Sample size calculation in non-inferiority trials with binary endpoint: scale matters

Nan van Geloven

Leiden University Medical Center
n.van_geloven@lumc.nl

PSI, 3rd June 2019

Outline

Noninferiority trial

- Traditionally: proving non-inferiority versus active control indirectly proves superiority versus placebo
- Nowadays: new treatments that are less invasive, less side effects, cheaper etc only have to prove no less effective than current treatment. Examples: lower doses, different administration routes
- number of NI trials increased by a factor 6 from 2005 to 2015[1]

[1] Mauri and D'Agostino, NEJM 2017

Noninferiority margin

- 'no less effective' is defined by the non-inferiority margin: a boundary of acceptable difference
- In case of binary outcome this can be specified on three different scales
- π_c the success rate in the control arm and π_t the success rate in the treatment arm

risk difference (RD)	$\pi_t - \pi_c$
relative risk (RR)	π_t / π_c
odds ratio (OR)	$(\frac{\pi_t}{1-\pi_t})/(\frac{\pi_c}{1-\pi_c})$

What is used in practice?

review Hilton (2010)

- using criteria ‘New Engl J Med[JO] AND noninferiority’
- 32 distinct RCTs published between 2001-2009
- 21 RCTs with binary outcomes, among these: RD-17, RR-2, OR-2

review Buurkes (2019)

- using criteria ‘New Engl J Med[JO] AND noninferiority’
- 63 distinct RCTs published between 2016-2019
- 24 RCTs with binary outcomes, among these: RD-16, RR-2, OR-6

Outline

INES trial design[2]

- couples with unexplained subfertility
- 3-arm trial, randomised between two types of IVF and one type of IUI, here simplify to 2 arms
- primary outcome: 12 month single birth rate
- IVF expected to have lower multiple birth rates, therefor INES aimed to show non-inferiority of IVF arms for single birth rate compared to IUI

[2] Bensdorp et al, 2015 BMJ

INES trial results

- *Design paper:* ...expected birth rate of 40%, 190 patients per group are required to exclude a difference of 12.5% or more to the detriment of IVF (one sided $\alpha = 0.05$, power=80%)

INES trial results

- *Design paper:* ...expected birth rate of 40%, 190 patients per group are required to exclude a difference of 12.5% or more to the detriment of IVF (one sided $\alpha = 0.05$, power=80%)
- *Main paper:* rates IVF-MNC 43%, IUI 47%
  ...corresponds to a risk, relative to IUI, of 0.91 (0.73 to 1.14) for IVF-MNC. This 95% confidence interval does not extend below the predefined threshold of 0.69 for inferiority.

INES trial - recalculation of results

	95% CI	90% CI	NI
RR	0.91 (0.73 to 1.13)		met
RD		-4% (-12% to 4%)	met
RR		0.91 (0.76 to 1.10)	met
RD	-4% (-14% to 6%)		failed

RD, RR or OR in the INES trial

- If designed on 40% vs 27.5% on RR ($\delta_{RR} = 0.69$) in stead of RD only 135 patients per arm would be needed in stead of 190
- Note that this uses 'success rates' (higher is better). If formulated with 'failure rates' (60% vs 72.5% $\rightarrow \delta_{RR} = 1.21$), RR requires 235 patients
- Had they used OR (40% vs 27.5% $\rightarrow \delta_{OR} = 0.57$), 165 patients per arm would be needed

Outline

Switching margin scale

- risk difference:

$$H_0 : \pi_t - \pi_c \leq \delta_{RD}$$
$$H_1 : \pi_t - \pi_c > \delta_{RD}$$

- risk ratio

$$H_0 : \ln(\pi_t) - \ln(\pi_c) \leq \delta_{\ln RR}$$
$$H_1 : \ln(\pi_t) - \ln(\pi_c) > \delta_{\ln RR}$$
$$\delta_{\ln(RR)} = \ln(\pi_c + \delta_{RD}) - \ln(\pi_c)$$

