

Frequentist thoughts about borrowing of historical control data

Steven Teerenstra (Radboudumc)

Joint work with

Leonie Theis (MHH)

Anika Großhennig (MHH)

Armin Koch (MHH)

Background

- Many methods for borrowing historical data proposed in the last decade(s)
- Many of those were in Bayesian framework
 - no (full) type I error control (e.g. only in ‘sweet’ spots)
 - tuning [bias/type I error inflation] for [power/precision]
- Adaptive lasso can be tuned to achieve similar trade-off
 - Frequentist
- Borrowing => no experimental type I error => ‘type I error’
 - ‘type I error’=probability false positive conclusion based on mixed experimental and observational data
- Can we borrow essentially frequentist?
 - at the same (full) ‘t1e’ control as an RCT would have?

Setting

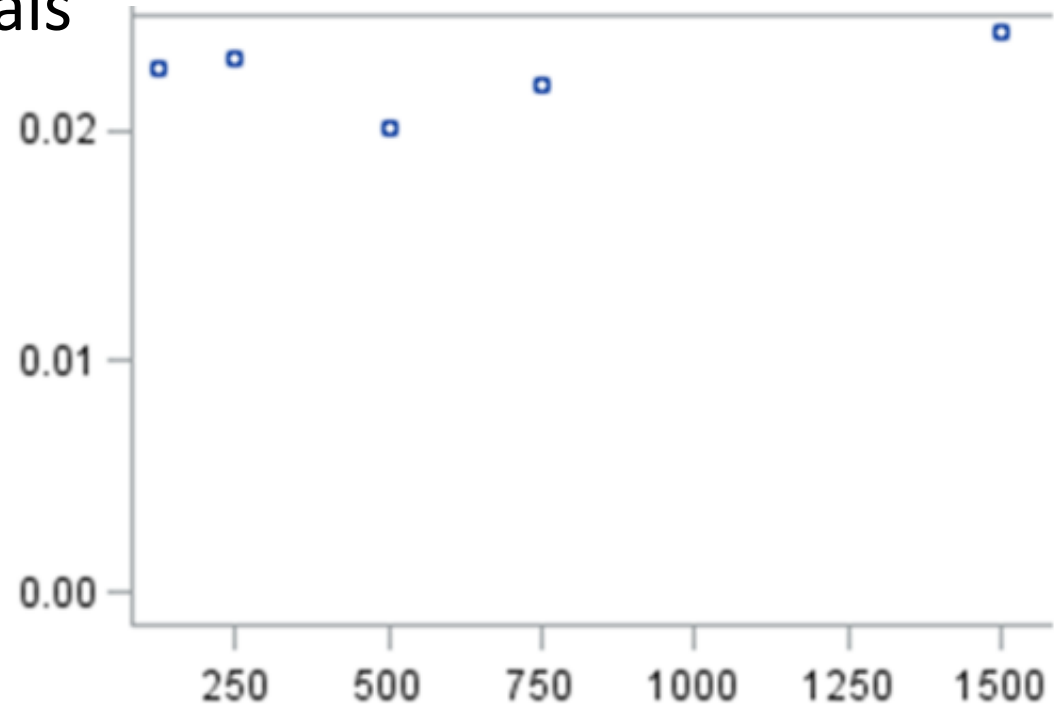
- Binary outcome parameter: “response”
 - Higher is better
- Parallel group, two arms: experimental vs. control
- Randomize (2:1)
 - experimental arm: 250 patients
 - Control arm: 125 patients
- Historical controls:
 - 125, 250, 500, 750, 1500 patients

Setting

- Randomize 250 vs 125 patients
 - Historical controls: 125, 250, 500, 750, 1500 patients
- ‘Type I error’ under $H_0: p_{exp} = p_{rand.ctrl} = p_{hist.ctrl}$
 - $p=0.1, 0.3, \mathbf{0.5}$ ← focus on, others similar
 - Chi-square test with continuity-correction
 - 10.000 simulated trials

