Can we trust claims from clinical trial reports?

A statistician’s perspective

Stuart Pocock

London School of Hygiene and Tropical Medicine
Reporting of a Major Randomized Trial

company press release?
↓
conference presentation
↓
journal publication
↓
regulatory submission to FDA, EMA
↓
company advertising

Assurance re Quality of Reports

journals: peer review, CONSORT guidelines
regulators: totality of evidence, ICH guidelines
Internal validity

Randomisation OK? often inadequately reported

Masking (blinding) implemented OK?

Size: were enough patients included?

Patient follow-up: problems of drop-outs, non-compliance

Results: correct analysis? emphasis on pre-defined aims? clear and informative presentation?

Conclusions: compatible with results? balanced account of efficacy and side-effects limitations assessed
External validity

Relevant patients
beware of inappropriate extrapolations

Treatment regimens: implementable and cost effective
appropriate control group

Outcome measures: relevant to patient well-being
treatment and follow-up long enough

Other evidence: adequate account of other trials
biological rationale
constructive critical appraisal

no trial is perfect!
Statistics in trial reports

beware of **positive spin**: post hoc emphasis on good news

Andrew Lang  \(1844-1912\)

“He uses statistics as a drunken man uses a lamp-post: for support rather than illumination”

obsession with \(P<.05\) \(\Rightarrow\) a “positive” trial

lots of data to play with \(\Rightarrow\) data dredging
disappointing result: try data dredging?!

**BEAUTIFUL trial** [Lancet 2008; 372 p 807]
ivabradine in 10,917 patients with stable coronary disease

primary composite outcome: CV death, MI, heart failure
median 19 months follow-up

<table>
<thead>
<tr>
<th></th>
<th>ivabradine</th>
<th>placebo</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary</td>
<td>15.4%</td>
<td>15.3%</td>
<td>.94</td>
</tr>
<tr>
<td>subgroup with heart rate ≥ 70:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>primary</td>
<td>17.2%</td>
<td>18.5%</td>
<td>.17</td>
</tr>
<tr>
<td>myocardial infarction</td>
<td>3.1%</td>
<td>4.9%</td>
<td>.001</td>
</tr>
</tbody>
</table>

hypothesis generating, but highlighted in abstract
a short history of the P-value

R A Fisher (1925)
Statistical Methods for Research Workers

introduced hypothesis testing and P-values

tables for specially selected values of P
eg. P=.1 .05 .02 .01 .001
but no particular emphasis on .05

"when P is between .02 and .05 the result must be judged significant but barely so. The data do not demonstrate this point beyond reasonable doubt"

P<.05 is **not** strong evidence, P<.001 **is**
R A Fisher, the founder of statistical inference, working on a mechanical calculator
Jerzy Neyman and Egon Pearson (1933)
Philosophical Transactions of the Royal Society

hypothesis testing formulation
null hypothesis versus alternative hypothesis
P<.05 reject null hypothesis
P>.05 accept null hypothesis

Fisher’s strength of evidence interpretation
versus
Neyman & Pearson’s decision approach

introduced Type II error, power calculations

Fisher: “the calculation is absurdly academic, for no scientific worker has a fixed level of significance at which he rejects hypotheses”
P-values are about “Shades of Grey”

Weak evidence against the null hypothesis

Increasing evidence against the null hypothesis with decreasing P value

Strong evidence against the null hypothesis

“A P-value is no substitute for a brain”

beware of big effects in small trials

Perioperative beta-blocker use in non-cardiac surgery

**DECREASE trial of bisoprolol**  [NEJM 1999;341 p 1789-]

<table>
<thead>
<tr>
<th></th>
<th>bisoprolol</th>
<th>control</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>59</td>
<td>53</td>
</tr>
<tr>
<td>death</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>myocardial infarction</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

“too good to be true”?

**POISE trial of metoprolol**  [Lancet 2008;371 p 1839-]

<table>
<thead>
<tr>
<th></th>
<th>metoprolol</th>
<th>placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>4174</td>
<td>4177</td>
</tr>
<tr>
<td>death</td>
<td>129</td>
<td>97</td>
</tr>
<tr>
<td>myocardial infarction</td>
<td>176</td>
<td>239</td>
</tr>
</tbody>
</table>

**ESC/ESA Guidelines 2014**: evidence inconclusive
beware of small effects in big trials

**IMPROVE-IT trial**  [NEJM 2015;372 p2387-]

18,144 acute coronary syndrome patients on simvastatin ezetimibe vs placebo

composite primary endpoint: CV death, MI, stroke, unstable angina, coronary revasc.

