OR, RR)

FIXED EFFECTS MODEL

POOLED RESULT
ESTIMATED
TREATMENT
EFFECT

SINGLE
TRUE
TREATMENT
EFFECT

RESULTS OF
MULTIPLE CLINICAL
TRIALS RANDOMLY
DISTRIBUTED
AROUND THE TRUE
TREATMENT EFFECT

TREATMENT EFFECTS (RD, OR, RR)
RANDOM EFFECTS MODEL

MULTIPLE TRUE TREATMENT EFFECTS (distribution of treatment effects)
RANDOM EFFECTS MODEL

TREATMENT EFFECTS (RD, OR, RR)

MULTIPLE TRUE TREATMENT EFFECTS
(distribution of treatment effects)

RESULTS OF MULTIPLE CLINICAL TRIALS
RANDOMLY DISTRIBUTED AROUND EACH OF THE TRUE TREATMENT EFFECT

POOLED RESULT
SINGLE ESTIMATED TREATMENT EFFECT
General Formula - Weighted Average Effect Size

$$d_+ = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i=1}^{k} w_i}$$

- $d_i = \text{effect size of study i}$
- $w_i = \text{weight of study i}$
- $k = \text{number of studies}$
- $s_i = \text{within study variance}$
- $\tau^2 = \text{between study variance}$

**Fixed Effect Weight**
$$W_i = 1/s_i$$

**Random Effect Weight**
$$W_i = 1/[s_i + \tau^2]$$
Fixed Effects Meta-Analysis

Trial

1 2 3 4 5 6 7 8 9 10 11 12

-1 0 1

common (fixed) effect

random error

Effect estimate

Treatment better  Control better
Random Effects Meta-Analysis

Trial
1 2 3 4 5 6 7 8 9 10 11 12

study-specific effect
random error

distribution of effects

Effect estimate
Treatment better ← → Control better

-1 Θ 0 τ 1
Comparison: 13 Magnesium vs placebo
Outcome: 03 Mortality

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment n/N</th>
<th>Control n/N</th>
<th>OR (95%CI Fixed)</th>
<th>Weight %</th>
<th>OR (95%CI Fixed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abraham</td>
<td>1 / 48</td>
<td>1 / 46</td>
<td></td>
<td>0.0</td>
<td>0.96[0.06,15.77]</td>
</tr>
<tr>
<td>Bertschat</td>
<td>0 / 22</td>
<td>1 / 21</td>
<td></td>
<td>0.1</td>
<td>0.30[0.01,7.88]</td>
</tr>
<tr>
<td>Morton</td>
<td>1 / 40</td>
<td>2 / 36</td>
<td></td>
<td>0.1</td>
<td>0.44[0.04,5.02]</td>
</tr>
<tr>
<td>Ceremuzynski</td>
<td>1 / 25</td>
<td>3 / 23</td>
<td></td>
<td>0.1</td>
<td>0.28[0.03,2.88]</td>
</tr>
<tr>
<td>Pereira</td>
<td>1 / 27</td>
<td>7 / 27</td>
<td></td>
<td>0.3</td>
<td>0.11[0.01,0.97]</td>
</tr>
<tr>
<td>Smith</td>
<td>2 / 200</td>
<td>7 / 200</td>
<td></td>
<td>0.3</td>
<td>0.28[0.06,1.36]</td>
</tr>
<tr>
<td>Feldstedt</td>
<td>10 / 150</td>
<td>8 / 148</td>
<td></td>
<td>0.3</td>
<td>1.25[0.48,3.26]</td>
</tr>
<tr>
<td>Thogersen</td>
<td>4 / 130</td>
<td>8 / 122</td>
<td></td>
<td>0.4</td>
<td>0.45[0.13,1.54]</td>
</tr>
<tr>
<td>Golf</td>
<td>5 / 23</td>
<td>13 / 33</td>
<td></td>
<td>0.4</td>
<td>0.43[0.13,1.44]</td>
</tr>
<tr>
<td>Shechter 90</td>
<td>1 / 59</td>
<td>9 / 56</td>
<td></td>
<td>0.4</td>
<td>0.09[0.01,0.74]</td>
</tr>
<tr>
<td>Singh</td>
<td>6 / 76</td>
<td>11 / 75</td>
<td></td>
<td>0.5</td>
<td>0.50[0.17,1.43]</td>
</tr>
<tr>
<td>Shechter 95</td>
<td>4 / 107</td>
<td>17 / 108</td>
<td></td>
<td>0.8</td>
<td>0.21[0.07,0.64]</td>
</tr>
<tr>
<td>Rasmussen</td>
<td>9 / 135</td>
<td>23 / 135</td>
<td></td>
<td>1.0</td>
<td>0.35[0.15,0.78]</td>
</tr>
<tr>
<td>LIMIT-2</td>
<td>90 / 1159</td>
<td>118 / 1157</td>
<td></td>
<td>5.1</td>
<td>0.74[0.56,0.99]</td>
</tr>
<tr>
<td>ISIS-4</td>
<td>2216 / 29011</td>
<td>2103 / 29039</td>
<td></td>
<td>90.2</td>
<td>1.06[1.00,1.13]</td>
</tr>
</tbody>
</table>

