Back to Basics: Explaining sample size in outcome trials, are statisticians doing a thorough job?

Kevin J Carroll

Contents

- A typical conversation
- The importance of transparency around trial sizing and what can be expected in terms of actual outcomes
- Some recommendations

A typical conversation…

- Medic:
  “We need to show that ‘Efektiv’ is better than the current standard of care treatment in terms of clinical outcome. We need at least a 15% reduction in risk of the outcome to convince the medical community, regulators and formularies alike that ‘Efectiv’ is a genuine candidate to replace the current standard in managing patients. How big a trial are we looking at?’

A typical conversation…

- Statistician, after 10 minutes on nQuery:
  “We’ll need 1,591 events to provide 90% power at the 2-sided 5% significance level. Further, assuming 10% of patients have an outcome event after 1 year, and with plans for a 1 year accrual period and a 1 year minimum follow-up period, 11 800 patients will need to be randomised.”

3.5 years later….

- Risk reduction 10%
- p=0.036, 95% CI (1%, 18%)
- Positive trial
- …technically
- …but the Development Team (DT) are confused because the trial was powered to detect a 15% risk reduction, and yet a 10% risk reduction has come out significant…
- …did the statistician make a mistake?
- …commercial are grumbling since a 10% risk reduction is considered by key customer groups as, at best, marginal and will do nothing to help secure reimbursement…
- …nor provide sufficient differentiation vs drugs on the market and, it is feared, vs competitors who are also trialling against the current standard

So what went wrong?

- Nothing
- …at least, that is, from a purely mathematical point of view
- …but there was perhaps a failure early on to clearly and transparently lay out what the sizing and power calculation actually meant in terms of what differences would and would not reach significance
- …a failure, perhaps, manage expectations
- …a probable failure to have an open DT dialogue at very the outset regarding the Target Product Profile to ensure that the requirements vs differentiation and reimbursement were taken into account…

Some sample size fundamentals

- $\hat{\theta}$ is the estimated log HR, $\hat{\theta} \sim N\left(\theta, \frac{d}{4}\right)$, and d=total number of events
- H_0: $\theta=0$ vs H_1: $\theta=\Delta$
- Decision rule reject H_0 if $T \geq \theta_{crit}$;
  do not reject otherwise
- $d = \frac{4\left(z_\alpha + z_\beta\right)^2}{\theta^2}$
- $e^{\theta_{crit}} = e^{-2z_\alpha \sqrt{d^{-1}}}$ is the threshold value for the HR that flips the outcome between $p \leq 2\alpha$ and $p > 2\alpha$

Scope for confusion

- Medic clear that a 15% reduction in risk is required to convince physicians to change their usual practice
- In response, the statistician has sized the trial to test the hypothesis H_0: $\theta=0$ vs. H_1: $\theta=\ln(0.85)$
- Now, it is not uncommon for the medic to assume that this means that if the trial delivers a HR of 0.85 or better, then $p\leq0.05$, and otherwise the result will not be significant.
- However, the trial will in fact yield $p\leq0.05$ for a HR of 0.906 or better, i.e. a 9.4% risk reduction; if a 15% risk reduction is observed, then p= $2\left(1-\phi(z_\alpha+z_\beta)\right)$ =0.0012
- Thus, there is a danger that the trial will yield a statistically significant but clinically irrelevant result since the medic is blissfully unaware that differences smaller than the minimum difference required to be clinically persuasive will yield $p\leq0.05$.

So what might we do a little differently?

