Step 3 in EBM: appraisal

1. Formulate an answerable question
2. Track down the best evidence
3. Critically appraise the evidence for:
   • Validity
   • Impact (size of the benefit)
   • Applicability
4. Integrate with clinical expertise and patient values
5. Evaluate our effectiveness and efficiency
   • keep a record; improve the process

A CHECKLIST FOR APPRAISING RANDOMIZED CONTROLLED TRIALS

1. Was the objective of the trial sufficiently described?
2. Was a satisfactory statement given of the diagnostic criteria for entry to the trial?
3. Were random samples of patients allocated to treatment groups?
4. Was the treatment well defined?
5. Were the groups comparable in relevant measures? Was a primary outcome measure identified?
6. Were follow-up losses acceptable?
7. Were the statistical analyses clearly described and appropriate for the data analyzed?
8. What conclusions can be drawn from the statistical analyses?

Hydroxychloroquine in Rheumatoid Arthritis

Clinical Question
In people who take long-haul flights does wearing graduated compression stockings prevent DVT?
APPRAISAL OF RCTs

**The PECOT acronym: the 5 parts of every epidemiological study**

- **Participants**
- **Exposure Group**
- **Comparison Group**
- **Outcome**
- **Time**

All epidemiological studies can be hung on the GATE frame.

---

**Was it a fair race?**

1. Fair start?
2. Few drop outs?
3. Fair finish?

---

**Using the PICO to orient us**

**Clinical Question**
In people who take long-haul flights does wearing graduated compression stockings prevent DVT?

**Frequency and prevention of symptomless deep-vein thrombosis in long-haul flights: a randomised trial**

Scurr et al, Lancet 2001; 357:1485-89

---

**Use the RAMMbo to check validity**

**Was the Study valid?**

1. **Representativeness**
   - Who did the subjects represent?
2. **Allocation**
   - Was the assignment to treatments randomised?
   - Were the groups similar at the trial’s start?
3. **Maintenance**
   - Were the groups treated equally?
   - Were outcomes ascertained & analysed for most patients?
4. **Measurements blinded OR objective**
   - Were patients and clinicians “blinded” to treatment? OR
   - Were measurements objective & standardised?

Study statistics (p-values & confidence intervals)

User Guide. JAMA, 1993
**Participants**

Study Setting: volunteers, UK, ? 1990s

Eligible Participants: no previous DVT, > 50 yrs, planned economy air travel 2 sectors > 8 hours

Participants: 200, mean age 61-62 years

---

**Appraisal checklist - RAMMbo**

1. **Recruitment**
   - Who did the subjects represent?

2. **Allocation**
   - Was the assignment to treatments randomised?
   - Were the groups similar at the trial’s start?

3. **Maintainance**
   - Were the groups treated equally?
   - Were outcomes ascertained & analysed for most patients?

4. **Measurements**
   - Were patients and clinicians “blinded” to treatment? OR
   - Were measurements objective & standardised?

Study statistics (p-values & confidence intervals):

Guyatt. JAMA, 1993

---

**Exposure & Comparison Groups**

Exposure or Intervention Group (EG): Below knee compression stockings

Comparison or Control Group (CG): no stockings

Exposure or Intervention Group (EG): 115

Comparison or Control Group (CG): 116

100 100

---

**Comparable Groups: the only difference should be the treatments**

✓

(i) I C

(ii) I C

Is the difference between I and C because of (i) the intervention or (ii) because the groups were not comparable in the first place?

---

**Appraisal of RCTs**

**DVT in long-haul flights:** Lancet 2001;357:1485-9

Participants

Study Setting: volunteers, UK, ? 1990s

Eligible Participants: no previous DVT, > 50 yrs, planned economy air travel 2 sectors > 8 hours

Participants: 200, mean age 61-62 years

---

**Comparable Groups: the only difference should be the treatments**

✓

(i) I C

(ii) I C

Is the difference between I and C because of (i) the intervention or (ii) because the groups were not comparable in the first place?

---

**Fair Allocation to treatments**

How do we get comparable groups?

- Was assignment to treatments randomised?
- Was the allocation process tamper proof? OR
- Were the groups similar at start of trial?
Benefits of Randomisation (and Allocation Concealment)

- Minimises confounding - known and unknown potential confounders are evenly distributed between study groups
  - reduces bias in those selected for treatment
  - guarantees treatment assignment will not be based on patients’ prognosis

Fair Allocation – balance achieved?

Were the groups similar at the start?

