Utilities and Pitfalls of Composite and Surrogate Endpoints in Clinical Trials

Helen Parise, ScD

Cardiovascular Research Foundation
Columbia University Medical Center
Common CV Clinical Endpoints

- Death (Cardiac, Non-Cardiac)
- Myocardial Infarction (Q-Wave, NQWMI)
- Stent Thrombosis (subcategories)
- Target Lesion Revascularization
- Target Vessel Revascularization
• What is a surrogate endpoint?
  - An outcome measure which may occur sooner or with greater frequency than the “true” outcome of interest
  - Predictive of the “true” clinical endpoint
  - In the causal pathway

Treatment → Surrogate → Clinical Endpoint
Surrogate Endpoints in DES Studies

• Surrogates in DES Studies
  - Usually continuous variables that are objectively quantifiable and reproducible
• IVUS: Neointimal Volume
• Angiographic:
  - Late Lumen Loss
  - Percent Diameter Stenosis
Surrogate Endpoints

Why use surrogate endpoints?

- Reduce sample size with adequate power
  - fewer patients + shorter follow-up = reduced costs

- Avoid/minimize randomizing patients to a therapy which might not provide benefit or possibly cause harm

- Test new technologies and be able to anticipate their outcomes
How do I know if a surrogate is good enough to use in my clinical trial?
Access to patient level data from multiple well conducted trials is required

Analysis of 11 trials evaluating DES v. BMS

- TAXUS-IV, TAXUS-V, TAXUS-VI, SIRIUS, E-SIRIUS, C-SIRIUS, RAVEL, DELIVER, REALITY, ENDEAVOR II, ENDEAVOR III

In-Segment and In-Stent Late Loss (LL) and %DS were evaluated
Statistical Criteria for Evaluating a Surrogate

1. Consistent evidence of treatment differences within each trial

Pocock, et al., JACC 2008; 51 p 23-32
## 1. Evidence of treatment difference

### Z-Scores for Treatment Difference

<table>
<thead>
<tr>
<th></th>
<th>TAX IV</th>
<th>TAX V</th>
<th>TAX VI</th>
<th>SIR-IUS</th>
<th>E-SIR-IUS</th>
<th>C-SIR-IUS</th>
<th>RA-VEL</th>
<th>DE-LIVER</th>
<th>RE-ALITY</th>
<th>END II</th>
<th>END III</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In-Stent</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LL</td>
<td>11.6</td>
<td>10.5</td>
<td>10.6</td>
<td>18.7</td>
<td>14.5</td>
<td>8.2</td>
<td>13.7</td>
<td>3</td>
<td>7.4</td>
<td>9.2</td>
<td>8.3</td>
</tr>
<tr>
<td>%DS</td>
<td>12.5</td>
<td>10</td>
<td>10.5</td>
<td>18.5</td>
<td>14</td>
<td>8.1</td>
<td>11.9</td>
<td>2.7</td>
<td>3.3</td>
<td>8.5</td>
<td>6.5</td>
</tr>
<tr>
<td><strong>In-Segment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LL</td>
<td>8.9</td>
<td>7.3</td>
<td>7.2</td>
<td>13.1</td>
<td>11.1</td>
<td>5.8</td>
<td>9.8</td>
<td>2.3</td>
<td>4.5</td>
<td>7.7</td>
<td>4.4</td>
</tr>
<tr>
<td>%DS</td>
<td>9.4</td>
<td>6.2</td>
<td>8.2</td>
<td>13.3</td>
<td>11</td>
<td>6.8</td>
<td>7.2</td>
<td>2.3</td>
<td>2.5</td>
<td>7.3</td>
<td>3.4</td>
</tr>
<tr>
<td>TLR</td>
<td>4.7</td>
<td>4</td>
<td>3.6</td>
<td>7.1</td>
<td>6.1</td>
<td>3.1</td>
<td>3.8</td>
<td>1.3</td>
<td>1.1</td>
<td>4</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Pocock, et al., JACC 2008; 51 p 23-32
Statistical Criteria for Evaluating a Surrogate