- Point out that if a specific log HR advantage needs to be realised of at least θ , with a result less than θ not being clinically persuasive even if it reached statistical significance, then the need is to hypothesise not θ but $\theta' = (1 + z_\beta z_\alpha^{-1})\theta$ under the alternative
- Provide $e^{\theta_{crit}}$
- Translate $e^{\theta_{crit}}$ into more meaningful terms by stating what this means in terms of the anticipated split of events between E and C

$$d_C = \frac{N}{2}\left\{\left(1 - \frac{2d_C}{N}\right)^{e^{\theta_{crit}}} - 1\right\} + d \quad \text{and} \quad d_E = d - d_C$$

Saying a little more

- Returning to the hypothetical dialogue between the medic and the statistician, after *“...11,800 patients will need to be randomised”* the statistician could add *“…though you should realise that in this trial I’m hypothesising a risk reduction of 15% which means that a lesser observed difference at the end of the trial would give p≤0.05; a risk reduction of at least 9.4%, corresponding to a difference in events of at least 832 (14.1%) versus 759 (12.9%), i.e. a difference of at least 73 (1.2%) events, would be significant. If the end result was actually a 15% risk reduction [corresponding to 855 (14.5%) versus 736 (12.5%) events, a difference of 119 (2%) events] the p-value would be well below 0.05; it would be around 0.0012. I’ve prepared a table with the details and a few other scenarios to look at...”*

Some scenarios

Total number of events required, d, and θ_{crit}					The overall probability of an event, π_{bar}, and the total number of patients, N					The number (%) of events on experimental, E, and control, C, to give θ_{crit}		
1-sided α level	Power	HR hypothesised under H_1, θ	Number of events req'd, d	HR (RR) result, θ_{crit}, required to achieve $p \leq 2\alpha$	Expected fraction of control patients with an event at 1 year, q_i	Accrual, A (yrs)	Minimum follow-up, F (yrs)	Prob. of an event over the trial period, π_{bar}	Total number of patients required, N	Number (%) of events on C	Number (%) of events on E	Difference (%) in events
2.5%	90%	0.750	508	0.840 (16.0%)	10%	1	1	0.1264	4019	275 (13.7%)	233 (11.6%)	42 (2.1%)
2.5%	90%	0.800	844	0.874 (12.6%)	10%	1	1	0.1307	6458	448 (13.9%)	396 (12.3%)	52 (1.6%)
2.5%	90%	0.850	1591	0.906 (9.4%)	10%	1	1	0.1348	11803	832 (14.1%)	759 (12.9%)	73 (1.2%)
2.5%	90%	0.764	580	0.850 (15%)	10%	1	1	0.1276	4545	312 (13.7%)	268 (11.8%)	44 (1.9%)

The best laid plans of mice and men…

- Event over the first six months lower than expected.
- Translates to a shortfall in the 1-year rate from 10% to 7.5%.
- A (reduced) total of 1200 events are now expected by the end of the trial.
- The question for the DT is what should be done?

The best laid plans of mice and men…

1. Keep N at 11,800 but extend follow-up to compensate to achieve 1591 events – question for the statistician is how much should follow-up be extended?
2. Keep the overall trial duration at A+F and but increase N to compensate to achieve 1591 events – question for the statistician is how much should N be increased?
3. Some combination of 1 and 2.
4. Do nothing. Check again in 6 months and hope the event rate has picked up.
5. Accept the lower event rate, do not change N or extend follow-up, and settle for the lower number of events, now expected to be 1200 – question for the statistician is what are the implications in terms of achieving a positive outcome for the trial?

Options 1, 2 and 3

- Event rate is small and reduced by 100ω percent

1. Increase minimum follow-up $F \rightarrow F + \left(\frac{A}{2} + F\right)(\omega - 1)^{-1}$
   - $\Rightarrow$ follow-up should increase by approximately 0.33 × 18 = 6 months, taking the overall duration from 24 months to 30 months
2. Increase number of patients $N \rightarrow N(1 - \omega)^{-1}$
   - $\Rightarrow$ number of patients should increase by approximately 33%, i.e. by 3900 patients, to around 15,700 pts.
3. A hybrid approach might be to increase *N* to 13,500 and follow for an additional 3 months which, based on the observed event rate, would deliver 1591 events by the end of the trial.