Total(95%CI) 2351 / 31212 | 2331 / 31226
Weight % 100.0
OR 1.01[0.95,1.07]

Test for heterogeneity chi-square=40.18 df=14 p=0.0002
Test for overall effect z=0.36 p=0.7
**Comparison: 13 Magnesium vs placebo**

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment n/N</th>
<th>Control n/N</th>
<th>OR (95%CI Random)</th>
<th>Weight %</th>
<th>OR (95%CI Random)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abraham</td>
<td>1 / 48</td>
<td>1 / 46</td>
<td>1.6</td>
<td></td>
<td>0.96[0.06,15.77]</td>
</tr>
<tr>
<td>Bertschat</td>
<td>0 / 22</td>
<td>1 / 21</td>
<td>1.2</td>
<td></td>
<td>0.30[0.01,7.88]</td>
</tr>
<tr>
<td>Morton</td>
<td>1 / 40</td>
<td>2 / 36</td>
<td>2.1</td>
<td></td>
<td>0.44[0.04,5.02]</td>
</tr>
<tr>
<td>Ceremuzynski</td>
<td>1 / 25</td>
<td>3 / 23</td>
<td>2.3</td>
<td></td>
<td>0.28[0.03,2.88]</td>
</tr>
<tr>
<td>Pereira</td>
<td>0 / 27</td>
<td>7 / 27</td>
<td>2.6</td>
<td></td>
<td>0.11[0.01,0.97]</td>
</tr>
<tr>
<td>Smith</td>
<td>2 / 200</td>
<td>7 / 200</td>
<td>4.3</td>
<td></td>
<td>0.28[0.06,1.36]</td>
</tr>
<tr>
<td>Feldstedt</td>
<td>10 / 150</td>
<td>8 / 148</td>
<td>8.6</td>
<td></td>
<td>1.25[0.48,3.26]</td>
</tr>
<tr>
<td>Thogersen</td>
<td>4 / 130</td>
<td>8 / 122</td>
<td>6.3</td>
<td></td>
<td>0.45[0.13,1.54]</td>
</tr>
<tr>
<td>Golf</td>
<td>5 / 23</td>
<td>13 / 33</td>
<td>6.4</td>
<td></td>
<td>0.43[0.13,1.44]</td>
</tr>
<tr>
<td>Shechter 90</td>
<td>1 / 59</td>
<td>9 / 56</td>
<td>2.7</td>
<td></td>
<td>0.09[0.01,0.74]</td>
</tr>
<tr>
<td>Singh</td>
<td>6 / 76</td>
<td>11 / 75</td>
<td>7.7</td>
<td></td>
<td>0.50[0.17,1.43]</td>
</tr>
<tr>
<td>Shechter 95</td>
<td>4 / 107</td>
<td>17 / 108</td>
<td>7.1</td>
<td></td>
<td>0.21[0.07,0.64]</td>
</tr>
<tr>
<td>Rasmussen</td>
<td>9 / 135</td>
<td>23 / 135</td>
<td>10.2</td>
<td></td>
<td>0.35[0.15,0.78]</td>
</tr>
<tr>
<td>LIMIT-2</td>
<td>90 / 1159</td>
<td>118 / 1157</td>
<td>17.5</td>
<td></td>
<td>0.74[0.56,0.99]</td>
</tr>
<tr>
<td>ISIS-4</td>
<td>2216 / 29011</td>
<td>2103 / 29039</td>
<td>19.4</td>
<td></td>
<td>1.06[1.00,1.13]</td>
</tr>
</tbody>
</table>

*Total (95%CI):* 2351 / 31212 2331 / 31226

Test for heterogeneity chi-square = 40.18 df=14 p=0.0002
Test for overall effect z=-3.34 p=0.0008

**RE gives less ‘contrasted’ weights between big and small studies**
Identifying Heterogeneity

- Visualize data
- Statistical test
  - Low power since usually very few studies
  - But has excessive power to detect clinically unimportant heterogeneity with many studies

Quantify amount of heterogeneity
- Between-study variance
- Test Statistic
- Percent of Total Variation Between Studies
Different subgroups representing various patient characteristics and disease manifestations may have different responses to a treatment.

Different inclusion criteria, patient recruitment, and random variations may result in study cohorts consisting of different distributions and combinations of subgroups in RCTs.

Protocol differences, study design and reporting flaws, and publication bias contribute to bias or exclusion of some studies in a meta-analysis.