1. Consistent evidence of treatment differences within each trial
2. Strong relationship with clinical outcome

Pocock, et al., JACC 2008; 51 p 23-32
2. Relationship to clinical outcome

- c-statistics summarizing the strength of association to TLR in each trial

<table>
<thead>
<tr>
<th>Trial</th>
<th>In-Stent</th>
<th>In-Segment</th>
<th>Trial</th>
<th>In-Stent</th>
<th>In-Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAXUS-IV</td>
<td>0.9</td>
<td>0.94</td>
<td>0.92</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>TAXUS-V</td>
<td>0.88</td>
<td>0.9</td>
<td>0.91</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>TAXUS-VI</td>
<td>0.86</td>
<td>0.87</td>
<td>0.89</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>SIRIUS</td>
<td>0.88</td>
<td>0.9</td>
<td>0.92</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>E-SIRIUS</td>
<td>0.94</td>
<td>0.95</td>
<td>0.92</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>C-SIRIUS</td>
<td>0.86</td>
<td>0.86</td>
<td>0.93</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>RAVEL</td>
<td>0.98</td>
<td>0.93</td>
<td>0.91</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>DELIVER</td>
<td>0.85</td>
<td>0.86</td>
<td>0.86</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>REALITY</td>
<td>0.9</td>
<td>0.93</td>
<td>0.93</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>END. II</td>
<td>0.88</td>
<td>0.91</td>
<td>0.88</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>END. III</td>
<td>0.77</td>
<td>0.8</td>
<td>0.93</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Averaged</td>
<td>0.88</td>
<td>0.9</td>
<td>0.91</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>

Pocock, et al., JACC 2008; 51 p 23-32
Statistical Criteria for Evaluating a Surrogate

1. Consistent evidence of treatment differences within each trial
2. Strong relationship with clinical outcome
3. Treatment difference in clinical outcome statistically explained by the surrogate within each trial [Prentice criterion]

Pocock, et al., JACC 2008; 51 p 23-32
### 3. Treatment Differences explained within trial [Prentice criterion]

#### Percent treatment effect explained* (largest 5 trials shown)

<table>
<thead>
<tr>
<th>Trial</th>
<th>In-Stent</th>
<th>In-Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LL</td>
<td>%DS</td>
</tr>
<tr>
<td>TAXUS-IV</td>
<td>98</td>
<td>138</td>
</tr>
<tr>
<td>TAXUS-V</td>
<td>101</td>
<td>94</td>
</tr>
<tr>
<td>TAXUS-VI</td>
<td>115</td>
<td>124</td>
</tr>
<tr>
<td>SIRIUS</td>
<td>132</td>
<td>142</td>
</tr>
<tr>
<td>ENDEAVOR II</td>
<td>100</td>
<td>119</td>
</tr>
</tbody>
</table>

*from logistic model for log odds of TLR after potential surrogate (e.g., LL) is added as predictor variable.

*Pocock, et al., JACC 2008; 51 p 23-32*
Statistical Criteria for Evaluating a Surrogate

1. Consistent evidence of treatment differences within each trial
2. Strong relationship with clinical outcome
3. Treatment difference in clinical outcome statistically explained by the surrogate *within* each trial [Prentice criterion]
4. Magnitude of treatment difference in clinical outcome clearly linked to magnitude of treatment difference in surrogate *across* trials [Hughes criterion]

*Pocock, et al., JACC 2008; 51 p 23-32*
4. Relating size of treatment effect across trials [Hughes criterion]

Pocock, et al., JACC 2008; 51 p 23-32
Example: XIENCE V

- SPIRIT III trial
- Primary endpoint: in-segment late loss at 8 months

<table>
<thead>
<tr>
<th></th>
<th>Xience V</th>
<th>Taxus</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>669</td>
<td>332</td>
</tr>
<tr>
<td>Mean Late Loss</td>
<td>0.14mm</td>
<td>0.26mm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P=0.003</td>
</tr>
<tr>
<td>Target Lesion Revasc.</td>
<td>3.4%</td>
<td>5.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P=0.12</td>
</tr>
<tr>
<td>1 year MACE</td>
<td>6.0%</td>
<td>10.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P=0.02</td>
</tr>
</tbody>
</table>

- Larger long term trials (and registries) underway to assess patient safety re death, MI, stent thrombosis

Stone, et al. JAMA 2008; 299 p 1903-1913
Surrogates

- Reduced trial size → more efficient progress
- LL and % DS both look good as surrogates for TLR in DES
- Use can speed up the regulatory process …
  IF the surrogate is acceptable to the regulatory agency
Surrogate Endpoints – Some Pitfalls

- May not predict clinical events in “real world”
- Surrogates for efficacy may not (and likely do not) predict safety
- Can lead to premature approval of potentially unsafe treatment
- Missing data issues: e.g. incomplete follow-up in angiographic endpoints may bias results
Composite Endpoints