Options 4 and 5

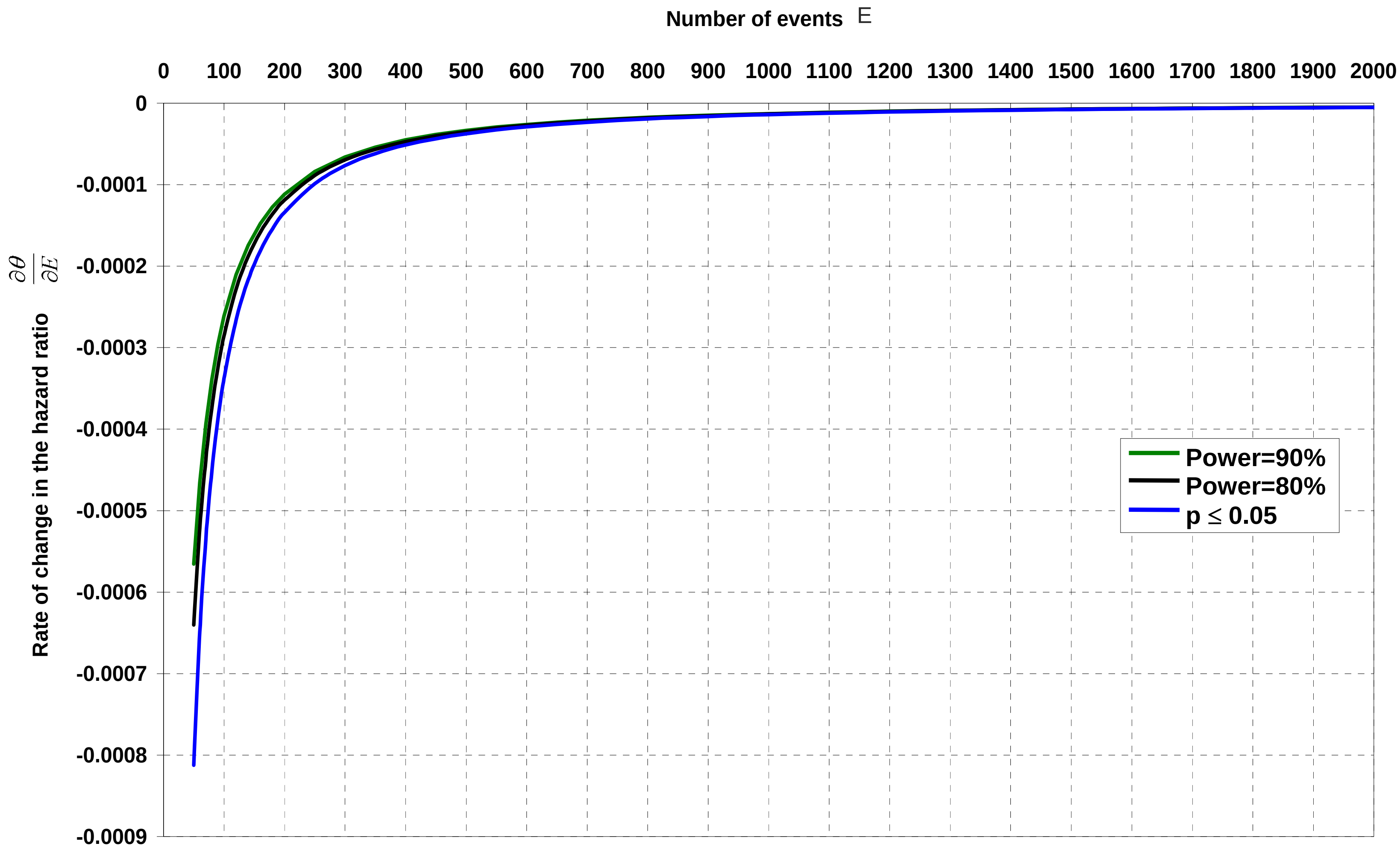
- Option 4 is a realistic option given relatively little data are available, but unlikely in practice.
- It’s more likely that some change to N and/or minimum follow-up would be proposed with the event rate continuously monitored such that if it was to increase, plans to increase trial size and/or follow-up of could be revisited.