5314 primary events over mean 5.4 years follow-up

the definitive study of ezetimibe?
Primary Endpoint — ITT

Cardiovascular death, MI, documented unstable angina requiring rehospitalization, coronary revascularization (≥30 days), or stroke

Simva — 34.7%
2742 events
HR 0.936 CI (0.887, 0.988)
p=0.016
NNT= 50

EZ/Simva — 32.7%
2572 events

7-year event rates
on top of simvastatin, ezetimibe had a modest mean reduction in LDL-C (16.7 mg/dl)

modest impact on cardiovascular primary events:
relative risk reduction 6.4% (95% CI 2.2% to 11.3%)
absolute risk reduction 2.0% over 7 years

is this a worthwhile benefit?
ODYSSEY OUTCOMES Trial [ACC March 2018]

alirocumab vs placebo in 18,924 ACS patients with LDL-C ≥70 mg/dl and on high-dose statin

this PCSK9 inhibitor reduces LDL-C by ~60%

composite primary outcome: CHD death, MI, ischemic stroke, hospitalized unstable angina

median 2.8 years follow-up
ODYSSEY trial: primary efficacy endpoint MACE

![Graph showing MACE rates over years for Placebo and Alirocumab groups.](image)

- **Placebo**
  - Initial number at risk: 9462
  - Years: 0, 1, 2, 3, 4
  - MACE rates: 0%, 3%, 6%, 9%, 12%

- **Alirocumab**
  - Initial number at risk: 9462
  - Years: 0, 1, 2, 3, 4
  - MACE rates: 0%, 3%, 6%, 9%, 12%

**Statistical Analysis**

- Hazard Ratio (HR) 0.85 (95% CI 0.78, 0.93)
- **P-value**: 0.0003

**Additional Information**

- **ARR**: 1.6%

**Number at Risk**

- **Placebo**: 9462, 8805, 8201, 3471, 629
- **Alirocumab**: 9462, 8846, 8345, 3574, 653
### ODYSSEY Primary Endpoint and its Components

<table>
<thead>
<tr>
<th>Endpoint, n (%)</th>
<th>Alirocumab (N=9462)</th>
<th>Placebo (N=9462)</th>
<th>Hazard Ratio (95% CI)</th>
<th>Log-rank P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACE (primary)</td>
<td>903 (9.5)</td>
<td>1052 (11.1)</td>
<td>0.85 (0.78, 0.93)</td>
<td>0.0003</td>
</tr>
<tr>
<td>CHD death</td>
<td>205 (2.2)</td>
<td>222 (2.3)</td>
<td>0.92 (0.76, 1.11)</td>
<td>0.38</td>
</tr>
<tr>
<td>Non-fatal MI</td>
<td>626 (6.6)</td>
<td>722 (7.6)</td>
<td>0.86 (0.77, 0.96)</td>
<td>0.006</td>
</tr>
<tr>
<td>Ischemic stroke</td>
<td>111 (1.2)</td>
<td>152 (1.6)</td>
<td>0.73 (0.57, 0.93)</td>
<td>0.01</td>
</tr>
<tr>
<td>Unstable angina</td>
<td>37 (0.4)</td>
<td>60 (0.6)</td>
<td>0.61 (0.41, 0.92)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