- ‘Type I error’
Naive pooling

$$p_{exp} = p_{rand.ctrl} = p_{hist.ctrl}$$



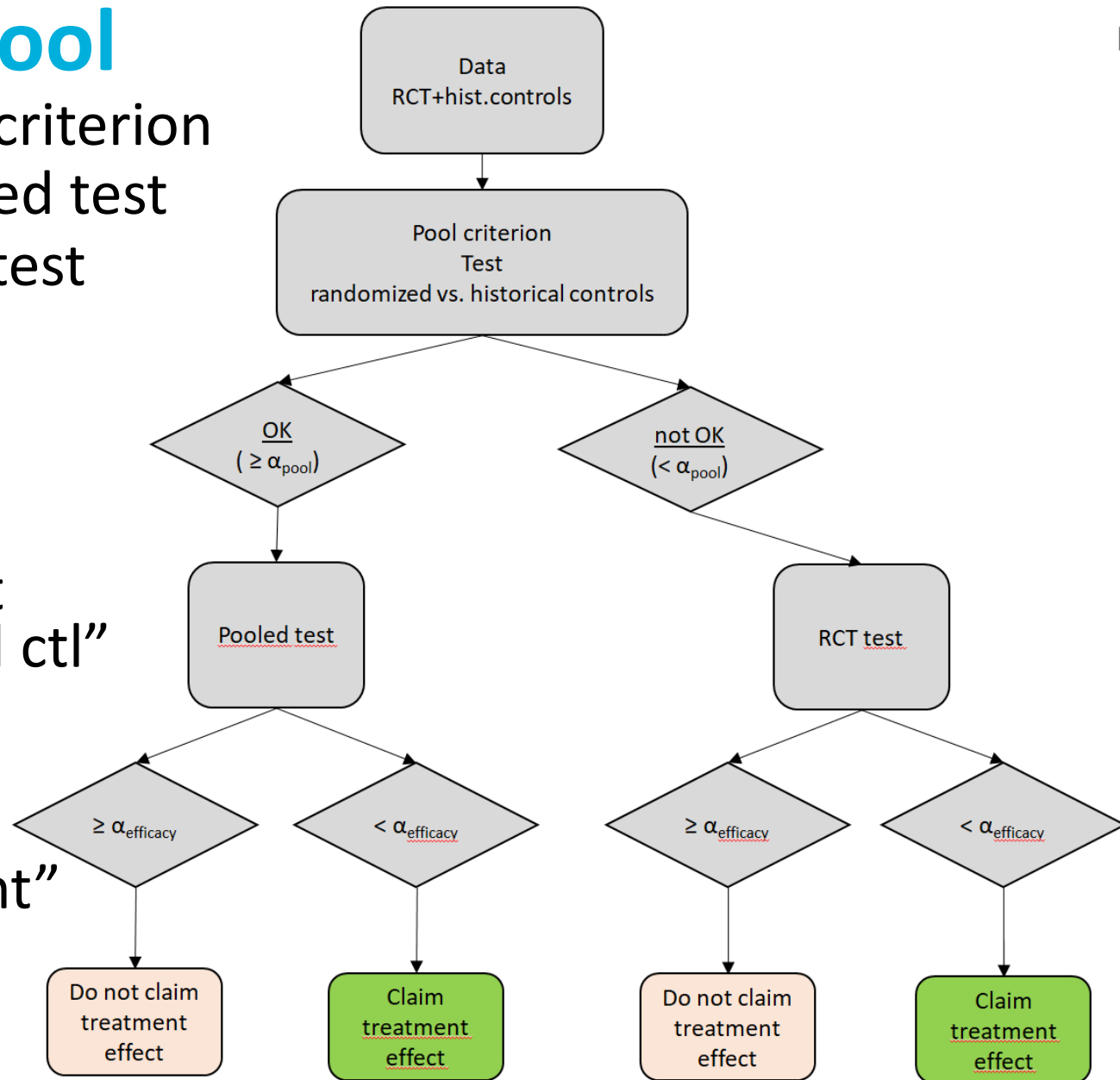
Test-then-pool

- First assess pool criterion
 - OK: then pooled test
 - ~~OK~~: then RCT test

- Pool criterion: “OK” =

“Historical ctl not sign. different from randomized ctl”

