Statistical assessment of consistency of treatment effect in multiregional clinical trials

Kevin J Carroll
Content and flow

• Background and key questions
• Design MRCT
• Analysis MRCT
• Examples
• Summary
Background

• Modern drug development is impossible without the use of MRCTs

• However, there are challenges as reflected in ICH E5 relating to possible inconsistency of effects from region to region
  • ‘intrinsic’ = race, biology, genetics
  • ‘extrinsic’ = cultural, social, economic, medical practice, quality of care, experience in clinical trials, quality of trial conduct and monitoring

• With more recent regulatory guidelines from PMDA and EMA together with notable test cases, MRCTs subject to a lot of attention within and outside of statistics
Content and flow

- Background and key questions
- Design MRCT
- Analysis MRCT
- Examples
- Summary
Construct

- Trial designed to detect a true overall effect $\theta$ with $\alpha$ 1-sided Type I error and power $1-\beta$. The information content is therefore

$$V = \frac{(z_\alpha + z_\beta)^2}{\theta^2}$$

- $i=1$ to $r$ regions each with a fraction $f_i$ patients. The estimated treatment effect estimate in $i^{th}$ region is therefore

$$\hat{\theta}_i \sim N\left(\theta, \frac{1}{f_i V}\right)$$

so that the overall estimate

$$\hat{\theta} \sim \sum_{i=1}^{r} f_i \hat{\theta}_i \sim N\left(\theta, \frac{1}{V}\right)$$

- However, if regional effects are considered as random then $\theta_i \sim N(\theta, \tau^2)$

so that

$$\hat{\theta}_i \sim N\left(\theta, \tau^2 + \frac{1}{f_i V}\right)$$

and

$$\hat{\theta} \sim \sum_{i=1}^{r} f_i \hat{\theta}_i \sim N\left(\theta, \tau^2 \sum_{i=1}^{r} f_i^2 + \frac{1}{V}\right)$$
Hung (2010)

• If, in truth, $\tau^2 > 0$, then Type I error will be inflated

$$1 - \Phi \left[ z_\alpha \left( 1 + V\tau^2 \sum_{i=1}^{r} f_i^2 \right)^{-0.5} \right]$$

• $V$ should be increased to

$$\tilde{V} = \left( \frac{\theta^2}{(z_\alpha + z_\beta)^2} - \tau^2 \sum_{i=1}^{r} f_i^2 \right)^{-1}$$

so that

$$\frac{V}{\tilde{V}} = 1 - \frac{\tau^2 (z_\alpha + z_\beta)^2}{\theta^2} \sum_{i=1}^{r} f_i^2$$

and power = \Phi \left[ -z_\alpha + (z_\alpha + z_\beta) \left( 1 + V\tau^2 \sum_{i=1}^{r} f_i^2 \right)^{-0.5} \right]

• For $V = \frac{N}{\sigma^2}$ sample size inflation is:

![Graph showing sample size ratio N/N₀ versus (δ/σ).](image)
Type I error inflation for $V=N/\sigma^2$, $r=3$, $f_i=1/3$, $\beta=0.1$
Issues and Questions (1)

- Ignoring $\tau^2$ increases Type I error and reduces power.
- Should N in an MRCT be routinely increased, allowing for $\tau^2$?
- In trial planning, what value for $\tau^2$ (or $\tau / \sigma$) should be assumed? On what is the choice based? Is it logical to assume $\tau^2 > 0$?
- At what value of $\tau / \sigma$ does a MRCT become infeasible in terms of trial size?
  - impossible to interpret clinically due to excessive between region variation in treatment effect?
Issues and Questions (2)

• How do we define ‘region’? Should there be a regulatory standard agreed to cover all trials?

• What should be the allocation of N across regions / countries? How do we determine This?

• What is meant by ‘consistency’? How do we define this? How do we assess it? What is the value and role of routine homogeneity testing of regional results? And graphical methods?

• Should a random effects analysis be the standard in MRCTs? What are the consequences if so?
What allocation of patient to regions?

