Network meta-analysis

LSHTM, 31st January 2014

Ian White
MRC Biostatistics Unit, Cambridge, UK
Assessing inconsistency in multiple treatments meta-analysis

LSHTM
Wednesday 24 November 2010

Ian White
MRC Biostatistics Unit, Cambridge

Have you heard it all before?

No, thanks to:

Maurice Belz
(1897–1975)
Plan

- Systematic review
- Meta-analysis
- Indirect comparisons
- Network meta-analysis
  - models allowing for heterogeneity
  - models allowing for inconsistency
  - model estimation
  - examples
  - controversies
Systematic review

- Define a clinical question
  - typically: how good is this intervention? (often drugs, but also e.g. psychological therapy)
- Obtain all papers relevant to the question using a systematic search strategy
  - typically restricted to randomised controlled trials (RCTs)
- Record study characteristics including quality
- Extract quantitative study results
- If appropriate, perform a statistical summary

statistician’s roles

See e.g.
- The Cochrane Collaboration, http://www.cochrane.org/
Pairwise meta-analysis: data from 15 randomised trials

<table>
<thead>
<tr>
<th>study</th>
<th>dA</th>
<th>nA</th>
<th>dC</th>
<th>nC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>140</td>
<td>23</td>
<td>140</td>
</tr>
<tr>
<td>6</td>
<td>75</td>
<td>731</td>
<td>363</td>
<td>714</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>106</td>
<td>9</td>
<td>205</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>549</td>
<td>237</td>
<td>1561</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>33</td>
<td>9</td>
<td>48</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>100</td>
<td>31</td>
<td>98</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>31</td>
<td>26</td>
<td>95</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>39</td>
<td>17</td>
<td>77</td>
</tr>
<tr>
<td>13</td>
<td>95</td>
<td>1107</td>
<td>134</td>
<td>1031</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>187</td>
<td>35</td>
<td>504</td>
</tr>
<tr>
<td>15</td>
<td>78</td>
<td>584</td>
<td>73</td>
<td>675</td>
</tr>
<tr>
<td>16</td>
<td>69</td>
<td>1177</td>
<td>54</td>
<td>888</td>
</tr>
<tr>
<td>17</td>
<td>64</td>
<td>642</td>
<td>107</td>
<td>761</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>62</td>
<td>8</td>
<td>90</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>234</td>
<td>34</td>
<td>237</td>
</tr>
</tbody>
</table>

**Aim** is to compare effectiveness of individual counselling (“C”) with no contact (“A”) in helping smokers to quit.

**Data** in arm A, C:
- dA, dC = # who quit smoking
- nA, nC = # randomised

Data display: Forest plot

Forest plot shows odds ratio (95% confidence interval) for C vs. A for each of the 15 studies.

Shaded blocks represent amount of information (area $\propto 1/se^2$)
Pairwise meta-analysis:
“fixed-effect” model

• Say we’re interested in the log odds ratio
• Assume there is a “true log odds ratio” \( \mu \)
• Express the results from study \( i \) as
  - \( y_i \) = estimated log odds ratio
  - \( s_i \) = its standard error
• Model: \( y_i \sim N(\mu, s_i^2) \)
  - approximation, valid for moderate/large counts

• (We are using a two-stage estimation procedure: compute the \( y_i \), then estimate \( \mu \). We can also do one-stage estimation – see later.)
Forest plot again

Note the high degree of heterogeneity between studies.

Ideally we’d explain it – e.g. if study 6 was in people who had just had a major diagnosis.

But often we need to model it instead.
Pairwise meta-analysis: random-effects model

- Model for “true log odds ratio in study i”: \( \mu_i \sim N(\mu, \tau^2) \)
- Parameters of interest:
  - \( \mu \) is the overall mean treatment effect
  - \( \tau^2 \) is the between-studies (heterogeneity) variance

- Two-stage estimation procedure
- Model for point estimate: \( y_i \sim N(\mu_i, s_i^2) \)
  - \( y_i \) = estimated log odds ratio in study \( i \)
  - \( s_i \) = its standard error
- Estimate \( \tau^2 \) (and hence \( \mu \)) by
  - method of moments – very popular
  - or restricted maximum likelihood (REML)
Forest plot showing meta-analysis result