**absolute reduction in first MACE event:**
5.62 per 1000 patient years (95% CI 2.35 to 8.89)

**no. needed to treat (NNT):**
63 patients for median 2.8 years (95% CI 41 to 141)

**no apparent benefit in first 12 months**
**ODYSSEY Secondary Endpoints** in order of hierarchical testing

<table>
<thead>
<tr>
<th>Endpoint, n (%)</th>
<th>Alirocumab (N=9462)</th>
<th>Placebo (N=9462)</th>
<th>Hazard Ratio (95% CI)</th>
<th>Log-rank P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD event</td>
<td>1199 (12.7)</td>
<td>1349 (14.3)</td>
<td>0.88 (0.81, 0.95)</td>
<td>0.001 ✓</td>
</tr>
<tr>
<td>Major CHD event</td>
<td>793 (8.4)</td>
<td>899 (9.5)</td>
<td>0.88 (0.80, 0.96)</td>
<td>0.006 ✓</td>
</tr>
<tr>
<td>CV event</td>
<td>1301 (13.7)</td>
<td>1474 (15.6)</td>
<td>0.87 (0.81, 0.94)</td>
<td>0.0003 ✓</td>
</tr>
<tr>
<td>Death, MI, ischemic stroke</td>
<td>973 (10.3)</td>
<td>1126 (11.9)</td>
<td>0.86 (0.79, 0.93)</td>
<td>0.0003 ✓</td>
</tr>
<tr>
<td>CHD death</td>
<td>205 (2.2)</td>
<td>222 (2.3)</td>
<td>0.92 (0.76, 1.11)</td>
<td>0.38 X</td>
</tr>
<tr>
<td>CV death</td>
<td>240 (2.5)</td>
<td>271 (2.9)</td>
<td>0.88 (0.74, 1.05)</td>
<td>0.15 (X)</td>
</tr>
<tr>
<td>All-cause death</td>
<td>334 (3.5)</td>
<td>392 (4.1)</td>
<td>0.85 (0.73, 0.98)</td>
<td>0.026* (√)</td>
</tr>
</tbody>
</table>

all-cause death not formally significant, it’s exploratory

no mortality signal in FOURIER trial of evolocumab
**Subgroup Claim**

alirocumab more effective if LDL-C $\geq$ 100 mg/dl:
24% reduction in MACE, 29% reduction in death

but no significant interactions ($P=0.09$ and $P=0.12$ respectively)

beware of such “positive spin”

more plausible is that absolute benefit for MACE is greater in higher risk patients

need to stratify patients by overall risk (not just LDL-C) and concentrate on absolute (not relative) reduction
General Issues arising from ODYSSEY

Estimating the magnitude of treatment effect

**relative risk reduction** useful, but quantify uncertainty:

eg risk ratio, odds ratio, hazard ratio
  plus 95% confidence interval

**absolute risk reduction** more useful

eg. difference in %, difference in rates, NNT
  again plus 95% CI

absolute risk reduction greater in higher-risk patients
The meaning of LIFE [Lancet March 23, 2002]

losartan vs atenolol, 9193 patients with hypertension
primary endpoint: death, MI or stroke over 4.8 years

no of events  508  losartan vs  588  atenolol
4605                    4588

relative risk reduction 13%  95% CI 2% to 23% P=.021

“losartan prevents more cardiovascular events than atenolol”

but not overwhelming evidence
also what about absolute risk reduction?
Absolute Risk Reduction in the LIFE trial

primary endpoint: death, MI or stroke

<table>
<thead>
<tr>
<th>losartan</th>
<th>atenolol</th>
</tr>
</thead>
<tbody>
<tr>
<td>it occurred in</td>
<td>508</td>
</tr>
<tr>
<td>rate per 1000 patient years</td>
<td>23.8</td>
</tr>
</tbody>
</table>

difference 4.1 per 1000 patient years
with 95% CI 1.1 to 7.1 per 1000 patient years

No. Needed to Treat (NNT)

244 patient years of treatment to prevent one event
95% CI 141 to 909 patient years
a small gain imprecisely estimated
Subgroup analyses: interpret with caution

1) patients are not homogeneous
   response to treatment may well vary
   legitimate to explore in subgroup analyses

2) trials usually not large enough
   lack power to detect subgroup effects

3) many possible subgroups
   guard against data dredging/false positive

4) do not rely on subgroup P-values
   use interaction tests instead
A second **LIFE study** report on **diabetic subgroup**
[Lancet 23 March 2002]
“the benefits of losartan were more marked in this group”

<table>
<thead>
<tr>
<th></th>
<th>losartan</th>
<th>atenolol</th>
<th>relative risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>diabetics N=1195</td>
<td>103</td>
<td>139</td>
<td>0.77</td>
</tr>
<tr>
<td>non-diabetics N=7798</td>
<td>405</td>
<td>449</td>
<td>0.89</td>
</tr>
</tbody>
</table>

is relative risk greater in diabetics?
interaction test (test for heterogeneity) P=0.22

**insufficient evidence that more effective in diabetic subgroup**
Can we be more sensibly cautious re subgroup claims

CURRENT OASIS 7

Standard vs double dose clopidogrel in 25,087 ACS patients

NEJM 2 Sept 2010

“in patients referred for an invasive strategy, there was no significant difference between double-dose and standard dose”

Lancet on-line 1 Sept 2010

“In patients undergoing PCI double dose was associated with a reduction in CV events”

Confused?!
CURRENT OASIS 7

Primary Composite Outcome: CV Death, MI, stroke at 30 days
Overall 4.4% vs 4.2% P=0.37
no evidence that double dose is beneficial?