• If we want to observe a positive treatment effect in each region, then
  \[ \Pr(\hat{\theta}_i > 0, \forall i = 1 \text{ to } r) = \prod_{i=1}^{r} \Phi\left(\sqrt{f_i(z_\alpha + z_\beta)}\right) \]

• But need also to observe an overall significant effect, so then we want
  \[ \Pr(\hat{\theta}_i > 0, \forall i = 1 \text{ to } r \cap \hat{\theta} > z_\alpha V^{-0.5}) \]
  where
  \[
  \left(\hat{\theta}, \hat{\theta}_1, ..., \hat{\theta}_r\right) \sim N(\theta, \Lambda)
  \]

\[
\Lambda = \begin{bmatrix}
\frac{1}{V} & \frac{1}{V} & \frac{1}{V} \\
\frac{1}{V} & \frac{1}{V} & 0 \\
\frac{1}{V} & \frac{1}{V} & 0 \\
\end{bmatrix}
\]

• Chuang (2008) considers the most typical MRCT with 3 regions, to reflect US, EU and Japan concluding the smallest region could be as low as 15% and yet still achieve \(= 0.80\).

\[
\Pr(\hat{\theta}_i > 0, \forall i = 1 \text{ to } r \cap \hat{\theta} > z_\alpha V^{-0.5})
\]
Pr observing a positive treatment effect in each region
Pr observing a positive treatment effect in each
Pr observing a positive treatment effect in each

Pr exceeds 0.85 so long as smallest region is > 15%

Pr exceeds 0.90 so long as smallest region is > 25%
Pr observing a positive treatment effect in each region and $p<0.05$ for the overall effect

Based on 10,000 simulations for each combination of $p_1$ and $p_2$
Pr observing a positive treatment effect in each region and $p<0.05$ for the overall effect

Based on 10,000 simulations for each combination of $p_1$ and $p_2$

Pr exceeds 0.80 so long as smallest region is > 15%

Pr no longer can reach 0.90
Looking at the ‘reference’ region of interest

- Denote $\hat{\theta}_{\text{ref}} \sim N\left(\theta, \frac{1}{fv}\right)$ and $\hat{\theta}_{\text{nref}} \sim N\left(\theta, \frac{1}{(1-f)v}\right)$ then consistency between the reference region and non-reference regions may be defined in several ways:

1. $\hat{\theta}_{\text{ref}} - 0.5\hat{\theta} > 0$
2. $\hat{\theta}_{\text{ref}} - 0.5\hat{\theta}_{\text{nref}} > 0$
3. $\left(\hat{\theta}_{\text{ref}} - \hat{\theta}_{\text{nref}}\right) + \frac{\theta}{2} > 0$
4. $\hat{\theta}_{\text{ref}} - \left(\hat{\theta} - z_\alpha V^{-0.5}\right) > 0$
5. $\hat{\theta}_{\text{ref}} > 0$ and $\hat{\theta}_{\text{nref}} > 0$

- Denoting criteria as $C_j$ then $\Pr(\text{meeting criteria and achieving } \hat{\theta} > z_\alpha V^{-0.5})$ can be determined by noting

\[
\begin{pmatrix}
C_1 \\
\hat{\theta}
\end{pmatrix} \sim N\left(\theta(1-k), \frac{1}{fv} + \frac{k(k-2)}{V} \frac{1-k}{V} \right)
\]

\[
\begin{pmatrix}
C_2 \\
\hat{\theta}
\end{pmatrix} \sim N\left(\theta(1-k), \frac{1}{fv} - \frac{k^2}{V(1-f)} \frac{1-k}{V} \right)
\]

\[
\begin{pmatrix}
C_3 \\
\hat{\theta}
\end{pmatrix} \sim N\left(z_\alpha V^{-0.5}, \frac{(1-f)}{fv} + \frac{1}{V} \right)
\]
Pr(meeting criteria 1 to 5)

![Graph showing the probability of meeting criteria 1 to 5 against the fraction of patients in the reference region.](image-url)
So what?

• Of these possible alternative criteria, only criterion #4 represents some improvement over the MHLW guideline criterion in terms of the fraction of reference region patients required in a MRCT.

• Criterion #4 requires around 1/3 fewer patients than Criterion #1 for about the same overall power.
“Effect Reversals”, where the treatment effect is positive overall but numerically negative in some regions are to be expected in a large multiregional trials

In a trial with 90% (80%) power and 7 (6) regions, the probability of at least one effect reversal >50%.