The random-effects analysis gives an estimate of the overall mean allowing for heterogeneity and a prediction interval (effect in a new study).
A note on terminology

• “Fixed-effect” is unfortunate terminology
• Elsewhere in statistics, “fixed effects” means lots of free parameters
• Could consider $y_i \sim N(\mu_i, \sigma_i^2)$ with these 3 models:

<table>
<thead>
<tr>
<th>Model</th>
<th>Standard name</th>
<th>Better name?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_i = \mu$</td>
<td>Fixed-effect</td>
<td>Common effect</td>
</tr>
<tr>
<td>$\mu_i \sim N(\mu, \tau^2)$</td>
<td>Random-effects</td>
<td>Random effects</td>
</tr>
<tr>
<td>$\mu_i$ all separate</td>
<td>(not used)</td>
<td>Fixed effects</td>
</tr>
</tbody>
</table>
Other issues in (pairwise) meta-analysis

- Study quality
- Study-level covariates → “meta-regression”
- Publication bias
  - small trials more likely to be published if they show statistically significant effects?
  - see next
Exploring publication bias: “funnel plot”
Actually the data are more complicated ...

<table>
<thead>
<tr>
<th>study</th>
<th>dA</th>
<th>nA</th>
<th>dB</th>
<th>nB</th>
<th>dC</th>
<th>nC</th>
<th>dD</th>
<th>nD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>140</td>
<td>23</td>
<td>140</td>
<td>10</td>
<td>138</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>78</td>
<td>12</td>
<td>85</td>
<td>29</td>
<td>170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>79</td>
<td>702</td>
<td>77</td>
<td>694</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>671</td>
<td>21</td>
<td>535</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>116</td>
<td>19</td>
<td>146</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>75</td>
<td>731</td>
<td>363</td>
<td>714</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>106</td>
<td>9</td>
<td>205</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>549</td>
<td>237</td>
<td>1561</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>33</td>
<td>9</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>100</td>
<td>31</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>31</td>
<td>26</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>39</td>
<td>17</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>95</td>
<td>1107</td>
<td>134</td>
<td>1031</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>187</td>
<td>35</td>
<td>504</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>78</td>
<td>584</td>
<td>73</td>
<td>675</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>69</td>
<td>1177</td>
<td>54</td>
<td>888</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>64</td>
<td>642</td>
<td>107</td>
<td>761</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>62</td>
<td>8</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>234</td>
<td>34</td>
<td>237</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>20</td>
<td>9</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>49</td>
<td>16</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>7</td>
<td>66</td>
<td>32</td>
<td>127</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>12</td>
<td>76</td>
<td>20</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>9</td>
<td>55</td>
<td>3</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

24 trials compared 4 different interventions to help smokers quit:
A="No contact"
B="Self help"
C="Individual counselling"
D="Group counselling"
Actually the data are more complicated ...

<table>
<thead>
<tr>
<th>study</th>
<th>dA</th>
<th>nA</th>
<th>dB</th>
<th>nB</th>
<th>dC</th>
<th>nC</th>
<th>dD</th>
<th>nD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>140</td>
<td>23</td>
<td>140</td>
<td>10</td>
<td>138</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>11</td>
<td>78</td>
<td>12</td>
<td>85</td>
<td>29</td>
<td>170</td>
</tr>
<tr>
<td>3</td>
<td>79</td>
<td>702</td>
<td>12</td>
<td>85</td>
<td>29</td>
<td>170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>671</td>
<td>21</td>
<td>535</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>116</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>75</td>
<td>731</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>363</td>
<td>714</td>
</tr>
<tr>
<td>7</td>
<td>79</td>
<td>702</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>85</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>549</td>
<td></td>
<td></td>
<td>237</td>
<td>1561</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>671</td>
<td></td>
<td></td>
<td>21</td>
<td>535</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>116</td>
<td></td>
<td></td>
<td>19</td>
<td>146</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td>77</td>
</tr>
<tr>
<td>13</td>
<td>75</td>
<td>731</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>363</td>
<td>714</td>
</tr>
<tr>
<td>14</td>
<td>58</td>
<td>549</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>237</td>
<td>1561</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td>77</td>
</tr>
<tr>
<td>17</td>
<td>58</td>
<td>549</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>237</td>
<td>1561</td>
</tr>
<tr>
<td>18</td>
<td>15</td>
<td>187</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td>77</td>
</tr>
<tr>
<td>19</td>
<td>64</td>
<td>642</td>
<td></td>
<td></td>
<td>107</td>
<td>761</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>62</td>
<td></td>
<td></td>
<td>34</td>
<td>237</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>76</td>
</tr>
<tr>
<td>22</td>
<td>20</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>76</td>
</tr>
<tr>
<td>23</td>
<td>9</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>20</td>
</tr>
</tbody>
</table>