Subgroup analysis

clopidogrel dose

<table>
<thead>
<tr>
<th></th>
<th>standard</th>
<th>double</th>
<th>hazard ratio</th>
<th>interaction test</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCI (N=17232)</td>
<td>4.5%</td>
<td>3.9%</td>
<td>0.85</td>
<td>P=.039 P=.016</td>
</tr>
<tr>
<td>no PCI (N=7855)</td>
<td>4.2%</td>
<td>4.9%</td>
<td>1.17</td>
<td>P=.14</td>
</tr>
</tbody>
</table>

“double dose clopidogrel reduced CV events in PCI patients”?! beware: 1) PCI an improper subgroup
2) qualitative interactions are rare/implausible
3) such secondary evidence is weak
Analysis by Intention to Treat (ITT)
Analyse all randomised patients in their allocated groups
An unbiased comparison of strategies
Dilution bias for “pure” treatment effect

Per protocol analysis of compliers
Potential bias
Nearer to “pure” treatment effects

Often do both, but emphasize ITT
ROCKET-AF trial [NEJM 2011 365 p883-]

rivaroxaban vs warfarin in atrial fibrillation

14264 patients with 1.9 years median follow-up

primary endpoint: stroke or systemic embolism

Analysis by Intention to Treat

269 vs 306 hazard ratio 0.88 95% CI 0.74 to 1.03, P=0.12

Per Protocol Analysis

188 vs 241 hazard ratio 0.79 95% CI 0.66 to 0.96, P=0.02

Conclusions concentrated on the former

The Journal did not permit a claim of superiority
When can per protocol analysis be of value?

CABANA trial

2204 patients with atrial fibrillation

Ablation therapy  vs  Drug therapy

↓

9.2% not ablated  27.5% ablated

Primary endpoint: death, stroke, serious bleed, cardiac arrest over mean 4 years follow-up

ITT analysis “diluted” by crossovers
Primary Endpoint (Death, Disabling Stroke, Serious Bleeding, or Cardiac Arrest) (ITT)

Ablation vs. Drug
Hazard ratio: 0.86 (95% CI, 0.65–1.15)
P=0.303

Event rate (%)

Months since randomization

Number at risk

<table>
<thead>
<tr>
<th></th>
<th>Drug</th>
<th>Ablation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1096</td>
<td>1108</td>
</tr>
<tr>
<td>6</td>
<td>1036</td>
<td>1045</td>
</tr>
<tr>
<td>12</td>
<td>1066</td>
<td>1021</td>
</tr>
<tr>
<td>18</td>
<td>970</td>
<td>996</td>
</tr>
<tr>
<td>24</td>
<td>880</td>
<td>915</td>
</tr>
<tr>
<td>30</td>
<td>763</td>
<td>793</td>
</tr>
<tr>
<td>36</td>
<td>652</td>
<td>700</td>
</tr>
<tr>
<td>42</td>
<td>578</td>
<td>614</td>
</tr>
<tr>
<td>48</td>
<td>499</td>
<td>535</td>
</tr>
<tr>
<td>54</td>
<td>418</td>
<td>432</td>
</tr>
<tr>
<td>60</td>
<td>312</td>
<td>309</td>
</tr>
</tbody>
</table>

MAYO CLINIC  Duke Clinical Research Institute  NIH
Primary Endpoint (Death, Disabling Stroke, Serious Bleeding, or Cardiac Arrest (Per Protocol))

Ablation vs. Drug
Hazard ratio: 0.73 (95% CI, 0.54–0.99)
P=0.046

Event rate (%)

Months since randomization

Number at risk

<table>
<thead>
<tr>
<th>Drug</th>
<th>1096</th>
<th>987</th>
<th>968</th>
<th>937</th>
<th>918</th>
<th>918</th>
<th>849</th>
<th>735</th>
<th>648</th>
<th>566</th>
<th>396</th>
<th>330</th>
<th>275</th>
<th>204</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ablation</td>
<td>987</td>
<td>968</td>
<td>937</td>
<td>918</td>
<td>849</td>
<td>735</td>
<td>648</td>
<td>566</td>
<td>494</td>
<td>404</td>
<td>291</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MAYO CLINIC  Duke Clinical Research Institute  NIH  National Heart, Lung, and Blood Institute
My Conclusions re CABANA