Content and flow

• Background and key questions
• Design MRCT
• Analysis MRCT
• Examples
• Summary
Random effects analysis – good idea?

• If regional effects are considered as random then $\theta_i \sim N(\theta, \tau^2)$ and

$$\hat{\theta}_i \sim N\left(\theta, \tau^2 + \frac{1}{f_i V}\right) \quad \text{with} \quad Q = V \sum_{i=1}^{r} f_i \left(\hat{\theta}_i - \hat{\theta}\right)^2 \sim \chi^2_{r-1}
$$

• Then $\hat{\tau}^2 = \frac{Q - (r-1)}{1 - \sum_{i=1}^{r} f_i^2}$ and $w_i = \left(\hat{\tau}^2 + \frac{1}{f_i V}\right)^{-1}$ so that

$$\hat{\theta}_{RE} \sim \sum_{i=1}^{r} w_i \hat{\theta}_i \sim N\left(\theta, \left(\sum_{i=1}^{r} w_i\right)^{-1}\right)
$$

• Example, Chen (2010):

$\hat{\theta} = 0.89$ 95% CI (0.79, 0.99), $p=0.037$

$\hat{\theta}_{RE} = 0.91$ (0.76, 1.08), $p=0.29$

<table>
<thead>
<tr>
<th>Table 1. PURSUIT: Efficacy by Region.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>Overall</td>
</tr>
<tr>
<td>North America</td>
</tr>
<tr>
<td>Western Europe</td>
</tr>
<tr>
<td>Latin America</td>
</tr>
<tr>
<td>Eastern Europe</td>
</tr>
</tbody>
</table>

N: sample size. Odds ratio: eptifibatide vs placebo, lower the better [19].
Random effects analysis – good idea?

• Example, Chen (2010):
  \( \hat{\tau}^2 = 0.016, \frac{\hat{\tau}}{\hat{\sigma}} = 1.06 \)
  \( \chi^2 = 6.3, p=0.096, I^2=0.52 \)
  \( \hat{\theta} = 0.89 \text{ CI (0.79, 0.99), } p=0.037 \)
  \( \hat{\theta}_{RE} = 0.91 \text{ CI (0.76, 1.08), } p=0.29 \)

• \( \hat{\theta}_{RE} = 0.91 (0.68, 1.21), p=0.36 \) if Follmann\(^1\) adjustment applied which means using a t value on k-1 df for \( \hat{\theta}_{RE} \) as opposed to z value. This has been advocated in the recent past by FDA for meta-analyses.

• If applied this has some odd consequences for MRCTs.

\[
\begin{array}{cccccc}
\text{fi} & \text{RR} & \text{CL} & \text{CL} & \text{p-value} \\
\hline
\text{Region 1} & 0.33 & 0.80 & 0.65 & 0.99 & 0.0416 \\
\text{Region 2} & 0.33 & 0.80 & 0.65 & 0.99 & 0.0416 \\
\text{Region 3} & 0.33 & 0.80 & 0.65 & 0.99 & 0.0416 \\
\text{All} & & 0.80 & 0.71 & 0.91 & 0.0004 \\
\text{RE} & & & & & 0.0718 \\
\end{array}
\]

1 Follmann (1999)
Content and flow

• Background and key questions
• Design MRCT
• Analysis MRCT
• Examples
• Summary
MERIT-HF

• Examined metoprolol CR/XL vs placebo in patients with chronic heart failure and decreased ejection fraction

• Randomised 3991 across 14 countries

• The study stopped early on the recommendation of the IDMC. All-cause mortality was lower in the metoprolol CR/XL group than in the placebo group (145 [7.2%] vs 217 deaths [11.0%]) RR = 0.66 [95% CI 0.53–0.81]; p=0.00009.