We have trials of different designs:
- A vs C vs D
- B vs C vs D
- A vs B (x3)
- A vs C (x14)
- A vs D
- B vs C
- B vs D
- C vs D (x2)
Evidence network: the smoking data

14 trials compared A with C
“design AC”
1 trial compared A, C and D
“design ACD”

etc.
Indirect comparisons

- Let’s now focus on comparing B with C.
- Evidence from B vs C and B vs C vs D trials is “direct evidence.”
- Can we use indirect evidence to compare B with C?
  - e.g. combining A vs B trials with A vs C trials.
- The maths is easy: using indirect evidence only, $\hat{\delta}_{BC} = \hat{\delta}_{AC} - \hat{\delta}_{AB}$
  with variance $\text{var}(\hat{\delta}_{BC}) = \text{var}(\hat{\delta}_{AC}) + \text{var}(\hat{\delta}_{AB})$
  - where $\hat{\delta}_{BC}$ = effect of C compared to B, etc.
- But the assumptions are tricky: must assume the 3 designs (A vs B, A vs C, B vs C) are comparable.
Bias in indirect comparisons (1)

- Suppose B and C are equally beneficial compared to A
  - B was trialled in the 1990s in a wide range of smokers
  - C was trialled in the 2000s in smokers who had failed in previous quit attempts
- So C is likely to show smaller benefit than B
- Quit rates might be:

<table>
<thead>
<tr>
<th>Trial</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990s</td>
<td>10%</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>2000s</td>
<td>10%</td>
<td>15%</td>
<td>15%</td>
</tr>
</tbody>
</table>

- But what if all 3 interventions had been tried?
- Can regard C in 1990s and B in 2000s as “missing groups” – and data are missing not at random
Bias in indirect comparisons (2)

- If the overall event rates differ, then there are also problems with the scale on which intervention effects are measured. Suppose:

<table>
<thead>
<tr>
<th>Trial</th>
<th>A vs. B</th>
<th>A vs. C</th>
<th>Comparison with A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk difference</td>
<td>Risk ratio</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>A vs. B</td>
<td>20%</td>
<td>30%</td>
<td>+10%</td>
</tr>
<tr>
<td>A vs. C</td>
<td>10%</td>
<td>18%</td>
<td>+8%</td>
</tr>
</tbody>
</table>

B best | C best | C best

- Extrapolation problem – no easy answer
Network meta-analysis

- Despite these problems, I’ll proceed to combine all the evidence – indirect and direct – in order to get our best estimates of the value of all the interventions
- This is called network meta-analysis
  - multiple treatments meta-analysis
  - mixed treatment comparisons
- Network meta-analysis addresses the real clinical question: which intervention is best for the patient?
  - may additionally require modelling covariates
- Much used by NICE (National Institute for Clinical Excellence) in comparing interventions

Aims of network meta-analysis

1. **Use all the data** & thus get
   - better estimates of treatment effects
   - opportunity to identify the best treatment

2. Assess whether the evidence is **consistent**
   - i.e. does the indirect evidence agree with the direct evidence?

The main statistical challenges are
   - formulating and fitting models that allow for heterogeneity and inconsistency
   - assessing inconsistency and (if found) finding ways to handle it

Less-statistical challenges include defining the scope of the problem: which treatments to include, what patient groups, what outcomes
Models for network meta-analysis: consistency model (1)