Analysis by Intention to Treat

no significant reduction on primary endpoint

results affected by cross-overs and low event rates

ablation reduced mortality or CV hospitalization by 17% (P=0.002)

Per Protocol Analysis

33% reduction in primary endpoint (P=0.046)

40% reduction in mortality (P=0.005)

but not clear what method was used

potential bias, controversy remains
Strategies for handling Secondary Endpoints

pre-define a limited set of key secondary endpoints

other endpoints become exploratory

pre-define either:

1) a hierarchy of secondary testing

or 2) correction for multiple testing (eg Bonferroni)

strict control of type I error: good or bad?

it matters to FDA, less to scientific knowledge

flexibility without cheating

all-cause death is different?
Interpretation of Secondary Endpoint Surprises

**safety concern**: heart failure in SAVOR trial

**efficacy bonus**: mortality in EMPA-REG trial

my prior statistical perspective

any unexpected finding (good or bad) is prone to be an exaggeration

note one of multiple hypotheses across secondary endpoints

collect more data (if you can) and expect regression to the truth

is it a real effect or just due to chance? often impossible to tell
SAVOR-TIMI 53 trial [NEJM 2013; 369 p 1317-]

Saxagliptin vs Placebo in 16,492 high risk type II diabetics

788 sites in 26 countries, median 2.1 years follow-up

<table>
<thead>
<tr>
<th>Treatment</th>
<th>N</th>
<th>Hazard Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saxagliptin</td>
<td>8280</td>
<td>1.00 (0.89 to 1.12)</td>
</tr>
<tr>
<td>Placebo</td>
<td>8212</td>
<td></td>
</tr>
</tbody>
</table>

- **Primary endpoint (CV death, MI, stroke)**
  - Saxagliptin: 613
  - Placebo: 609
  - Hazard ratio: 1.00 (0.89 to 1.12)

- **Heart failure hospitalization**
  - Saxagliptin: 289
  - Placebo: 228
  - Hazard ratio: 1.27 (1.07 to 1.51) \( \uparrow \)
  - P = .007

**Primary endpoint:** non-inferiority established, but no benefit

**Heart failure:** given multiple testing, a false positive?
Combining Evidence from 3 Related Trials

<table>
<thead>
<tr>
<th>Trial</th>
<th>Hazard Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAVOR (saxagliptin)</td>
<td>1.27 (1.07, 1.50)</td>
</tr>
<tr>
<td>EXAMINE (alogliptin)</td>
<td>1.07 (0.79, 1.45)</td>
</tr>
<tr>
<td>TECOS (sitagliptin)</td>
<td>1.00 (0.83, 1.20)</td>
</tr>
<tr>
<td>Overall (fixed effect)</td>
<td>1.13 (1.01, 1.27)</td>
</tr>
<tr>
<td>Overall (random effect)</td>
<td>1.12 (0.95, 1.32)</td>
</tr>
<tr>
<td>heterogeneity P=0.16</td>
<td></td>
</tr>
</tbody>
</table>

Hazard ratio for heart failure hospitalisation
EMPÁ-REG OUTCOME trial  [NEJM 17 Sept 2015]  
empagliflozin 10mg vs 25mg vs placebo in 7020 type 2 diabetics  
primary endpoint: CV death, MI, stroke over median 3.1 years  

<table>
<thead>
<tr>
<th></th>
<th>empagliflozin combined [N=4687]</th>
<th>Placebo [N=2333]</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary</td>
<td>10.5%</td>
<td>12.1%</td>
<td>.04</td>
</tr>
<tr>
<td>heart failure hospn.</td>
<td>2.7%</td>
<td>4.1%</td>
<td>.002</td>
</tr>
<tr>
<td>all-cause death</td>
<td>5.7%</td>
<td>8.3%</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

how do we interpret such impressive secondary findings?
Moderating “Too Good To Be True” Results
Bayesian Analysis of All Cause Mortality
EMPA REG OUTCOME Trial (Empagliflozin in T2 DM)

<table>
<thead>
<tr>
<th>Group</th>
<th>OUTCOME +</th>
<th>EVENT -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Con</td>
<td>194</td>
<td>2139</td>
<td>2333</td>
</tr>
<tr>
<td>Ept</td>
<td>269</td>
<td>4418</td>
<td>4687</td>
</tr>
<tr>
<td>Total</td>
<td>463</td>
<td>6557</td>
<td>7020</td>
</tr>
</tbody>
</table>