- “not sign. different” tested at level α_{pool}



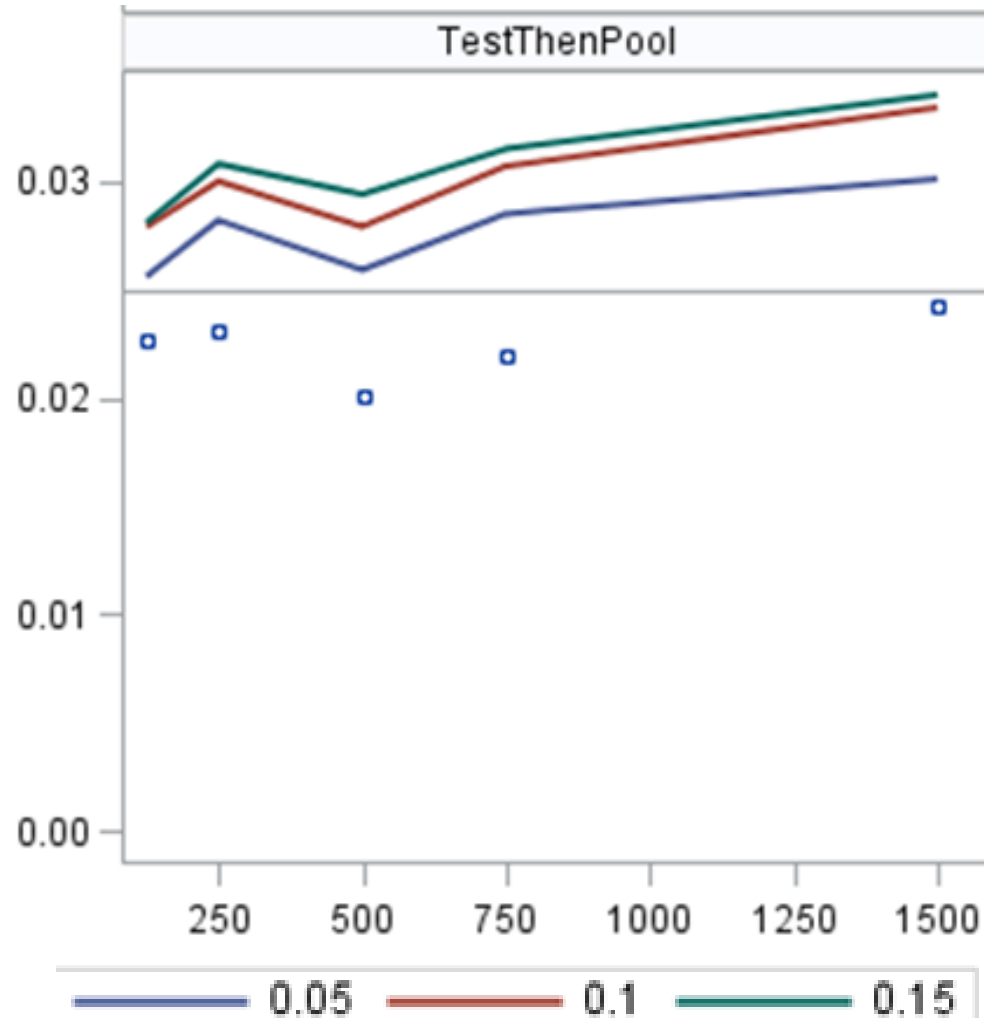
Type I error $H_0: p_{exp} = p_{hist} = p_{rand.ctl} = 0.5$

- Viele et al.:
 - larger difference
hist. vs rand. controls
increases type I error
- Here: what is impact of
 - sample size
 - α_{pool}
 under the ‘ideal’ situation

$$p_{exp} = p_{hist} = p_{rand.ctl}$$

=> ‘type I error’ inflation, if

- Increasing with sample size
historical controls;
- Increasing when more strict on
when to pool
(i.e. increase α_{pool} from 0.05 to 0.15)