Interpreting the results of meta-analysis of RCTs depends on how the data are synthesized: weighted average, regression, or individual patient data modeling.
Dealing With Heterogeneity

HETEROGENEOUS TREATMENT EFFECTS

- IGNORE (insensitive)
  - FIXED EFFECTS MODEL
- ESTIMATE
  - DO NOT COMBINE WHEN HETEROGENEITY IS PRESENT
- ACCOUNT FOR
  - RANDOM EFFECTS MODEL
- EXPLAIN
  - SUBGROUP ANALYSES
  - META-REGRESSION (control rate, covariates)
OVERALL ESTIMATE
combining summary data

META-REGRESSION
modeling summary data

RESPONSE SURFACE
modeling individual patient data
Basic Concept of Cumulative Meta-Analysis

Studies ordered chronologically or by covariates

- Study 1
- Study 2
- Study 3
- Study 4

- Pool Studies 1 to 2 → Cumulative M-A 1
- Pool Studies 1 to 3 → Cumulative M-A 2
- Pool Studies 1 to 4 → Cumulative M-A 3

- Study n-1
- Study n

- Pool Studies 1 to n-1 → Cumulative M-A n-2
- Pool Studies 1 to n → Cumulative M-A n-1
Intravenous Streptokinase Therapy in Acute Myocardial Infarction

Individual RCT and Overall Meta-Analysis Results

Cumulative Mantel-Haenszel method

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Patients</th>
<th>Odds Ratio (Log Scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fletcher</td>
<td>1959</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Dewar</td>
<td>1963</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>European 1</td>
<td>1969</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td>European 2</td>
<td>1971</td>
<td>730</td>
<td></td>
</tr>
<tr>
<td>Heikinheimo</td>
<td>1971</td>
<td>428</td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>1971</td>
<td>321</td>
<td></td>
</tr>
<tr>
<td>Australian 1</td>
<td>1973</td>
<td>517</td>
<td></td>
</tr>
<tr>
<td>Frankfurt 2</td>
<td>1973</td>
<td>206</td>
<td></td>
</tr>
<tr>
<td>NHLBI SMIT</td>
<td>1974</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>1975</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>Valere</td>
<td>1975</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Klein</td>
<td>1976</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>UK-Collab</td>
<td>1976</td>
<td>595</td>
<td></td>
</tr>
<tr>
<td>Austrian</td>
<td>1977</td>
<td>728</td>
<td></td>
</tr>
<tr>
<td>Australian 2</td>
<td>1977</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>Lasieria</td>
<td>1977</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>N Ger Collab</td>
<td>1977</td>
<td>483</td>
<td></td>
</tr>
<tr>
<td>Witchitz</td>
<td>1977</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>European 3</td>
<td>1979</td>
<td>315</td>
<td></td>
</tr>
<tr>
<td>ISAM</td>
<td>1986</td>
<td>1741</td>
<td></td>
</tr>
<tr>
<td>GISSI-1</td>
<td>1986</td>
<td>11712</td>
<td></td>
</tr>
<tr>
<td>Olson</td>
<td>1986</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>Barcofo</td>
<td>1986</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Schreiber</td>
<td>1986</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Cribier</td>
<td>1986</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Sainsous</td>
<td>1986</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Durand</td>
<td>1987</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1987</td>
<td>219</td>
<td></td>
</tr>
<tr>
<td>Bassand</td>
<td>1987</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>Vlay</td>
<td>1988</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Kennedy</td>
<td>1988</td>
<td>368</td>
<td></td>
</tr>
<tr>
<td>ISIS-2</td>
<td>1988</td>
<td>17187</td>
<td></td>
</tr>
<tr>
<td>Wisenberg</td>
<td>1988</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>36974</td>
<td></td>
<td>z = -8.16 p &lt; 0.000001</td>
</tr>
</tbody>
</table>

Favors Treatment    Favors Control

Cumulative odds ratio: z = -8.16 p < 0.000001

Z-scores:
- GISSI-1: z = -4.98 p < 0.000001
- Overall: z = -8.16 p < 0.000001
Findings of Cumulative Meta-analysis

- Clinical experts’ recommendations often are unreliably synchronized with developing RCT evidence.

- Large clinical trials often echo findings from meta-analyses of several smaller studies.

- Trends established by cumulative meta-analyses of previous studies are unlikely to be reversed.