• However....
MERIT-HF: Inconsistent regional effects?
MERIT-HF: Mortality by region

usa
uk
hollswiz
sweden
poland
norway
iceland
hungary
germany
denfin
czech
belgium
MERIT-HF: Mortality by region vs all other regions

<table>
<thead>
<tr>
<th>Region</th>
<th>$t^2$ for vs</th>
<th>$f_i$ for vs</th>
<th>$t^2*(1-f_i)$</th>
<th>% total $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>usa</td>
<td>8.39</td>
<td>27.6%</td>
<td>6.07</td>
<td>34.4%</td>
</tr>
<tr>
<td>uk</td>
<td>0.70</td>
<td>3.6%</td>
<td>0.67</td>
<td>3.8%</td>
</tr>
<tr>
<td>hollswiz</td>
<td>0.60</td>
<td>11.0%</td>
<td>0.54</td>
<td>3.0%</td>
</tr>
<tr>
<td>sweden</td>
<td>2.74</td>
<td>3.0%</td>
<td>2.65</td>
<td>15.0%</td>
</tr>
<tr>
<td>poland</td>
<td>0.80</td>
<td>4.4%</td>
<td>0.76</td>
<td>4.3%</td>
</tr>
<tr>
<td>norway</td>
<td>0.06</td>
<td>4.7%</td>
<td>0.05</td>
<td>0.3%</td>
</tr>
<tr>
<td>iceland</td>
<td>0.35</td>
<td>1.1%</td>
<td>0.35</td>
<td>2.0%</td>
</tr>
<tr>
<td>hungary</td>
<td>0.45</td>
<td>12.4%</td>
<td>0.39</td>
<td>2.2%</td>
</tr>
<tr>
<td>germany</td>
<td>0.14</td>
<td>13.8%</td>
<td>0.12</td>
<td>0.7%</td>
</tr>
<tr>
<td>denfin</td>
<td>0.51</td>
<td>6.6%</td>
<td>0.48</td>
<td>2.7%</td>
</tr>
<tr>
<td>czech</td>
<td>0.36</td>
<td>7.2%</td>
<td>0.33</td>
<td>1.9%</td>
</tr>
<tr>
<td>belgium</td>
<td>5.48</td>
<td>4.4%</td>
<td>5.24</td>
<td>29.6%</td>
</tr>
</tbody>
</table>

$\chi^2$ p-value = 0.090

% total $\chi^2$
MERIT-HF: Mortality:
HR=0.65, 95% CI (0.49. 0.86), $\tau^2=0.038$, $s^2=0.14$, $I^2=0.21$
MERIT-HF: Mortality:
HR=0.65, 95% CI (0.49, 0.86), $\tau^2=0.038$, $s^2=0.14$, $I^2=0.21$
## MERIT-HF: Mortality vs Primary

### Mortality

<table>
<thead>
<tr>
<th>Country</th>
<th>Drug Events</th>
<th>Placebo Events</th>
<th>N</th>
<th>x</th>
<th>v_i</th>
<th>x_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>belgium</td>
<td>3</td>
<td>13</td>
<td>66</td>
<td></td>
<td>0.3804</td>
<td>-1.4962</td>
</tr>
<tr>
<td>czech</td>
<td>9</td>
<td>17</td>
<td>124</td>
<td></td>
<td>0.1537</td>
<td>-0.6279</td>
</tr>
<tr>
<td>Denm/Fin</td>
<td>11</td>
<td>13</td>
<td>164</td>
<td></td>
<td>0.1555</td>
<td>-0.1486</td>
</tr>
<tr>
<td>germany</td>
<td>19</td>
<td>31</td>
<td>247</td>
<td></td>
<td>0.0769</td>
<td>-0.5096</td>
</tr>
<tr>
<td>hungary</td>
<td>16</td>
<td>29</td>
<td>212</td>
<td></td>
<td>0.0875</td>
<td>-0.5900</td>
</tr>
<tr>
<td>iceland</td>
<td>2</td>
<td>2</td>
<td>22</td>
<td></td>
<td>0.9019</td>
<td>0.1466</td>
</tr>
<tr>
<td>norway</td>
<td>6</td>
<td>11</td>
<td>105</td>
<td></td>
<td>0.2377</td>
<td>-0.5269</td>
</tr>
<tr>
<td>poland</td>
<td>8</td>
<td>8</td>
<td>102</td>
<td></td>
<td>0.2304</td>
<td>0.0000</td>
</tr>
<tr>
<td>sweden</td>
<td>2</td>
<td>9</td>
<td>46</td>
<td></td>
<td>0.5637</td>
<td>-1.3390</td>
</tr>
<tr>
<td>holl/swiz</td>
<td>14</td>
<td>26</td>
<td>291</td>
<td></td>
<td>0.1031</td>
<td>-0.6462</td>
</tr>
<tr>
<td>uk</td>
<td>4</td>
<td>9</td>
<td>83</td>
<td></td>
<td>0.3376</td>
<td>-0.8580</td>
</tr>
<tr>
<td>usa</td>
<td>51</td>
<td>49</td>
<td>539</td>
<td></td>
<td>0.0363</td>
<td>0.0531</td>
</tr>
</tbody>
</table>