- Option 5 is likely to be viewed as least favourable.
  - Team concerns regarding loss of power
  - Steering Committee concerns regarding power and perceptions
  - However, anxiety around settling for fewer events is often founded upon a poor appreciation of what is really being lost in terms of outcomes to yield significance or, conversely, what is really to be gained by pushing for the original target number at all costs.

Option 5

	Number of events at end of trial	HR (RR) result required to achieve p≤0.05	Number (%) of events on control	Number (%) of events on experimental	Difference (%) in events	Power for Hypothesised HR=0.85	Hypothesised HR for power=90%
As originally planned	1591	0.906 (9.4%)	832 (14.1%)	759 (12.9%)	73 (1.2%)	90%	0.85
Accepting fewer events	1200	0.893 (10.7%)	632 (10.7%)	568 (9.6%)	64 (1.1%)	80.4%	0.83

Rate of change in the ‘detectable’ hazard ratio vs total number of events



Recommendations (1 of 3)

1. Avoid the use of loose and imprecise language when describing basic sample size and power; avoid statements such as “the trial is sized to detect a difference of Δ with 90% power at the 2-sided 5% significance level” and rather use more correct language such as “**the trial is sized to test the null hypothesis H_0:the true difference=0 versus the alternative H_1:the true difference=Δ with Type I and II errors of 5% and 10% respectively [or with a 5% 2-sided significance level and 90% power]**”.

2. At the outset, **statisticians should proactively lead the DT in a dialogue around the Target Product Profile**, and, thus help avoid confusion and concern when the trial subsequently delivers a significant result for an observed outcome less than that which was hypothesised.

3. Always and routinely provide the critical value, θ_{crit}, of the test, when describing sample size and power. Take time to carefully explain to the DT the difference between the outcome and the hypothesis.

Recommendations (2 of 3)

4. Be clear that, if θ is hypothesised, then there is a 50% chance the observed difference will exceed θ with $p\leq0.0012$; a 40% chance that the observed difference will be less than θ but still significant, with $0.05\leq p<0.0012$, and with a difference of $0.6\times\theta$ yielding $p=0.05$; and a 10% chance that the observed difference will be less than $0.6\times\theta$ with p=NS.

5. Point out that if a specific advantage of, say, at least θ, needs to be realised to be persuasive, then the need is to hypothesise not θ but rather $\theta' = (1 + z_\beta z_\alpha^{-1})\theta$ and if E events are needed when θ is hypothesised, $E' = E\left(1 + z_\beta z_\alpha^{-1}\right)^{-2}$ events are needed when θ' is hypothesized.

   The implications of hypothesising θ' as opposed to q should to be carefully explained to the DT as the jump in expectation could be considered biologically implausible to such an extent as to render the entire trial nonviable.

Recommendations (3 of 3)

6. For time to event trials, translate the sample size, hypothesized effect Δ and θ_{crit} into more meaningful terms by stating what this means in terms of the anticipated split of events between drug and control

7. Point out that, when the event rate is relatively low, and early blinded trial data suggest the event rate may be reduced by 100ω percent relative to initial expectations, then to compensate either the minimum follow-up *F* must be increased by approximately $(\omega - 1)^{-1}$ times the overall duration of the trial, *A+F,* or the target number of events must be increased by approximately 100$(\omega - 1)^{-1}$ percent

8. Highlight that once around 1000 to 1200 events are achieved, the practical gain in accumulating further events is marginal such that substantial jumps in the number of additional events beyond 1000 to 1200 events are required to make a meaningful difference to the ‘detectable’ log HR.

Back ups