- Cumulative meta-analysis is an example of Bayesian updating.
Hierarchical Meta-Analysis Model

- $Y_i$ observed treatment effect (e.g. odds ratio) and $\theta_i$ unknown true treatment effect from $i^{th}$ study

- First level describes variability of $Y_i$ given $\theta_i$

$$Y_i \sim N\left(\theta_i, \sigma_i^2\right)$$

- Within-study variance often assumed known

- But could use common variance estimate if studies are small

- If data are binary, use binomial distribution here
Hierarchical Meta-Analysis Model

Second level describes variability of study-level parameters $\theta_i$

$$\theta_i \sim N(\theta, \tau^2)$$

in terms of population level parameters: $\theta$ and $\tau^2$

**Fixed Effects**

$$\theta_i = \theta \quad (\tau^2 = 0)$$

**Random Effects**

$$\theta_i \sim N(\theta, \tau^2)$$

$$\Rightarrow Y_i \sim N(\theta_i, \sigma_i^2 + \tau^2)$$
Bayesian Hierarchical Model

Placing priors on *hyperparameters* \((\theta, \tau^2)\) makes Bayesian model

Posterior distribution of random effects is

\[
\theta_i | y_i, \theta, \sigma^2 \sim N(\theta_i^*, V_i (1 - B_i))
\]

where

\[
\theta_i^* = (1 - B_i) y_i + B_i \theta
\]

\[
B_i = V_i / (V_i + \tau^2)
\]

Each study’s conditional mean is weighted average of observed study mean and overall mean

- Inferences sensitive to prior on \(\tau^2\)
Shrinkage

\[ B_i = \frac{V_i}{(V_i + \tau^2)} \] are shrinkage factors

- Larger \( B_i \) shrink \( \theta_i^* \) more back to the grand mean \( \theta \)
- Well-estimated studies (small within-study variances) weighted most
- Bigger within-study variances lead to more shrinkage
- Smaller within-study variances lead to less shrinkage
- Increased between-study variance weights studies more evenly
Example: Magnesium for AMI

- Infamous because random effects and fixed effects analysis lead to different conclusions
  
  Random effects OR = 0.59
  Fixed effects OR = 1.02

- Very large, influential clinical trial showed no treatment benefit

- Contradicted earlier MA with large trial showing large benefit
# Meta-analysis for Magnesium Studies

## Mortality

<table>
<thead>
<tr>
<th>Study</th>
<th>Treated</th>
<th>Control</th>
<th>Est</th>
<th>95% PI</th>
<th>Est</th>
<th>95% PI</th>
<th>Pr(OR&lt;1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morton</td>
<td>1/40</td>
<td>2/36</td>
<td>0.44</td>
<td>0.0, 5.0</td>
<td>0.54</td>
<td>0.2, 1.6</td>
<td>0.89</td>
</tr>
<tr>
<td>Smith</td>
<td>2/200</td>
<td>7/200</td>
<td>0.28</td>
<td>0.1, 1.5</td>
<td>0.46</td>
<td>0.1, 1.1</td>
<td>0.96</td>
</tr>
<tr>
<td>Abraham</td>
<td>1/48</td>
<td>1/46</td>
<td>0.96</td>
<td>0.1, 15.8</td>
<td>0.61</td>
<td>0.2, 1.9</td>
<td>0.84</td>
</tr>
<tr>
<td>Feldstedt</td>
<td>10/150</td>
<td>8/148</td>
<td>1.25</td>
<td>0.5, 3.3</td>
<td>0.86</td>
<td>0.4, 1.9</td>
<td>0.70</td>
</tr>
<tr>
<td>Rasmussen</td>
<td>9/135</td>
<td>23/135</td>
<td>0.35</td>
<td>0.2, 0.8</td>
<td>0.43</td>
<td>0.2, 0.9</td>
<td>0.99</td>
</tr>
<tr>
<td>Ceremuz.</td>
<td>1/25</td>
<td>3/23</td>
<td>0.28</td>
<td>0.0, 2.9</td>
<td>0.49</td>
<td>0.1, 1.4</td>
<td>0.92</td>
</tr>
<tr>
<td>Shechter I</td>
<td>1/59</td>
<td>9/56</td>
<td>0.09</td>
<td>0.0, 0.7</td>
<td>0.38</td>
<td>0.1, 1.4</td>
<td>0.97</td>
</tr>
<tr>
<td>LIMIT 2</td>
<td>90/1159</td>
<td>118/1157</td>
<td>0.74</td>
<td>0.6, 1.0</td>
<td>0.73</td>
<td>0.6, 1.0</td>
<td>0.99</td>
</tr>
<tr>
<td>ISIS-4</td>
<td>2216/29011</td>
<td>2103/29039</td>
<td>1.06</td>
<td>1.0, 1.1</td>
<td>1.06</td>
<td>1.0, 1.1</td>
<td>0.04</td>
</tr>
<tr>
<td>Shechter II</td>
<td>2/89</td>
<td>12/80</td>
<td>0.13</td>
<td>0.0, 0.6</td>
<td>0.36</td>
<td>0.1, 0.9</td>
<td>0.99</td>
</tr>
<tr>
<td>Singh</td>
<td>6/76</td>
<td>11/75</td>
<td>0.50</td>
<td>0.2, 1.4</td>
<td>0.54</td>
<td>0.2, 1.1</td>
<td>0.95</td>
</tr>
</tbody>
</table>