### Primary (mortality + hospitalisation)

<table>
<thead>
<tr>
<th>Country</th>
<th>Drug Events</th>
<th>Placebo Events</th>
<th>N</th>
<th>x</th>
<th>v_i</th>
<th>x_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>belgium</td>
<td>31</td>
<td>31</td>
<td>66</td>
<td></td>
<td>0.0347</td>
<td>-0.0299</td>
</tr>
<tr>
<td>czech</td>
<td>35</td>
<td>50</td>
<td>124</td>
<td></td>
<td>0.0324</td>
<td>-0.3486</td>
</tr>
<tr>
<td>Denm/Fin</td>
<td>64</td>
<td>64</td>
<td>164</td>
<td></td>
<td>0.0189</td>
<td>0.0185</td>
</tr>
<tr>
<td>germany</td>
<td>88</td>
<td>100</td>
<td>247</td>
<td></td>
<td>0.0134</td>
<td>-0.1479</td>
</tr>
<tr>
<td>hungary</td>
<td>57</td>
<td>72</td>
<td>212</td>
<td></td>
<td>0.0220</td>
<td>-0.2289</td>
</tr>
<tr>
<td>iceland</td>
<td>6</td>
<td>10</td>
<td>22</td>
<td></td>
<td>0.1686</td>
<td>-0.3642</td>
</tr>
<tr>
<td>norway</td>
<td>41</td>
<td>48</td>
<td>105</td>
<td></td>
<td>0.0254</td>
<td>-0.0784</td>
</tr>
<tr>
<td>poland</td>
<td>26</td>
<td>25</td>
<td>102</td>
<td></td>
<td>0.0589</td>
<td>0.0392</td>
</tr>
<tr>
<td>sweden</td>
<td>15</td>
<td>27</td>
<td>46</td>
<td></td>
<td>0.0563</td>
<td>-0.4227</td>
</tr>
<tr>
<td>holl/swiz</td>
<td>68</td>
<td>95</td>
<td>291</td>
<td></td>
<td>0.0185</td>
<td>-0.3615</td>
</tr>
<tr>
<td>uk</td>
<td>26</td>
<td>29</td>
<td>83</td>
<td></td>
<td>0.0494</td>
<td>-0.1563</td>
</tr>
<tr>
<td>usa</td>
<td>184</td>
<td>216</td>
<td>539</td>
<td></td>
<td>0.0063</td>
<td>-0.1473</td>
</tr>
</tbody>
</table>

Given the overall result and based on distribution of events across countries, expect 2 countries to show effects > 0
MERIT-HF: Primary (mortality + hospitalisation):
HR=0.85 95% CI (0.77, 0.93), $\tau^2=0$, $\sigma^2=0.024$, $I^2=0$
MERIT-HF:
Mortality: HR=0.64 95% CI (0.51, 0.83) by bootstrap
MERIT-HF:
Mortality: $\tau^2=0.031$ 95% CI (0. 0.13)
MERIT-HF: Primary (mortality + hospitalisation): HR=0.85 95% CI (0.79 to 0.90) by bootstrap
MERIT-HF: Primary (mortality + hospitalisation):
$\tau^2=0$, 95% CI (0, 0.0025)
Wedel and DeMets, (Am Heart J 2001;142:502-11.)

• Just as we must be extremely cautious in over-interpreting positive effects in subgroups, even those that are predefined, we must also be cautious in focusing on subgroups with an apparent neutral or negative trend.

• We should examine subgroups to obtain a general sense of consistency, which is clearly the case in MERIT-HF.

• We should expect some variation of the treatment effect around the overall estimate as we examine a large number of subgroups because of small sample size in subgroups and chance.

• Thus the best estimate of the treatment effect on total mortality for any subgroup is the estimate of the hazard ratio for the overall trial.
PLATO

• Randomized double-blind study comparing BRILINTA (N=9333) to clopidogrel (N=9291), both given in combination with aspirin, in patients with acute coronary syndromes.