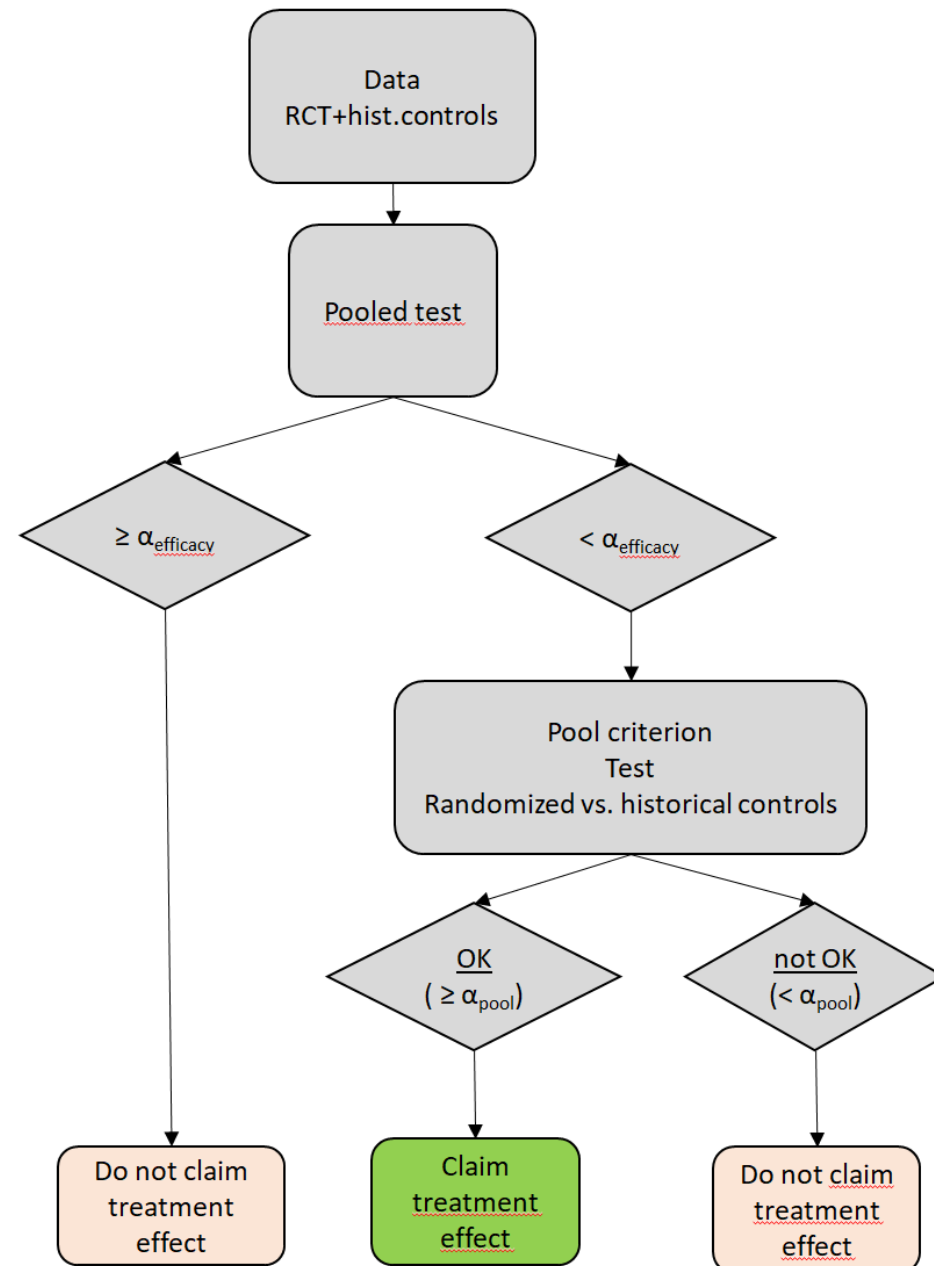
Can test-then-pool be repaired?