- Usually Table 1 in Results section
- Do imbalances favour one treatment?

Allocation Concealment OR

Demonstrated baseline balance

BEST – most valid technique
  - Central computer randomization

DOUBTFUL
  - Envelopes, etc

NOT RANDOMISED
  - Date of birth, alternate days, etc

Appraisal checklist - RAMMbo

Study biases
1. Recruitment
   - Who did the subjects represent?
2. Allocation
   - Was the assignment to treatments randomised?
   - Were the groups similar at the trial’s start?
3. Maintenance
   - Were outcomes ascertained & analysed for most patients?
   - Were the groups treated equally?
4. Measurements
   - Were patients and clinicians “blinded” to treatment? OR
   - Were measurements objective & standardised?

Study statistics (p-values & confidence intervals)

Guyatt. JAMA, 1993

Effects of non-equal treatment

Apart from actual intervention - groups should receive identical care!

- Trial of Vitamin E in pre-term infants (1948)
- Vit E “prevented” retrolental fibroplasia (By removal from 100% Oxygen to give the frequent doses of Vit E!)
- Rx: Give placebo in an identical regime, and a standard protocol
## Equal treatment in DVT study?

<table>
<thead>
<tr>
<th>Number of Participants</th>
<th>No Stockings</th>
<th>Stockings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Hormone replacement therapy</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>Thyroxine</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Antihypertensives, including diuretics</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Antipenic ulcer drugs</td>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>

*Includes additions to usual drugs

Table 3: All drugs taken by volunteers who attended for examination before and after air travel*

Scurr et al, Lancet 2001; 357:1485-89

## How important are the losses?

- Equally distributed?
  - Stocking group: 6 men, 9 women - 15
  - No stocking group: 7 men, 9 women - 16

- Similar characteristics?
  - No information provided

## Maintaining the Randomisation

- **Principle 1 (Intention to treat)**
  - Once a patient is randomised, s/he should be analysed in the group randomised to - even if they discontinue, never receive treatment, or crossover.

- **Principle 2 (adequate followup)**
  - "5-and-20 rule of thumb"
  - 5% probably leads to little bias
  - >20% poses serious threats to validity

## Follow-up in DVT study?

- 231 randomised (115 to stockings; 116 none)
- 200 analysed
  - 27 were unable to attend for subsequent ultrasound
  - 2 were excluded from analysis because they were upgraded to business class
  - 2 were excluded from analysis because they were taking anticoagulants

See figure on page 1486

Scurr et al, Lancet 2001; 357:1485-89

## Appraisal checklist - RAMMbo

**Study biases**
1. Recruitment
   - Who did the subjects represent?
2. Allocation
   - Was the assignment to treatments randomised?
   - Were the groups similar at the trial's start?
3. Maintenance
   - Were outcomes ascertained & analysed for most patients?
   - Were the groups treated equally?
4. Measurements
   - Were patients and clinicians "blinded" to treatment?
   - Were measurements objective & standardised?

*Study statistics (p-values & confidence intervals)*

Guyatt. JAMA, 1993

## Measurement Bias - minimizing differential error

- Objective or Blinded
  - Participants?
  - Investigators?
  - Outcome assessors?
  - Analysts?
- Papers should report WHO was blinded and HOW it was done

Schulz and Grimes. Lancet, 2002
Summary
The true frequency of deep-vein thrombosis (DVT) during long-haul air travel is unknown. We sought to determine the frequency of DVT in the lower limb during long-haul economy-class air travel and the efficacy of graduated elastic compression stockings in its prevention.

Methods
We recruited 89 male and 142 female passengers over 50 years of age with no history of thromboembolic problems. Passengers were randomly allocated to one of two groups: one group wore class-I below-knee graduated elastic compression stockings, the other group did not. All the passengers made journeys lasting more than 8 h per flight (median total duration 24 h), returning to the UK within 6 weeks.

Duplex ultrasonography was used to assess the deep veins before and after travel. Blood samples were analysed for two specific common gene mutations, factor V Leiden (FVL) and prothrombin G20210A (PGM), which predispose to venous thromboembolism. A sensitive D-dimer assay was used to screen for the development of recent thrombosis.

Findings
12/116 passengers (10%; 95% CI 4·8–16·0%) developed symptomless DVT in the calf (five men, seven women). None of these passengers wore elastic compression stockings, and two were heterozygous for FVL. Four further patients who wore elastic compression stockings, had varicose veins and developed superficial thrombophlebitis. One of these passengers was heterozygous for both FVL and PGM. None of the passengers who wore class-I compression stockings developed DVT (95% CI 0–3·2%).