• Primary endpoint was time to first occurrence of CV death, MI or stroke.

• Randomisation across 41 countries.

• Primary endpoint met for BRILINTA 9.8% vs 11.7% events HR = 0.84
  95% CI 0.77–0.92]; p=0.0003.

• Benefit also seen in overall mortality 4.5% vs 5.9% events HR = 0.78
  95% CI 0.69–0.89]; p=0.0003.
PLATO: Ticagrelor Effect Apparently Inconsistent Across Geographic Regions

- 31 pre-specified subgroup tests conducted for consistency
- No $\alpha$-level adjustment for multiplicity
- Indication of qualitatively different outcomes by region
- Results in NA appear to be driven by US: HR 1.27 (0.92, 1.75)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total Patients</th>
<th>KM at Month 12</th>
<th>HR (95% CI)</th>
<th>Interaction p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tic</td>
<td>Clop</td>
<td></td>
</tr>
<tr>
<td>Asia / Australia</td>
<td>1714</td>
<td>11.4</td>
<td>14.8</td>
<td>0.80 (0.61, 1.04)</td>
</tr>
<tr>
<td>Cent / Sth America</td>
<td>1237</td>
<td>15.2</td>
<td>17.9</td>
<td>0.86 (0.65, 1.13)</td>
</tr>
<tr>
<td>Euro / Md E / Afr</td>
<td>13859</td>
<td>8.8</td>
<td>11.0</td>
<td>0.80 (0.72, 0.90)</td>
</tr>
<tr>
<td>North America</td>
<td>1814</td>
<td>11.9</td>
<td>9.6</td>
<td>1.25 (0.93, 1.67)</td>
</tr>
</tbody>
</table>
Galbraith plot

\[ \log(HR)/SE \]

Country

HR

HR-2*SE

HR+2*SE

1/SE

log(HR)/SE

0 1 2 3 4 5 6 7 8

0 1 2 3 4 5 6 7 8

-4 -3 -2 -1 0 1 2

0 1 2 3 4 5 6 7 8
## Probability of Observing At Least 1 Statistically Significant Treatment Interaction By Chance Alone is High

<table>
<thead>
<tr>
<th>Correlation between tests</th>
<th>Fraction of simulations with at least 1 significant result in 31 tests</th>
<th>Equivalent adjusted p-value to retain overall false-positive error rate at 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>79.1%</td>
<td>0.002</td>
</tr>
<tr>
<td>0.5</td>
<td>51.3%</td>
<td>0.004</td>
</tr>
<tr>
<td>0.9</td>
<td>17.3%</td>
<td>0.014</td>
</tr>
<tr>
<td>0.99</td>
<td>7.9%</td>
<td>0.031</td>
</tr>
<tr>
<td>1.0</td>
<td>5.0%</td>
<td>0.050</td>
</tr>
</tbody>
</table>
PLATO Pattern of effect reversals consistent with what would be expected in a large MRCT

<table>
<thead>
<tr>
<th>Expected no. countries with HR &gt;1</th>
<th>Actual no. countries with HR &gt;1</th>
<th>Expected no. countries with HR &gt;1.25</th>
<th>Actual no. countries with HR &gt;1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.9</td>
<td>12</td>
<td>6.2</td>
<td>3</td>
</tr>
</tbody>
</table>
Analysis by date of birth

- Late Apr
- Early Oct
- Late Nov
- Late Aug
- Late Dec

Log(HR)/SE vs 1/SE

- Date of Birth
- HR overall
- -2 SEs
- +2 SEs
PLATO: What Kind of Factors or Patient Characteristics Might Explain the US vs Non-US Result?