- Trials have different baseline risks: no assumptions on $\alpha_i$ ("fixed effects" for trial)
- Between-trials model: $\mu_i = (\mu_{iB}, \mu_{iC}) \sim N(\mu, \Sigma)$
  - heterogeneity (variation between trials): $\Sigma \neq 0$ ("random effects" for treatment*trial)
- Consistency: $\mu_i$ has same mean $\mu = (\mu_B, \mu_C)$ in each design, where $\mu_B, \mu_C$ = average effect of B, C vs A

<table>
<thead>
<tr>
<th>Design</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>$\alpha_i$</td>
<td>$\alpha_i + \mu_{iB}$</td>
<td>$\alpha_i + \mu_{iC}$</td>
</tr>
<tr>
<td>AB</td>
<td>$\alpha_i$</td>
<td>$\alpha_i + \mu_{iB}$</td>
<td>-</td>
</tr>
<tr>
<td>AC</td>
<td>$\alpha_i$</td>
<td>-</td>
<td>$\alpha_i + \mu_{iC}$</td>
</tr>
</tbody>
</table>
Models for network meta-analysis: consistency model (2)

True log odds in each group in trial \( i \)

<table>
<thead>
<tr>
<th>Design</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>( \alpha_i )</td>
<td>( \alpha_i + \mu_{iB} )</td>
<td>( \alpha_i + \mu_{iC} )</td>
</tr>
<tr>
<td>AB</td>
<td>( \alpha_i )</td>
<td>( \alpha_i + \mu_{iB} )</td>
<td>-</td>
</tr>
<tr>
<td>AC</td>
<td>( \alpha_i )</td>
<td>-</td>
<td>( \alpha_i + \mu_{iC} )</td>
</tr>
<tr>
<td>BC</td>
<td>( \alpha_i )</td>
<td>( \alpha_i + \mu_{iB} )</td>
<td>( \alpha_i + \mu_{iC} )</td>
</tr>
</tbody>
</table>

- What about trials with no arm A?
- Easiest to regard arm A in BC trials as “missing data”
- Design BC still contributes to estimating \( \mu_C - \mu_B \)
Full consistency model

- Notation:
  - interventions A (reference), B, C, D, ...
  - effect of intervention J vs. A:
    » estimate (from data) $y_{ij} \rightarrow y_i = (y_{iB}, y_{iC}, y_{iD}, ...)$
    » study-specific mean $\mu_{ij} \rightarrow \mu_i = (\mu_{iB}, \mu_{iC}, \mu_{iD}, ...)$
    » overall mean $\mu_j \rightarrow \mu = (\mu_B, \mu_C, \mu_D, ...)$
  - estimated variance-covariance matrix of $y_i$ is $S_i$

- Within-trial model: $y_i \sim N(\mu_i, S_i)$
- Between-trials model: $\mu_j \sim N(\mu, \Sigma)$
- Doesn’t matter that some $y_{ij}$ are missing

- This is a contrast-based model, cf. an arm-based model for summary outcomes $y_i^* = (y_{iA}^*, y_{iB}^*, y_{iC}^*, y_{iD}^*, ...)$
**Inconsistency model**

Inconsistency: treatment effects differ across designs
- “design-by-treatment interaction”
- regard the \( \omega \)'s as fixed effects

Heterogeneity

- Many networks are sparse
- e.g. a network meta-analysis of 8 thrombolytic treatments for AMI:

```
A Streptokinase
B Accelerated alteplase
C Alteplase
D = A + C
E Tenecteplase
F Reteplase
G Urokinase
H Anti-streptilase
```

2-arm trials
3-arm trials
Heterogeneity models

• Why does sparseness matter?