• What is a Composite Endpoint?
  - Common in Cardiovascular Trials
  - Usually a group of clinical outcomes considered together to form a single endpoint
  - Generally weighted equally
  - Experiencing any of the individual outcomes → experiencing the composite
Composite Endpoints: Examples

- Examples of Composite Endpoints used in Cardiovascular Trials
  - Cardiac death/MI
  - Device-oriented (TLF): cardiac death/MI/TLR
  - TVF: cardiac death/MI/TVR
  - MACE: a mixed bag (and varies from study to study!)
Composite Endpoints

- Why use Composite Endpoints?
  - Gain in statistical power
  - Simple summary of several outcomes
  - Captures broader range of treatment experience
  - Multiple outcomes tested simultaneously without alpha “hit” for multiplicity
  - Avoids (some) issues of competing risks
Considerations

- All components should be affected equally by treatment otherwise composite may be diluted
- All components should be clinically relevant
- All components should be defined as secondary and reported separately
Composite Endpoints

- In a review of 14 journals between 1 Jan 2000 and 1 Jan 2007
  - Of all cardiovascular trials (1231), 37% used composite endpoints
  - 98% of these included Mortality as a component
  - Typically 3 to 4 outcomes were included (based on the IQR)

“SYNTAX trial fails to show non-inferiority for DES” - Cardiovascular News, Sept 2008

- 1800 patients with left main/3 vessel disease comparing CABG vs. TAXUS
- MACCE = death/stroke/MI/Revascularizations
- 1-Year results were complex
- QOL results presented at ACC ’09
Example: SYNTAX (1-year Results)

<table>
<thead>
<tr>
<th></th>
<th>CABG</th>
<th>TAXUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>N Randomized</td>
<td>897</td>
<td>903</td>
</tr>
</tbody>
</table>
| MACCE (primary)      | 105 (12.4%)| 159 (17.8%)| P=.002
| Death                | 30 (3.5%)  | 39 (4.4%)  |
| Stroke               | 19 (2.2%)  | 5 (0.6%)   | P=.003
| MI                   | 28 (3.3%)  | 43 (4.8%)  |
| Death/MI/Stroke      | 65 (7.7%)  | 68 (7.6%)  | P=.98
| Revascularization    | 50 (5.9%)  | 120 (13.5%)| P<.0001
| PCI                  | 40 (4.7%)  | 102 (11.4%)|
| CABG                 | 11 (1.3%)  | 25 (2.8%)  |

### Example: HORIZONS (30 day Results)

- **Evaluating efficacy and safety in composite endpoint**

<table>
<thead>
<tr>
<th></th>
<th>Heparin +IIb/IIIa</th>
<th>Bivalirudin alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>1802</td>
<td>1800</td>
</tr>
<tr>
<td>Major Bleed</td>
<td>8.3%</td>
<td>4.9%</td>
</tr>
<tr>
<td>MACE*</td>
<td>5.5%</td>
<td>5.4%</td>
</tr>
<tr>
<td>Net Adverse Clinical Events</td>
<td>12.1%</td>
<td>9.2%</td>
</tr>
<tr>
<td>Death</td>
<td>3.1%</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

* MACE = Death/MI/Stroke/TLR

---

Stone, et al. JAMA 2008; 299 p 1903-1913
Survival analysis may count less severe outcomes over more severe (e.g. TLR vs. Mortality)

Example: MACE = Death/MI/Stroke/TLR

Patient A has TLR at 5 months but survives and feels well through end of trial (5 years)

→ Time to MACE = 5 months

Patient B has no TLR, MI or Stroke, but dies at 9 months

→ Time to MACE = 9 months

Q: Is patient A really worse off than patient B?
CompositeEndpoints: Some Pitfalls

• Can be misleading
  - MACE = Death/MI/Stroke/TLR
  - Why not
    MACE = TLR/MI/Death/Stroke?
  - Most important component listed first, but often has the lowest event rate (i.e. is the LEAST represented in the composite)
Composite Endpoints: Some Pitfalls

• All components tend to be treated equally
  - Can be weighted (but what are correct weights?)
• Some might presume that benefits relate to ALL components
• May dilute real treatment differences by including elements unaffected by treatment (impact on non-inferiority?)
• Can be difficult to interpret
CONCLUSIONS

• Both surrogates and composite endpoints are extremely useful for reducing sample sizes (increasing power)

• Both can get “better” treatments to patients faster

• But ...

• Proceed with CAUTION!