RRR 32% 95% CI 18%-43% P (2-tailed) <0.001

<table>
<thead>
<tr>
<th>Prior</th>
<th>Pb&gt;0</th>
<th>Pb&gt;10%</th>
<th>Pb&gt;15%</th>
<th>Pb&gt;20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noninformative</td>
<td>0.999</td>
<td>0.998</td>
<td>0.992</td>
<td>0.963</td>
</tr>
<tr>
<td>Skeptical</td>
<td>0.999</td>
<td>0.982</td>
<td><strong>0.918</strong></td>
<td>0.743</td>
</tr>
</tbody>
</table>

courtesy of Sanjay Kaul
COMPASS trial [NEJM 2017:377 p1319- ]

27,395 patients with stable cardiovascular disease

Rivaroxaban 2.5 mg bd + Aspirin 100mg
vs
Rivaroxaban 5 mg bd alone
vs
Aspirin 100 mg alone

primary outcome:
composite of death, myocardial infarction, stroke

trial stopped early for superiority of R + A

mean follow-up 23 months
**COMPASS trial Primary Outcome**

<table>
<thead>
<tr>
<th></th>
<th>R + A (N=9152)</th>
<th>R alone (N=9117)</th>
<th>A alone (N=9126)</th>
<th>P-value for R + A v Alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Outcome (CV death, MI, stroke)</td>
<td>379</td>
<td>448</td>
<td>496</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>CV death</td>
<td>160</td>
<td>195</td>
<td>203</td>
<td>P=0.01</td>
</tr>
<tr>
<td>Stroke</td>
<td>83</td>
<td>117</td>
<td>142</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>Myocardial Infarction</td>
<td>178</td>
<td>182</td>
<td>205</td>
<td>P=0.12</td>
</tr>
<tr>
<td>Major Bleeding</td>
<td>288</td>
<td>255</td>
<td>170</td>
<td>P&lt;0.001</td>
</tr>
</tbody>
</table>

strong evidence that R + A has superior efficacy and R + A increases bleeding risk
Feb 6 2017: 1st formal interim analysis

for primary efficacy outcome:

R + A vs A alone has $z = 4.592$, exceeds boundary

R alone vs A alone has $z = 2.44$, $P=.015$

DSMB recommends stopping

Aug 27 2017: results published in NEJM

R + A vs A alone $z = 4.126$

hazard ratio 0.76 (95% CI 0.66 to 0.86)  $P<0.0001$

R alone vs A alone $z = 1.575$

hazard ratio 0.90 (95% CI 0.79 to 1.03)  $P=0.12$

some regression to the truth
Issues re Stopping Early

extreme boundary means superiority of R + A believable
but may be exaggerated somewhat (random high)
mean follow-up restricted to 23 months
less assurance re long-term benefit
balancing efficacy and safety (bleeding): less data
regulatory consequences with FDA and EMA
a negative trial: can it be rescued?
TRUE-AHF trial  [NEJM 2017; 376 p1956- ]

Ularitide vs Placebo in 2157 acute heart failure patients

<table>
<thead>
<tr>
<th>Co-Primary Endpoints</th>
<th>Ularitide</th>
<th>Placebo</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cardiovascular Death</strong></td>
<td>21.7%</td>
<td>21.0%</td>
<td>0.75</td>
</tr>
<tr>
<td>(median 15 months)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Clinical Composite over 48 hrs:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>improved</td>
<td>48.6%</td>
<td>47.5%</td>
<td>0.82</td>
</tr>
<tr>
<td>unchanged</td>
<td>44.8%</td>
<td>44.2%</td>
<td></td>
</tr>
<tr>
<td>worse</td>
<td>6.6%</td>
<td>8.3%</td>
<td></td>
</tr>
</tbody>
</table>

seemingly, a “negative” trial
but 16.6% of patients were ineligible
also 3.7% were missing clinical composite
TRUE-AHF Results in 1799 Eligible Patients

<table>
<thead>
<tr>
<th></th>
<th>Ularitide</th>
<th>Placebo</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular Death</td>
<td>20.7%</td>
<td>20.8%</td>
<td>0.87</td>
</tr>
<tr>
<td>Clinical Composite over 48 hrs:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>improved</td>
<td>49.8%</td>
<td>45.8%</td>
<td>0.035</td>
</tr>
<tr>
<td>unchanged</td>
<td>43.9%</td>
<td>45.6%</td>
<td></td>
</tr>
<tr>
<td>worse</td>
<td>6.3%</td>
<td>8.6%</td>
<td></td>
</tr>
</tbody>
</table>

a weak signal of short-term benefit in eligible patients

such post hoc findings need cautious interpretation

concerns about quality of trial conduct
Were some countries deficient in trial conduct?