• Because between-trials variance $\Sigma = \text{var}(\mu_i)$ includes unidentified terms
  - e.g. $\text{var}(\mu_{iD} - \mu_{iE})$ and hence $\text{cov}(\mu_{iD}, \mu_{iE})$ isn’t identified without a D-E trial
  - nor is $\text{cov}(\mu_{iB}, \mu_{iE})$ with only 1 B-E trial

• Need modelling assumptions for $\Sigma$

• Commonest is “common heterogeneity assumption”: $\text{var}(\mu_{ij} - \mu_{i\ell}) = \tau^2$ for all treatment pairs $(I,J)$
Network meta-analysis: standard model

- Let $y_{di}^{IJ}$ be the estimated log odds ratio (or other measure) for treatment J vs. I in study i with design d
- Let $s_{di}^{IJ}$ be its standard error
- Consistency model: $y_{di}^{IJ} \sim N(\mu_{di}^{IJ},(s_{di}^{IJ})^2)$ ← approximation
  where $\mu_{di}^{IJ} \sim N(\delta^J - \delta^I, \tau^2)$
- $\delta^J$ is the mean effect of J vs. reference treatment A
  - we make sure that results don’t depend on the choice of reference treatment
- $\tau^2$ is the common heterogeneity (between-studies) variance
- Inconsistency model: $\mu_{di}^{IJ} \sim N(\delta^J - \delta^I + \omega_{di}^{IJ}, \tau^2)$
  - true treatment effects are different in every design
Network meta-analyses: estimation

- In the past, the models have been fitted using WinBUGS
  - because frequentist alternatives have not been available
  - has made network meta-analysis difficult for non-statisticians
- Now, consistency and inconsistency models can be fitted using multivariate meta-analysis and multivariate meta-regression
- Trials without the reference intervention are handled
  - by a trial-specific baseline intervention (complicates code); or
  - by “augmenting” these trials with a very small reference arm (e.g. 0.0001 successes out of 0.001)
Network meta-analysis: multi-arm trials

- Multi-arm trials contribute >1 log odds ratio
  - need to allow for their covariance
  - mathematically straightforward but complicates programming
- With only 2-arm trials, we can fit models using standard “meta-regression”
- Multi-arm trials complicate this – need suitable data formats and multivariate analysis
Example analyses
Smoking network

Log odds ratio

Study 1
Study 2
Study 3
Study 4
Study 5
Study 6
Study 7
Study 8
Study 9
Study 10
Study 11
Study 12
Study 13
Study 14
Study 15
Study 16
Study 17
Study 18
Study 19
Study 20
Study 21
Study 22
Study 23
Study 24

B vs. A
C vs. A
D vs. A
C vs. B
D vs. B
D vs. C
Smoking network: results

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Odds ratio (95% CI)</th>
<th>P(best)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (no contact)</td>
<td>1 (reference)</td>
<td>0.0%</td>
</tr>
<tr>
<td>B (self help)</td>
<td>1.49 (0.78-2.85)</td>
<td>3.1%</td>
</tr>
<tr>
<td>C (individual counselling)</td>
<td>2.02 (1.37-2.98)</td>
<td>31.9%</td>
</tr>
<tr>
<td>D (group counselling)</td>
<td>2.38 (1.14-4.97)</td>
<td>65.0%</td>
</tr>
</tbody>
</table>

- Between-trials SD on log OR scale: $\hat{t} = 0.674$ (large)
- D or C is likely to be best
- Test of inconsistency (Wald test in design-by-treatment interaction model): $\chi^2=5.11$ on 7 df, $p=0.65$
Smoking network

Test of consistency: chi2(7) = 5.11, P = 0.646
Thrombolitics network: results

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Odds ratio (95% CI)</th>
<th>P(best)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (streptokinase)</td>
<td>1 (reference)</td>
<td>0.0%</td>
</tr>
<tr>
<td>B (accelerated alteplase)</td>
<td>0.85 (0.78-0.93)</td>
<td>19.3%</td>
</tr>
<tr>
<td>C (alteplase)</td>
<td>1.00 (0.94-1.07)</td>
<td>0.1%</td>
</tr>
<tr>
<td>D (=A+C)</td>
<td>0.96 (0.87-1.05)</td>
<td>0.5%</td>
</tr>
<tr>
<td>E (tenecteplase)</td>
<td>0.86 (0.73-1.00)</td>
<td>22.4%</td>
</tr>
<tr>
<td>F (reteplase)</td>
<td>0.89 (0.79-1.01)</td>
<td>6.8%</td>
</tr>
<tr>
<td>G (urokinase)</td>
<td>0.82 (0.53-1.27)</td>
<td>50.9%</td>
</tr>
<tr>
<td>H (anti-streptilase)</td>
<td>1.01 (0.94-1.10)</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

- Between-trials SD on log OR scale: $\hat{t}=0.015$ (small)
- B, E or G is probably best
- Test of inconsistency: $\chi^2=8.61$ on 8 df, $p=0.38$
Thrombolytics network

- Studies
- Pooled overall
- Pooled within design

Test of consistency: $\chi^2(8)=8.61, P=0.377$
Some controversies
Controversies: what data to extract?