Lancet 2001; 357: 1485–89 See Commentary page 1461

Appraisal checklist - RAMMbo

Study biases
1. Recruitment
   • Who did the subjects represent?
2. Allocation
   • Was the assignment to treatments randomised?
   • Were the groups similar at the trial's start?
3. Maintenance
   • Were the groups treated equally?
   • Were outcomes ascertainment & analysed for most patients?
4. Measurements
   • Were patients and clinicians "blinded" to treatment? OR
   • Were measurements objective & standardised?

Study statistics (p-values & confidence intervals)

Guyatt. JAMA, 1993

Fundamental Equation of Error

Measure = Truth + Bias + Random Error

Two methods of assessing the role of chance

P-values (Hypothesis Testing)
• use statistical test to examine the 'null' hypothesis
• associated with "p values" - if p<0.05 then result is statistically significant

Confidence Intervals (Estimation)
• estimates the range of values that is likely to include the true value

Relationship between p-values and confidence intervals - if the value corresponding to 'no effect' (RR of 1 or treatment difference of 0) falls outside the CI then the result is statistically significant

P-values (Hypothesis Testing) - in DVT study

Incidence of DVT
• Stocking group - 0
• No Stocking group - 0.12

Risk difference = 0.12 - 0 = 0.12 (P=0.001)
The probability that this result would only occur by chance is 1 in 1000 → statistically significant
Confidence Intervals (Estimation)
- in DVT study

- Incidence of DVT
  - Stocking group - 0
  - No Stocking group - 0.12

  Risk difference = 0.12 - 0 = 0.12
  \( (95\% \text{ CI}, 0.058 - 0.20) \)

  The true value could be as low as 0.058 or as high as 0.20 - but is probably closer to 0.12

  Since the CI does not include the 'no effect' value of '0' \( \rightarrow \) the result is statistically significant

Placebo effect
- Trial in patients with chronic severe itching

Placebo effect - attributable to the expectation that the treatment will have an effect
A Systematic Review is a review of a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review.

Most reviews do not pass minimum criteria. A study of 158 reviews*

- Only 2 met all 10 criteria
- Median was only 1 of 10 criteria met

* McAlister Annals of Intern Med 1999

Is the review any good? FAST appraisal

- Question – What is the PICO?
  - Finding
    - Did they find most studies?
  - Appraisal
    - Did they select good ones?
  - Synthesis
    - What do they all mean?
  - Transferability of results

Do pedometers increase activity and improve health?

- Find: what is your search strategy?
  - Databases?
  - Terms?
  - Other methods?

METHODS
Data Sources and Search Strategies
In collaboration with a professional librarian, we developed individualized search strategies for 7 databases: MEDLINE (January 2006 to February 2007), and EMBASE, Sport Discus, PEDroNED, Cochrane Library, Thompson's (Thomson Scientific), CINAHL (Thomson), and ERIC (January 1999 to May 2000). We used search terms such as pedometer, activity monitor, and any variant. We also reviewed the bibliographies of retrieved articles and relevant conference proceedings and contacted experts in exercise physiology for additional studies.
FIND: Did they find all Studies?

- Check for existing systematic review?
- Good initial search
  - Terms (text and MeSH)
  - At least 2 Databases: MEDLINE, EMBASE, CINAHL, CCTR, ...
- Plus a Secondary search
  - Check references of relevant papers & reviews and
  - Find terms (words or MeSH terms) you didn't use
  - Search again! (snowballing)

Is finding all published studies enough?

- Negative studies less likely to be published than ‘Positive’
- How does this happen?
- Follow-up of 737 studies at Johns Hopkins*
  - Positive SUBMITTED more than negative (2.5 times)

Registered vs Published Studies

<table>
<thead>
<tr>
<th></th>
<th>Published</th>
<th>Registered</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. studies</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>Survival ratio</td>
<td>1.16</td>
<td>1.05</td>
</tr>
<tr>
<td>95% CI</td>
<td>1.06-1.27</td>
<td>0.98-1.12</td>
</tr>
<tr>
<td>P-Value</td>
<td>0.02</td>
<td>0.25</td>
</tr>
</tbody>
</table>

*Simes, J. Clin Oncol, 86, p1529

Which are biased? Which OK?