- odds ratio

$$H_0 : \ln(\frac{\pi_t}{1-\pi_t}) - \ln(\frac{\pi_c}{1-\pi_c}) \leq \delta_{\ln OR}$$
$$H_1 : \ln(\frac{\pi_t}{1-\pi_t}) - \ln(\frac{\pi_c}{1-\pi_c}) > \delta_{\ln OR}$$
$$\delta_{\ln(OR)} = \ln(1 + \frac{\delta_{RD}}{\pi_c(1-\pi_c-\delta_{RD})})$$

Sample size formulas based on z-tests

If we assume $\pi_t = \pi_c$, then n needed per group

RD scale

$$n^{RD} \geq \frac{2(z_{1-\alpha}+z_{1-\beta})^2 \pi_c(1-\pi_c)}{\delta^2_{RD}}$$

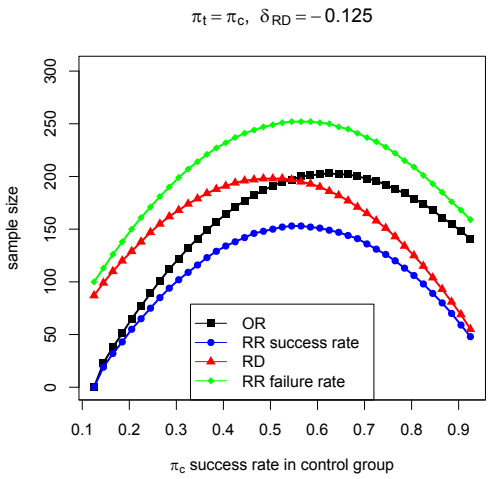
RR scale

$$n^{RR} \geq \frac{2(z_{1-\alpha}+z_{1-\beta})^2 \frac{(1-\pi_c)}{\pi_c}}{\delta^2_{\ln(RR)}}$$

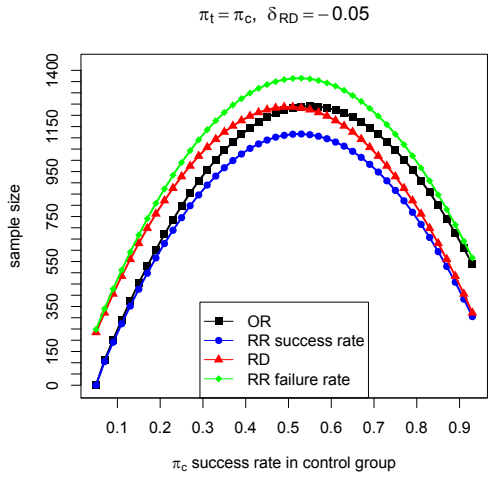
OR scale

$$n^{OR} \geq \frac{2(z_{1-\alpha}+z_{1-\beta})^2 \frac{1}{\pi_c(1-\pi_c)}}{\delta^2_{\ln(OR)}}$$

Comparison


$\pi_t = \pi_c,\ \delta_{RD} = -0.125$
sample size
0
50
100
150
200
250
300
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
π_c success rate in control group
OR
RR success rate
RD
RR failure rate

Also holds for smaller margins



n_{RD} always higher than n_{RR} with success rates ($\pi_t = \pi_c$)

$$\frac{n^{RD}}{n^{RR}} = \frac{2(z_{1-\alpha} + z_{1-\beta})^2 \pi_c (1-\pi_c)}{\delta_{RD}^2} \Big/ \frac{2(z_{1-\alpha} + z_{1-\beta})^2 \frac{(1-\pi_c)}{\pi_c}}{\delta_{\ln(RR)}^2}$$

$$= \frac{\pi_c^2}{(\delta_{RD})^2} (\ln(\pi_c + \delta_{RD}) - \ln(\pi_c))^2$$

$$= \frac{1}{(\frac{\delta_{RD}}{\pi_c})^2} (\ln(1 + \frac{\delta_{RD}}{\pi_c}))^2$$