• Both my examples have summarised each study as a 2x2 table: successes/total in each arm
  – the standard in Cochrane systematic reviews
  – has the advantage of avoiding authors’ tendency to “cherry-pick” the best results

• An alternative is to use the estimated treatment effect(s) in each trial’s report
  – may be adjusted for prognostic factors (increases power in RCTs)
  – essential in observational studies (where we have to trust the authors to adjust for confounders)
Controversies: are published data enough?

- Published data have limitations
- The ideal is to get the raw data from all studies (individual participant data, IPD)
- IPD is especially valuable when exploring phenomena which tend to be inconsistently analysed / reported:
  - interactions (subgroup effects)
  - adjustment for confounding in observational studies
- But it is much slower and much more expensive...
Controversies: the common heterogeneity model

- The common heterogeneity model assigns heterogeneity even when a contrast is estimated in a single study (e.g. B-E in thrombolytics) – must be good.
- But homogeneous parts of the network may become “contaminated” by more heterogeneous parts.
  - could in principle have:

![Network diagram](Image)

<table>
<thead>
<tr>
<th>Pairwise, B vs A:</th>
<th>Pairwise, C vs A:</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR = 0.8 (95% CI, 0.7-0.9)</td>
<td>OR = 1 (95% CI, 0.5-2.0)</td>
</tr>
<tr>
<td>( \hat{\tau}^2 = 0 )</td>
<td>( \hat{\tau}^2 = 2 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network, B vs A:</th>
<th>Network, C vs A:</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR = 0.8 (95% CI, 0.5-1.3)</td>
<td>OR = 1 (95% CI, 0.6-1.7)</td>
</tr>
<tr>
<td>( \hat{\tau}^2 = 1 ) “unfair”?!</td>
<td>( \hat{\tau}^2 = 1 )</td>
</tr>
</tbody>
</table>

Ideally want a model with \( \tau^2 \) exchangeable across comparisons
Controversies: defining inconsistency

- Our “design-by-treatment interaction model” has 3 inconsistency parameters
- Intuitively, should be only one per “loop”
  - but we haven’t found a sensible way to define it
  - model of Lu & Ades (2006) isn’t symmetrical with multi-arm trials

<table>
<thead>
<tr>
<th>Design</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>$\alpha_i$</td>
<td>$\alpha_i + \mu_{iB}$</td>
<td>$\alpha_i + \mu_{iC}$</td>
</tr>
<tr>
<td>AB</td>
<td>$\alpha_i$</td>
<td>$\alpha_i + \mu_{iB} + \omega_1$</td>
<td>-</td>
</tr>
<tr>
<td>AC</td>
<td>$\alpha_i$</td>
<td>-</td>
<td>$\alpha_i + \mu_{iC} + \omega_2$</td>
</tr>
<tr>
<td>BC</td>
<td>$\alpha_i$</td>
<td>$\alpha_i + \mu_{iB}$</td>
<td>$\alpha_i + \mu_{iC} + \omega_3$</td>
</tr>
</tbody>
</table>
Controversies: testing for inconsistency

- Test for inconsistency is a global test on many degrees of freedom
  - likely to have low power in practice
- Can we use substantive knowledge to define more targeted tests?
- Should we accept that inconsistency is present even when test is non-significant?
Controversies: allowing for inconsistency

What do we do if we decide we have inconsistency? Obviously we first try to explain it – “did the A-B trials recruit more severely ill patients?”, etc.