## Pooled Odds Ratio

<table>
<thead>
<tr>
<th>Study</th>
<th>Treated</th>
<th>Control</th>
<th>Est</th>
<th>95% PI</th>
<th>Est</th>
<th>95% PI</th>
<th>Pr(OR&lt;1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled</td>
<td>90/1159</td>
<td>118/1157</td>
<td>0.59</td>
<td>0.4, 0.9</td>
<td>0.55</td>
<td>0.3, 0.9</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Kernel Density Plots of Posteriors
Distribution of Between-Study Variance

Probability Density

Between-study variance

0.00 0.10 0.20 0.30 0.40
0.0 0.1 0.2 0.3 0.4

Distribution curve showing the probability density of between-study variance.
Meta-Regression

• Investigate sources of heterogeneity in meta-analysis

• Regression analysis to identify correlations between treatment effects (outcomes) and covariates of interest (predictors)

• Estimates *interaction* between covariate and treatment effect, i.e. how treatment effect is *modified* by covariate

• Unit of analysis is the individual study

• Correlation implies treatment interaction
Factors may be study-level or subject-level

Study-level factors: blinding, randomization, dosage, protocol

Subject-level factors: age, gender, race, blood pressure

Study effect is no longer a single value, but is a function of predictors

\[ \theta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + u_i \]

Or can use baseline risk level (control rate)
Meta-Regression with Study-Level Summary of Patient Level Covariates

- Weighted regression (need to make adjustment to program because weights known exactly)
- Data points proportional to study size
Problems with Meta-Regression

• Number of studies usually small
• Number of potential predictors may be large
• Data may be unavailable (not conceived or not reported)
• Covariates pre-selected (biased?)
• Little variation in range of mean predictor
• Subject-level factors can be affected by ecological bias
• Causality uncertain
Ecological Bias

• Group averages don't represent individuals well
  – E.g., what does percentage male/female mean?

• Averages have little between-study variation

• Averages do not account for within-study variation e.g. 40 year average age can mean different things

<table>
<thead>
<tr>
<th>Study</th>
<th>Mean Age</th>
<th>%&gt; 60</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>40</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>B</td>
<td>40</td>
<td>10</td>
<td>0.8</td>
</tr>
</tbody>
</table>

• Events concentrated in high-risk subgroup
  – May want to construct group-level variable to represent this
    E.g., percentage of elderly, rather than mean age
Baseline Risk Meta-Regression

- Control group event rate reflects multiple risk factors
  - different populations
  - underlying baseline risk of patients
  - length of study follow-up
  - treatment delivery

- Related to severity of illness but not interpretable for individual

- Data always available

- May signal multiple causes

- Standard weighted LS biased
  - ignores correlated measurement error
# Meta-Regression vs. Individual Patient Regression

<table>
<thead>
<tr>
<th></th>
<th>Meta-Regression</th>
<th>Individual Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost</strong></td>
<td>Cheap</td>
<td>Expensive</td>
</tr>
<tr>
<td><strong>Data Available</strong></td>
<td>Usually</td>
<td>Infrequently</td>
</tr>
<tr>
<td><strong>Factors</strong></td>
<td>Study</td>
<td>Patient and Study</td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td>Reported</td>
<td>Updated, Complete</td>
</tr>
<tr>
<td><strong>Data Cleaning</strong></td>
<td>Impossible</td>
<td>Possible</td>
</tr>
<tr>
<td><strong>Bias</strong></td>
<td>Reporting, Ecological</td>
<td>Reporting, Retrieval</td>
</tr>
<tr>
<td><strong>Interpretation</strong></td>
<td>Study-specific</td>
<td>Patient-specific</td>
</tr>
</tbody>
</table>
ACE Inhibitors for Non-Diabetic Renal Disease

- Meta-analysis of 11 RCTs of ACE inhibitors (ACEI) published in 1997 showed treatment effective for non-diabetics in preventing progression of disease.

- Is ACEI effect completely explained by its effect to lower blood pressure and urine protein?

- Do ACEI work equally well for all nondiabetic renal patients or are there treatment interactions?

- What is the optimal dosing of ACEI and what concomitant medications might improve its efficacy?

- With only 10 studies, need patient-level data to answer all these questions.
Meta-Regression with Summary Data

\[ \overline{Y}_{j1} - \overline{Y}_{j0} \sim N(\beta_j, \omega_j^2) \]

\[ \beta_j \sim N(\beta_0 + \beta_1 X_j, \tau_\beta^2) \]

- Fixed Effects if \( \tau_\beta^2 = 0 \)
- Can fit with standard weighted linear regression model
- With individual patient data, can fit by two-step process
Individual Patient Data Regression Model

\[ Y_{ij} \sim N(\alpha_j + \beta_j T_{ij} + \gamma Z_{ij} + \delta Z_{ij} \cdot T_{ij}, \sigma_j^2) \]

\[ \alpha_j \sim N(\alpha_0 + \alpha_1 X_j, \tau_\alpha^2) \]

\[ \beta_j \sim N(\beta_0 + \beta_1 X_j, \tau_\beta^2) \]