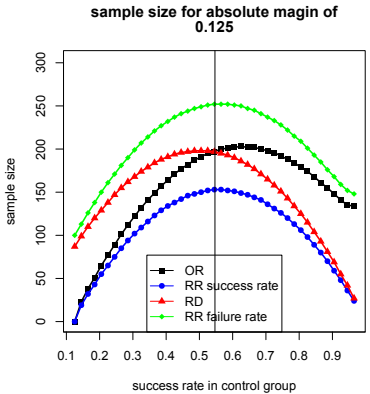
$$= (\frac{\ln(1+x)}{x})^2, \text{with x} = \frac{\delta_{\text{RD}}}{\pi_{\text{c}}} \in (-1, 0)$$

$$\in (1, +\infty)$$

n_{RD} lower than n_{OR} with large success rates ($\pi_t = \pi_c$)

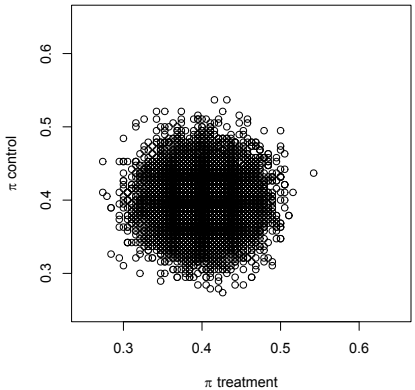
Rousson (2008) proves that $n_{RD} < n_{OR}$, when $\pi_c \geq \frac{1}{1-\delta_{OR}} + \frac{1}{\delta_{\ln OR}}$

in the INES example from $\pi_c \geq 0.55$



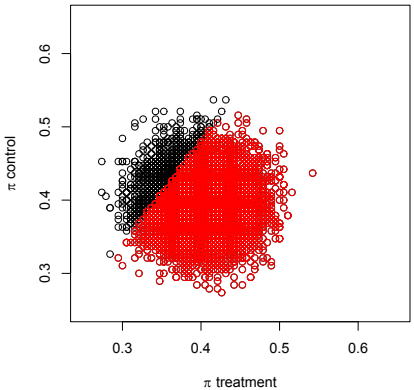
How come?

10000 trials simulated under $H_a : \pi_c = \pi_t = 0.4, \delta_{RD} = -0.125, n = 190$



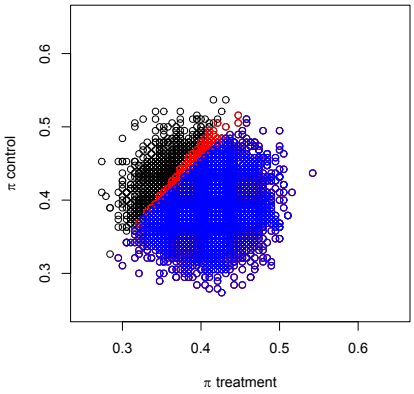
How come?

10000 trials simulated under $H_a : \pi_c = \pi_t = 0.4, \delta_{RD} = -0.125, n = 190$



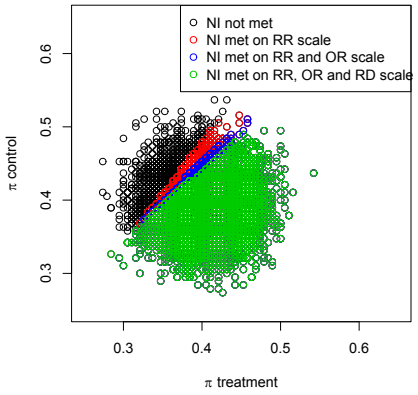
How come?