- To explain a meaningful fraction of the US/non-US interaction, a factor is needed that simultaneously:
  - (i) has a strong qualitative interaction with randomized treatment for the primary endpoint and
  - (ii) is strongly imbalanced between US and non US settings

- Weakly imbalanced prognostic factors will likely not be sufficient to explain the US result

- Visual inspection for imbalances of clinical concern needs to be supported by an objective and statistically rigorous analysis of the data
PLATO: No Factor Potentially Accounts for the Regional Interaction with the Exception of ASA Maintenance Dose During Therapy

80-100% of regional interaction explained by ASA maintenance dose
PLATO: The regional interaction is explained by an interaction with ASA maintenance dose

<table>
<thead>
<tr>
<th>ASA Dose (mg)</th>
<th>Ticagrelor</th>
<th>Clopidogrel</th>
<th>HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥300</td>
<td>464 68</td>
<td>492 50</td>
<td>1.45 (1.01, 2.09)</td>
</tr>
<tr>
<td>&gt;100 – &lt;300</td>
<td>525 64</td>
<td>527 65</td>
<td>0.99 (0.70, 1.40)</td>
</tr>
<tr>
<td>≤100</td>
<td>7733 565</td>
<td>7706 723</td>
<td>0.77 (0.69, 0.86)</td>
</tr>
</tbody>
</table>

χ²=16.1  
p=0.00006
PLATO: Similar Pattern of Treatment Effects in Relation to ASA Maintenance Dose in US and Non-US

<table>
<thead>
<tr>
<th>Region</th>
<th>ASA Dose (mg)</th>
<th>Ticagrelor</th>
<th>Clopidogrel</th>
<th>HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥300</td>
<td>324 40</td>
<td>352 27</td>
<td>1.62 (0.99, 2.64)</td>
<td></td>
</tr>
<tr>
<td>&gt;100 – &lt;300</td>
<td>22 2</td>
<td>16 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤100</td>
<td>284 19</td>
<td>263 24</td>
<td>0.73 (0.40, 1.33)</td>
<td></td>
</tr>
<tr>
<td>Non-US</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥300</td>
<td>140 28</td>
<td>140 23</td>
<td>1.23 (0.71, 2.14)</td>
<td></td>
</tr>
<tr>
<td>&gt;100 – &lt;300</td>
<td>503 62</td>
<td>511 63</td>
<td>1.00 (0.71, 1.42)</td>
<td></td>
</tr>
<tr>
<td>≤100</td>
<td>7449 546</td>
<td>7443 699</td>
<td>0.78 (0.69, 0.87)</td>
<td></td>
</tr>
</tbody>
</table>
Summary (1)

• MRCTs are essential in modern day drug development

• Some degree of observed variability is inevitable between regions – in a trial with 80% power and 4 equally sized regions, the chance of at least one reversal is ~30% and with 6 regions the chance is >50%. The corresponding figures are ~20% and 7 regions for a trial with 90% power.

• By definition, MRCTs are not designed, powered or intended to look statistically for true differences between regional effects – all the caveats, biases and pitfalls of subgroup analyses apply.

• Assuming at the planning stage that there is true heterogeneity in the treatment effect by region so ($\tau^2 >0$) is problematic
  
  • What value to choose for $\tau^2$ and on what basis? At what point does the MRCT become meaningless and impossible to interpret clinically due to excessive variability?
  
  • Sample size inflation will quickly render the trial infeasible.
Summary (2)

• A random effects analysis is not the answer especially if applying a Follmann adjustment

• More thought needs to be given to the consistency in defining ‘region’ across trials – “Brazil, Chile, Argentina and New Zealand”

• The optimum allocation of N across regions / countries and the criteria for ‘consistency’ are two sides of the same coin.

• For multiple regions, simplest, but least realistic, criteria is to require all regional effects > 0.

• For reference vs non-reference regions, most pragmatic criteria is to require 3 regions, is to require the point estimate for the reference region to be no less than the lower 100(1-2α)% CI estimate for the treatment effect in the total trial population.
  • provides some reassurance the treatment effect in the reference region is not worryingly less than the treatment effect seen in the total trial population.
Summary (3)

• Graphical methods are very informative when examining regional effects – Galbraith, Q-Q and forts plots are essential.

• Decomposition of the overall $\chi^2$ into a weighted sum of $t^2$ statistics for $i^{th}$ region vs all others is very informative and should be done routinely.

• If there is truly believed to be a region with a different result, extensive evaluation is required to assess play of chance, and to look for possible confounders in medical practice, medical care, quality of trial conduct, cultural, biologic and genetic factors that may be explanatory.

• In the end “Thus the best estimate of the treatment effect ... for any subgroup is the estimate of the ... [treatment effect] for the overall trial” Wedel and DeMets