**TOPCAT trial**  
[Circulation 2015; 131 p34-]

spironolactone vs placebo in preserved EF heart failure  
primary outcome: CV death, cardiac arrest or heart failure hospn.  
hazard ratio 0.89 (95% CI 0.77 to 1.04) P=0.14

patients in Russia and Georgia (N=1678)  
few events, not representative

in Americas subgroup (N=1767)  
hazard ratio 0.82 (95% CI 0.69 to 0.98) P=0.026

is this convincing enough to recommend spironolactone?
Regional Variation in TOPCAT

![Graph showing primary outcome for Placebo and Spironolactone in Americas and Russia/Georgia.](image)

- **Primary Outcome**
- **Years**
- **Number at risk**
  - Americas (P/S): 880/885
  - Russia/Georgia (P/S): 842/836

- **Lines**
  - Placebo
  - Spironolactone
Meta-analysis: the pros and cons

any one trial is 1) too small and 2) lacks generalisability

combining evidence from related trials is good in principle

but can meta-analyses be trusted?

3 key concerns re any meta-analysis:

breadth how similar are the trials?  
re patients, treatments, outcomes

quality which trials are good enough to include?  
use of individual patient data is better

representativeness can one identify all eligible studies?  
risk of publication bias

meta-analysis is a “growth industry”

Journals are wary: publish the good, ignore the bad!
Oseltamivir (tamiflu) treatment for influenza in adults

meta-analysis of individual patient data

from 9 randomised trials (all sponsored by Roche)

[Lancet 2015; 385 p 1729-]
Time to Symptom Alleviation: Accelerated Failure Time Meta-Analysis

<table>
<thead>
<tr>
<th>Trial</th>
<th>Tamiflu</th>
<th>Placebo</th>
<th>Estimated Difference in Median Time (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M76001</td>
<td>681</td>
<td>355</td>
<td>-27.4</td>
</tr>
<tr>
<td>WV15819+</td>
<td>223</td>
<td>254</td>
<td>-17.0</td>
</tr>
<tr>
<td>WV15670</td>
<td>157</td>
<td>161</td>
<td>-26.6</td>
</tr>
<tr>
<td>WV15812+</td>
<td>118</td>
<td>133</td>
<td>-6.5</td>
</tr>
<tr>
<td>JV15823</td>
<td>121</td>
<td>130</td>
<td>-21.5</td>
</tr>
<tr>
<td>WV15671</td>
<td>121</td>
<td>128</td>
<td>-33.1</td>
</tr>
<tr>
<td>WV16277</td>
<td>119</td>
<td>109</td>
<td>-19.7</td>
</tr>
<tr>
<td>WV15730</td>
<td>19</td>
<td>19</td>
<td>-69.1</td>
</tr>
<tr>
<td>WV15707</td>
<td>6</td>
<td>6</td>
<td>19.4</td>
</tr>
</tbody>
</table>

Overall

Favours tamiflu: 0.80 (95% CI 0.74, 0.85)
Favours placebo: -25.2 hours (95% CI -36.2, -16.0)
Estimated Benefits and Risks of Tamiflu

median time to alleviation of symptoms ↓ 25 hours

risk of lower respiratory infection ↓ 3.8%

risk of hospital admission (rare) ↓ 0.6%

incidence of nausea ↑ 3.7%  vomiting ↑ 4.7%

Controversy: validity of findings questioned  
our integrity challenged
Conclusions

clinical trial reports cannot be automatically trusted
constructive critical appraisal is always required
positive spin (being economical with the truth) is common
trialists (and sponsors) struggle with
1) the search for truth (honest science)
   versus
2) the desire for a “positive” trial
journal articles more reliable than conference presentations
only regulators see the full story
Further Reading

JACC 2018; 71: 2957-69
NEJM 2016; 375: 861-70 and 971-9
JACC 2014; 64; 1615-28