If we fail, then do we

- refuse to draw conclusions about treatment comparisons? (maybe we asked the wrong question?)
- infer treatment comparisons from the consistency model, with appropriate caveats?
- treat inconsistency as another random effect?
  - we’ve proposed a model for this (Jackson et al, under review)
  - it inflates std errors to “account for” inconsistency
  - just as the standard random-effects model inflates std errors to “account for” heterogeneity.
Controversies: estimation

- Network meta-analysis was in the past done using Bayesian methods (1-stage analysis, arm-based model, full binomial likelihood)
  - WinBUGS
  - rank treatments, give \( p(\text{treatment C is best}) \) etc.
- I’ve proposed frequentist methods based on multivariate meta-analysis (2-stage analysis, contrast-based model, Normal approximation to the likelihood)
  - faster and more accessible
  - don’t allow well for sparse binary data (e.g. smoking trial 9: 0/33 vs 9/48)
- Next slide compares the methods in the smoking data...
Smoking network: method comparison

<table>
<thead>
<tr>
<th>log OR: treatment vs. A</th>
<th>Two-stage frequentist</th>
<th>One-stage Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>std err</td>
</tr>
<tr>
<td>A (ref)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>0.398</td>
<td>0.331</td>
</tr>
<tr>
<td>C</td>
<td>0.702</td>
<td>0.199</td>
</tr>
<tr>
<td>D</td>
<td>0.866</td>
<td>0.376</td>
</tr>
<tr>
<td>( \tau ): between trials SD</td>
<td>0.674</td>
<td>0.140</td>
</tr>
</tbody>
</table>

- One-stage Bayesian analysis taken from Lu & Ades, JASA 2006; 101: 447–459.
- Differences between methods are mainly attributable to the approximation in the two-stage method.
Why is the two-stage method inaccurate?

- Because the standard error is correlated with the point estimate
  - more extreme estimates are down-weighted, causing bias towards null
- Problem appears to be restricted to binary data
A frequentist one-stage method for binary data?

• Should be able to fit a generalised linear mixed model (Stata `melogit`)
  – random effect for study*treatment interaction
  – (± fixed or random effect for design*treatment interaction)

• How do we handle main effect of study?
  – fixed effect? → one parameter per study → may underestimate heterogeneity variance & std error
  – random effect? but then results are contaminated by between-study information
  – eliminate it by conditioning on study margins? may be ideal but computationally difficult

Controversies: ranks

- Rankogram displays the posterior probability that each treatment is
  - ranked 1 (the best), ≤2, ≤3 etc.
- The argument is
  - a clinician wants to use the best treatment, so we maximise their chances
  - if best treatment isn’t available, want to maximise their chance of getting the 2\textsuperscript{nd} best

Controversies: ranks

- But is this the right way to choose a treatment?
- Decision theory suggests choosing the treatment which maximises the expected utility, e.g. $p(\text{quit smoking} \mid \text{treatment})$
  - would take account of uncertainty
  - best would depend on “baseline risk” $p(\text{quit smoking} \mid \text{no treatment})$
Resources

- Bayesian approach using WinBUGS: the NICE decision support unit has a series of useful documents at http://www.nicedsu.org.uk/Evidence-Synthesis-TSD-series%282391675%29.htm
- Frequentist approach using Stata: I have written network, a suite of programs to read in data, fit consistency and inconsistency models, and graph results
  - the consistency and inconsistency models are expressed as multivariate meta-analyses / meta-regressions and fitted using my mvmeta
  - net from http://www.mrc-bsu.cam.ac.uk/IW_Stata/
- Frequentist approach using R: Antonio Gasparrini has written an R counterpart to mvmeta
Network meta-analysis: summary

<table>
<thead>
<tr>
<th>Interpretation</th>
<th>Estimation</th>
<th>Model for</th>
<th>Extracting data</th>
<th>Identifying relevant papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>best treatment / decision theory</td>
<td>bayesian: exact likelihood, 1-stage, arm-based</td>
<td>study effect</td>
<td>2x2 table / treatment effect / IPD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>frequentist: 2-stage + normal approx? contrast-based?</td>
<td>treatment effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>heterogeneity</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>inconsistency</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>covariates</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>quality ...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thanks to Julian Higgins (U of Bristol), Dan Jackson (BSU) and Jessica Barrett (U of Cambridge) who worked with me on this.