- Multilevel model without aggregate effects
  \[ \tau_\alpha^2 = \tau_\beta^2 = 0 \]

- Can also assume common study variance \( \sigma^2 \)
Combining IPD and Summary Data

\[ Y_{ij}^* \sim N(D_j \alpha_j + \beta_j T_{ij}, V_j^*) \]

\[ \alpha_j \sim N(\alpha_0 + \alpha_1 X_j, \tau^2_{\alpha}) \]

\[ \beta_j \sim N(\beta_0 + \beta_1 X_j, \tau^2_{\beta}) \]

<table>
<thead>
<tr>
<th>IPD</th>
<th>Summary Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_{ij}^* )</td>
<td>( Y_{ij} )</td>
</tr>
<tr>
<td>( V_j^* )</td>
<td>( \sigma_j^2 )</td>
</tr>
<tr>
<td>( D_j )</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Within-Study Interaction
Issues With Meta-Analysis of Observational Studies

• Need to adjust for potential confounders

• Different studies may adjust for different confounders or may use different adjustment techniques
  
  – Some variables uncollected in original studies

• May want data on individual participants

• Misclassification and measurement of exposure

• Selection of subjects for control group may differ

• Lack of knowledge of study design characteristics
# Studies of Maternal Obesity & Stillbirth

## Cohort

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Time</th>
<th>Type</th>
<th>Source</th>
<th>Size</th>
<th>Normal</th>
<th>Overweight</th>
<th>Obese</th>
<th>Severely Obese</th>
</tr>
</thead>
</table>

## Case-control

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Time</th>
<th>Type</th>
<th>Source</th>
<th>Cases/C controls</th>
<th>Normal</th>
<th>Overweight</th>
<th>Obese</th>
<th>Severely Obese</th>
</tr>
</thead>
</table>
Network of 12 Antidepressants

19 meta-analyses published in the last two years
### Indirect Comparisons of Multiple Treatments

<table>
<thead>
<tr>
<th>Trial</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
</tbody>
</table>

- **Want to compare A vs. B**
  - Direct evidence from trials 1, 2 and 7
  - Indirect evidence from trials 3, 4, 5, 6 and 7

- **Combining all “A” arms and comparing with all “B” arms destroys randomization**

- **Use indirect evidence of A vs. C and B vs. C comparisons as additional evidence to preserve randomization and within-study comparison**
Indirect Comparisons

How do we make the indirect comparisons:

Calculate effect of A vs. C and B vs. C separately

\[ T_{AB} = T_{AC} - T_{BC} \]

with SE = square root of sum of variances

Strong Assumptions:

• All trials comparing pairs of tx arms estimate same effect

• Different sets of trials being used are similar
Measuring Inconsistency

Suppose we have AB, AC, BC direct evidence

Indirect estimate

\[
\hat{d}_{BC}^{\text{indirect}} = \hat{d}_{AC}^{\text{direct}} - \hat{d}_{AB}^{\text{direct}}
\]

Measure of inconsistency:

\[
\hat{\omega}_{BC} = \hat{d}_{BC}^{\text{indirect}} - \hat{d}_{BC}^{\text{direct}}
\]

Approximate test (normal distribution):

\[
z_{BC} = \frac{\hat{\omega}_{BC}}{\sqrt{V(\hat{\omega}_{BC})}}
\]

with variance

\[
V(\hat{\omega}_{BC}) = V(\hat{d}_{BC}^{\text{direct}}) + V(\hat{d}_{AC}^{\text{direct}}) + V(\hat{d}_{AB}^{\text{direct}})
\]
Example

- Population: Patients with cardiovascular disease
- Treatments: Statin treatment (different doses), fibrate
- Comparator: Conventional care or placebo
- Covariates: Baseline cholesterol, triglycerides
- Outcomes:
  - Myocardial infarction (fatal or non-fatal)
  - Stroke (fatal or non-fatal)
  - Death from all other causes
- Design: RCTs
Network