1. All positive studies
2. All studies with more than 100 patients
3. All studies published in BMJ, Lancet, JAMA or NEJM
4. All studies registered studies

Publication Bias: Solution

- All trials registered at inception,
  - The National Clinical Trials Registry: Cancer Trials
  - National Institutes of Health Inventory of Clinical Trials and Studies
  - International Registry of Perinatal Trials
- Meta-Registry of trial Registries
  - www.controlled-trials.com
**Flowchart**

345 identified → 91 duplicates → 254 screened → 223 not relevant → 31 retrieved in full → 17 excluded → 14 RCTs included.

**Selective Criticism of Evidence**

Biased appraisal increases polarization

- Capital punishment: beliefs and contradictory studies

**APPRAISE & select studies**

Did they select only the good quality studies?

**Selective Criticism of Evidence**

28 reviewers assessed one “study” results randomly positive or negative

<table>
<thead>
<tr>
<th></th>
<th>&quot;Positive&quot;</th>
<th>&quot;Negative&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>5.2</td>
<td>4.9</td>
</tr>
<tr>
<td>Methods</td>
<td>4.2</td>
<td>2.4</td>
</tr>
<tr>
<td>Presentation</td>
<td>4.3</td>
<td>2.6</td>
</tr>
<tr>
<td>Summary</td>
<td>3.2</td>
<td>1.8</td>
</tr>
</tbody>
</table>
**Assessment: How can you avoid biased selection of studies?**

- Assessment and selection should be: Standardized "Objective" OR Blinded to Results
- Cochrane Handbook has appraisal 'Risk of Bias' guide

* assessment of quality blind to study outcome

---

**What is a meta-analysis?**

*Optional part of a systematic review*

- Systematic reviews
- Meta-analyses

---

**| Study | Self-management | Control | OR (fixed) | Weight | OR (random) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Self 203</td>
<td>5.10</td>
<td>9,680</td>
<td>0.71</td>
<td>5.14</td>
<td>0.65</td>
</tr>
<tr>
<td>Self 202</td>
<td>5.10</td>
<td>9,680</td>
<td>0.71</td>
<td>5.14</td>
<td>0.65</td>
</tr>
<tr>
<td>Self 201</td>
<td>5.10</td>
<td>9,680</td>
<td>0.71</td>
<td>5.14</td>
<td>0.65</td>
</tr>
<tr>
<td>Self 200</td>
<td>5.10</td>
<td>9,680</td>
<td>0.71</td>
<td>5.14</td>
<td>0.65</td>
</tr>
</tbody>
</table>

- At the bottom there's a horizontal line. This is the scale measuring the treatment effect.
- The vertical line in the middle is where the treatment and control have the same effect - there is no difference between the two.
For each study there is an id. The data for each trial are here, divided into the experimental and control groups. This is the % weight given to this study in the pooled analysis.

The label above the graph tells you what statistic has been used. The data shown in the graph are also given numerically.

The pooled analysis is given a diamond shape, where the widest bit in the middle is located at the calculated best guess (point estimate), and the horizontal width is the confidence interval.

Note on interpretation
If the confidence interval crosses the line of no effect, this is equivalent to saying that we have found no statistically significant difference in the effects of the two interventions.
Weighting studies

- More weight to the studies which give us more information
  - More participants
  - More events
  - More precision

- Weight is proportional to the precision

Meta-analysis (Forest) plot

* Are the results similar across studies? 3 tests
  - Eyeball test — do they look the same?
  - Test of "Null hypothesis" of no variation (p-value)
  - Proportion of variation not due to chance (I²)

Are these trials different?

Risk of SIDS and sleeping position

Transferable? Use in my patients

Is the AVERAGE effect similar across studies?

- If NO, then WHY?
  - Study methods - biases
  - PICO

- If YES, then 2 questions
  - Effect in different individuals?
  - Which version of treatment?
Cumulative meta-analysis

When did we know that sleeping position affected mortality?

Conclusion

EBM and Systematic Review

- EBM (quick & dirty)
  - Ask Question
  - Search
  - Appraise
  - Apply
  - Time: 90 seconds
  - < 20 articles
  - This patient survives!

- Systematic Review
  - Ask Question
  - Search ++++ x 2
  - Appraise x 2
  - Synthesize
  - Apply
  - Time: 6 months, team
  - < 2,000 articles
  - This patient is dead

Find a systematic review! (and appraise it FAST)

Pros and cons of systematic reviews

- Advantages
  - Larger numbers & power
  - Robustness across PICOs
- Disadvantages
  - May conclude small biases are real effects

Pros and cons of systematic reviews