10000 trials simulated under $H_a : \pi_c = \pi_t = 0.4, \delta_{RD} = -0.125, n = 190$



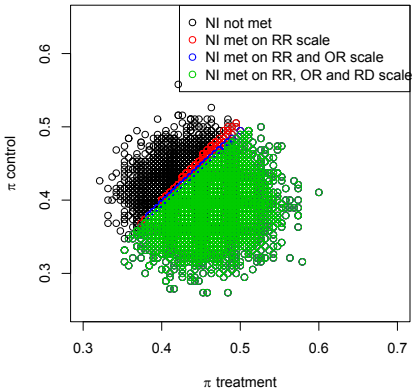
How come?

10000 trials simulated under $H_a : \pi_c = \pi_t = 0.4, \delta_{RD} = -0.125, n = 190$



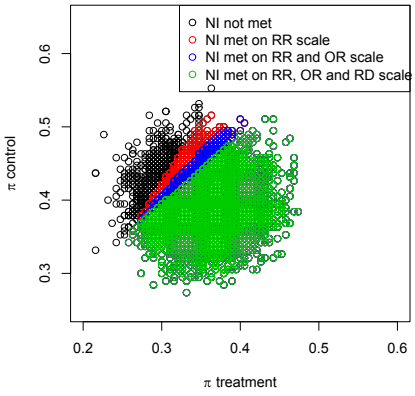
$\pi_t > \pi_c$ under H_a

10000 trials simulated under $H_a : \pi_c = 0.4, \pi_t = 0.45, \delta_{RD} = -0.075, n = 190$



$\pi_t < \pi_c$ under H_a

10000 trials simulated under $H_a : \pi_c = 0.4, \pi_t = 0.35, \delta_{RD} = -0.175, n = 190$



Outline

Other advantages / disadvantages

- RD is generally preferred by clinicians, e.g, author guideline Lancet Neurology

  "the abstract should include ... the difference between groups ... (absolute differences are more useful than relative ones)."

- RD is bounded by (-1,1), which may adversely affect inference, especially when success rates are close to the bounds
- RR not symmetric wrt formulating in success rates or failure rates
- RR with success rates: absolute margin increases when the control treatment has higher success rate, can be undesirable
- OR has statistical advantages: unbounded, stable and easy extension to covariate adjustments

RD, RR or OR: what to choose

- Hilton (StatMed 2008): design on relative failure rates $\rightarrow$ also sufficient power for analysis on absolute scale and for per protocol analysis with covariate adjustment
- Wellek (Biom J 2005): use RD in communication with clinicians, switch to more sensible OR scale directly after that
- Rousson (Biom J 2008): plan based on OR, then switch to RD $\rightarrow$ power gain if success rate in control group is large (>0.5-0.6)
- My advice: if control group turns out to have higher than anticipated success rate, how would treatment arm have to do to still be non-inferior?

Conclusions

- Scale matters when planning non-inferiority trials
- Switching scale in analysis phase (e.g. to accommodate covariate adjustment) affects power
- Bigger impact with larger margins
- Overview of different scenarios in the making, all feedback welcome (n.van_geloven@lumc.nl)

References

- Bensdorp AJ et al. Prevention of multiple pregnancies in couples with unexplained or mild male subfertility: randomised controlled trial of in vitro fertilisation with single embryo transfer or in vitro fertilisation in modified natural cycle compared with intrauterine insemination with controlled ovarian hyperstimulation. BMJ 2015
- Van Geloven N. Noninferiority or Superiority? Rapid response https://www.bmj.com/content/350/bmj.g7771/rr
- Rousson V and Seifert B. A mixed approach for proving non-inferiority in clinical trials with binary endpoints. Biom J 2008
- Wellek S. Statistical methods for the analysis of two-arm non-inferiority trials with binary outcomes. Biom J 2005
- Hilton JF. Noninferiority trial designs for odds ratios and risk differences. Stat Med 2008
- Mauri L and D'Agostino. Challenges in the design and interpretation of noninferiority trials. NEJM 2017