Low Dose Statins

High Dose Statins

Control

Fibrates

3

2

9

4
Data Setup

- Each study has 6 possible outcomes and 4 possible tx’s
- Not all tx’s carried out in each study
- Not all outcomes observed in each study
- Incomplete data with partial information from summary categories
- Can use available information to impute missing values
- Can build into Bayesian algorithm using multinomial model
<table>
<thead>
<tr>
<th></th>
<th>Other death</th>
<th>Non-fatal Stroke</th>
<th>Non-fatal MI</th>
<th>Fatal Stroke</th>
<th>Fatal MI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fibrate v Control</strong></td>
<td>1.03</td>
<td>0.90</td>
<td>0.69</td>
<td>0.80</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>(0.63 – 1.80)</td>
<td>(0.57 – 1.33)</td>
<td>(0.44 – 0.98)</td>
<td>(0.40 – 1.49)</td>
<td>(0.50 – 1.53)</td>
</tr>
<tr>
<td><strong>LDS v Control</strong></td>
<td>0.93</td>
<td>0.87</td>
<td>0.76</td>
<td>0.72</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>(0.60 – 1.41)</td>
<td>(0.56 – 1.42)</td>
<td>(0.46 – 1.08)</td>
<td>(0.32 – 1.69)</td>
<td>(0.39 – 1.04)</td>
</tr>
<tr>
<td><strong>HDS v Control</strong></td>
<td>0.84</td>
<td>0.72</td>
<td>0.66</td>
<td>0.74</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>(0.69 – 1.15)</td>
<td>(0.50 – 0.94)</td>
<td>(0.53 – 0.81)</td>
<td>(0.41 – 1.13)</td>
<td>(0.45 – 0.83)</td>
</tr>
<tr>
<td><strong>LDS v Fibrate</strong></td>
<td>0.88</td>
<td>0.95</td>
<td>1.11</td>
<td>0.88</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>(0.49 – 1.57)</td>
<td>(0.69 – 1.15)</td>
<td>(0.62 – 1.88)</td>
<td>(0.38 – 2.59)</td>
<td>(0.37 – 1.40)</td>
</tr>
<tr>
<td><strong>HDS v Fibrate</strong></td>
<td>0.81</td>
<td>0.80</td>
<td>0.97</td>
<td>0.92</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>(0.54 – 1.51)</td>
<td>(0.45 – 1.26)</td>
<td>(0.66 – 1.47)</td>
<td>(0.40 – 2.41)</td>
<td>(0.37 – 1.26)</td>
</tr>
<tr>
<td><strong>HDS v LDS</strong></td>
<td>0.94</td>
<td>0.80</td>
<td>0.87</td>
<td>1.01</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>(0.62 – 1.36)</td>
<td>(0.50 – 1.25)</td>
<td>(0.64 – 1.35)</td>
<td>(0.53 – 1.89)</td>
<td>(0.690 – 1.74)</td>
</tr>
</tbody>
</table>
Multivariate Model

Assume two outcomes \(Y_{i1}, Y_{i2}\) observed in \(I\) studies

\[
\begin{align*}
\left( \begin{array}{c}
Y_{i1} \\
Y_{i2}
\end{array} \right) & \sim N \left( \left[ \begin{array}{c}
\theta_{i1} \\
\theta_{i2}
\end{array} \right], \left[ \begin{array}{cc}
\rho_{W_i}s_{i1}s_{i2} & s_{i1}^2 \\
s_{i1}s_{i2} & s_{i2}^2
\end{array} \right] \right) \\
\left( \begin{array}{c}
\theta_{i1} \\
\theta_{i2}
\end{array} \right) & \sim N \left( \left[ \begin{array}{c}
\theta_1 \\
\theta_2
\end{array} \right], \left[ \begin{array}{cc}
\tau_1^2 & \rho_B\tau_1\tau_2 \\
\rho_B\tau_1\tau_2 & \tau_2^2
\end{array} \right] \right)
\end{align*}
\]

- May be difficult to estimate within-study correlations
- Instead, could reformulate problem to estimate only single correlation for marginal model adding within and between-study
- Or could estimate each outcome separately
- Ignoring within-study correlation gives biased estimates
Longitudinal Model

Each study has K measurements taken over time

$\theta_i$ is vector of K treatment effects at each time for $i^{th}$ study

$\theta$ is vector of average treatment effects at each time

$$Y_i \sim MVN\left(\theta_i, \Sigma_i\right)$$

$$\theta_i \sim MVN\left(X_i\theta, Z_i\text{DZ}_i\right)$$

- Often reporting times differ across studies
- Can aggregate
Longitudinal Model: Variance Structure

• $\Sigma_i$ usually assumed known
• May not have information reported on correlations
• Could assume $\Sigma_i$ diagonal or take $\Sigma_i = W_i^{-1/2}CW_i^{-1/2}$
• $W_i$ is diagonal matrix holding known within-study variances
• $C$ is correlation matrix constant across studies and estimated from data
• Could use autoregressive structure or allow different random effects at each time
  • E.g. $D$ is AR(1) with unequal variances
Uses of Diagnostic Tests

- Screen (mammography for breast cancer)
- Diagnose (ECG for acute myocardial infarction)
- Grade (stage of cancer)
- Monitor progression (recurrence)
- Monitor therapy (blood drug level) and therapeutic response (regression of tumor size)
- Guide treatments (arteriography for CABG)

False positive results may lead to unnecessary tests and treatments and possible harms

False negative results may prevent proper treatment
Defining Test Performance

\[\text{Prevalence} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}\]

\[\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}\]

\[\text{Sensitivity (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}\]

\[\text{Specificity (TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}}\]

\[\text{Predictive Value +} = \frac{\text{TP}}{\text{TP} + \text{FP}}\]

\[\text{Predictive Value -} = \frac{\text{TN}}{\text{TN} + \text{FN}}\]

\[\text{Likelihood Ratio +} = \frac{\text{TP} / (\text{TP} + \text{FN})}{\text{FP} / (\text{FP} + \text{TN})}\]

\[\text{Likelihood Ratio -} = \frac{\text{FN} / (\text{TP} + \text{FN})}{\text{TN} / (\text{FP} + \text{TN})}\]

\[\text{Odds Ratio} = \frac{\text{TP} \times \text{TN}}{\text{FP} \times \text{FN}}\]

\[= \frac{\text{Se} / (1 - \text{Se})}{(1 - \text{Sp}) / \text{Sp}}\]

\[= \text{LR +} / \text{LR -}\]
Changing diagnostic threshold or disease spectrum changes test performance.
Making ROC Curve from Multiple Test Thresholds

Multiple thresholds evaluated in test

Diseased

Not diseased

Sensitivity

1 - Specificity

a b c d
# Full Cycle of Diagnostic Test Evaluation: Magnetic Resonance Spectroscopy for Brain

<table>
<thead>
<tr>
<th>Level</th>
<th>Example of study purpose</th>
<th># studies</th>
<th># patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Technical feasibility</td>
<td>Ability to produce consistent spectra</td>
<td>85</td>
<td>2434</td>
</tr>
<tr>
<td>2: Test accuracy</td>
<td>Sensitivity and specificity</td>
<td>8</td>
<td>461</td>
</tr>
<tr>
<td>3: Diagnostic impact</td>
<td>Percentage of times clinicians’ subjective assessment of diagnostic probabilities changed after the test</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>4: Therapeutic impact</td>
<td>Percentage of times therapy planned before MRS changed after the test</td>
<td>2</td>
<td>105</td>
</tr>
<tr>
<td>5: Clinical outcomes</td>
<td>Percentage of patients who improved with MRS diagnosis compared with those without MRS</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6: Societal Impact</td>
<td>CEA: use of test in asymptomatic patients</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Diagnostic Technology Controversy: Screening Mammography RCTs

- 1999 study found no decrease in breast cancer mortality in Sweden, where screening has been recommended since 1985
- Reviewed methodological quality of mammography trials and repeated a meta-analysis
# Relative Risk of Death from Breast Cancer

<table>
<thead>
<tr>
<th></th>
<th>Screening</th>
<th>Control</th>
<th>Screening</th>
<th>Control</th>
<th>Relative risk (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Randomization adequate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malmo</td>
<td>21088</td>
<td>21195</td>
<td>63</td>
<td>66</td>
<td>0.96 (0.68-1.35)</td>
</tr>
<tr>
<td>Canada</td>
<td>44925</td>
<td>44910</td>
<td>120</td>
<td>111</td>
<td>1.08 (0.84–1.40)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>66013</td>
<td>66105</td>
<td>183</td>
<td>177</td>
<td>1.04 (0.84-1.27)</td>
</tr>
<tr>
<td><strong>Randomization not adequate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goteberg</td>
<td>11724</td>
<td>14217</td>
<td>18</td>
<td>40</td>
<td>0.55 (0.31-0.95)</td>
</tr>
<tr>
<td>Stockholm</td>
<td>40318</td>
<td>19943</td>
<td>66</td>
<td>45</td>
<td>0.73 (0.50-1.06)</td>
</tr>
<tr>
<td>Kopparberg</td>
<td>38589</td>
<td>18582</td>
<td>126</td>
<td>104</td>
<td>0.58 (0.45-0.76)</td>
</tr>
<tr>
<td>Ostergotland</td>
<td>38491</td>
<td>37403</td>
<td>135</td>
<td>173</td>
<td>0.76 (0.61-0.95)</td>
</tr>
<tr>
<td>New York</td>
<td>30131</td>
<td>30565</td>
<td>153</td>
<td>196</td>
<td>0.79 (0.64-0.98)</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>22926</td>
<td>21342</td>
<td>156</td>
<td>167</td>
<td>0.87 (0.70-1.08)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>182179</td>
<td>142052</td>
<td>654</td>
<td>725</td>
<td>0.75 (0.67-0.83)</td>
</tr>
</tbody>
</table>
Policy Results

- Switzerland decided to not cover screening mammography
- National Cancer Institute wavered on value of screening mammograms
- Women and doctors more confused about value of test
Conclusions

- Evidence-based medicine requires collaboration of doctors, statisticians, librarians, epidemiologists and other experts
- Goal is to provide scientific basis for clinical decisions
- Often requires sifting through extensive literature
- Systematic reviews more scientific than narrative reviews
- Determines validity of evidence and identifies research gaps
- Discovery of heterogeneity can improve interventions