- Inflation ‘type I error’ comes primarily because “not sign.different” allows cases where the effect is increased (under the null of no difference).
- Pool criterion: “not sign.different” => hist. ctl “not worse”



Pool-then-test

- First pool and test efficacy
- if success in pooled test then assess pool criterion:
 - OK: claim efficacy
 - ~~OK~~: do not claim
- Pool criterion:
“OK” =
historical controls
not sign.different
from randomized controls



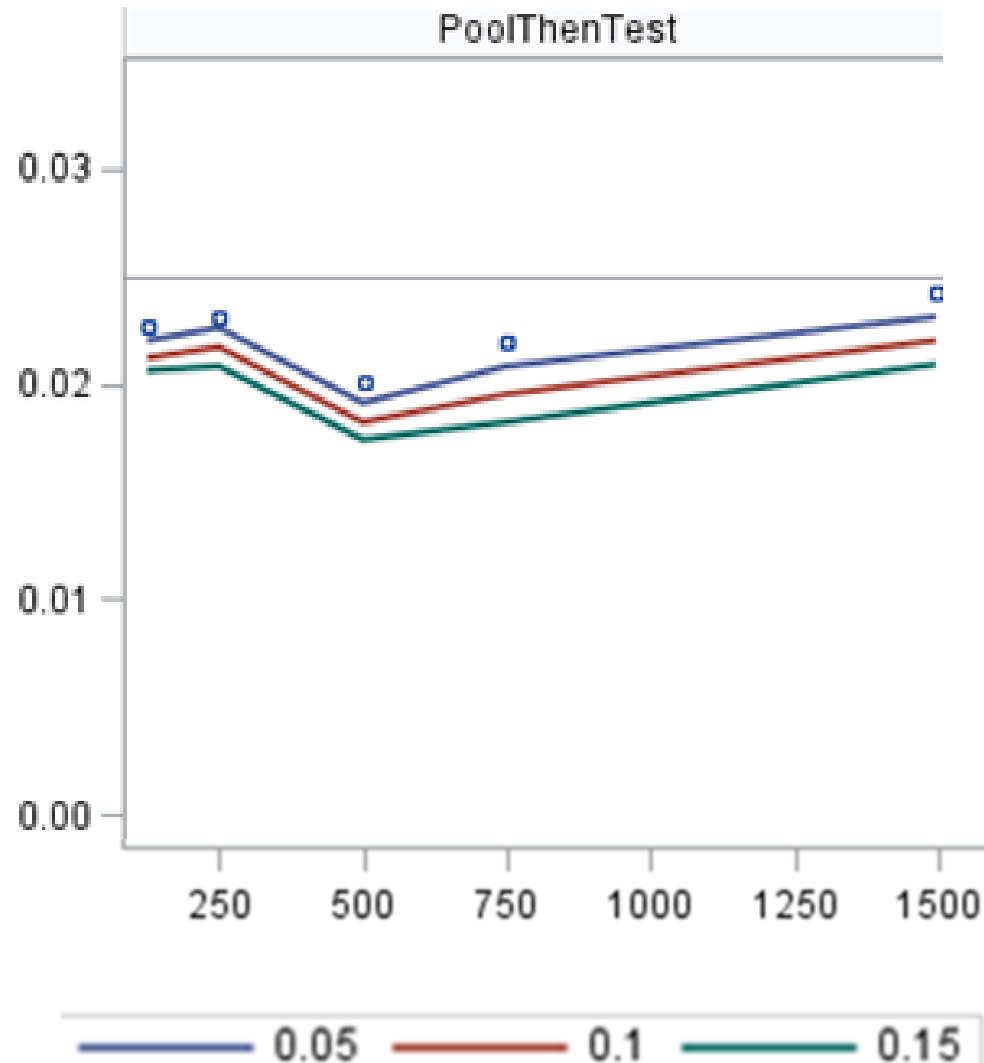
Type I error $H_0: p_{exp} = p_{hist} = p_{rand.ctl} = 0.5$

- Under 'ideal' situation

$$p_{exp} = p_{hist} = p_{rand.ctl}$$

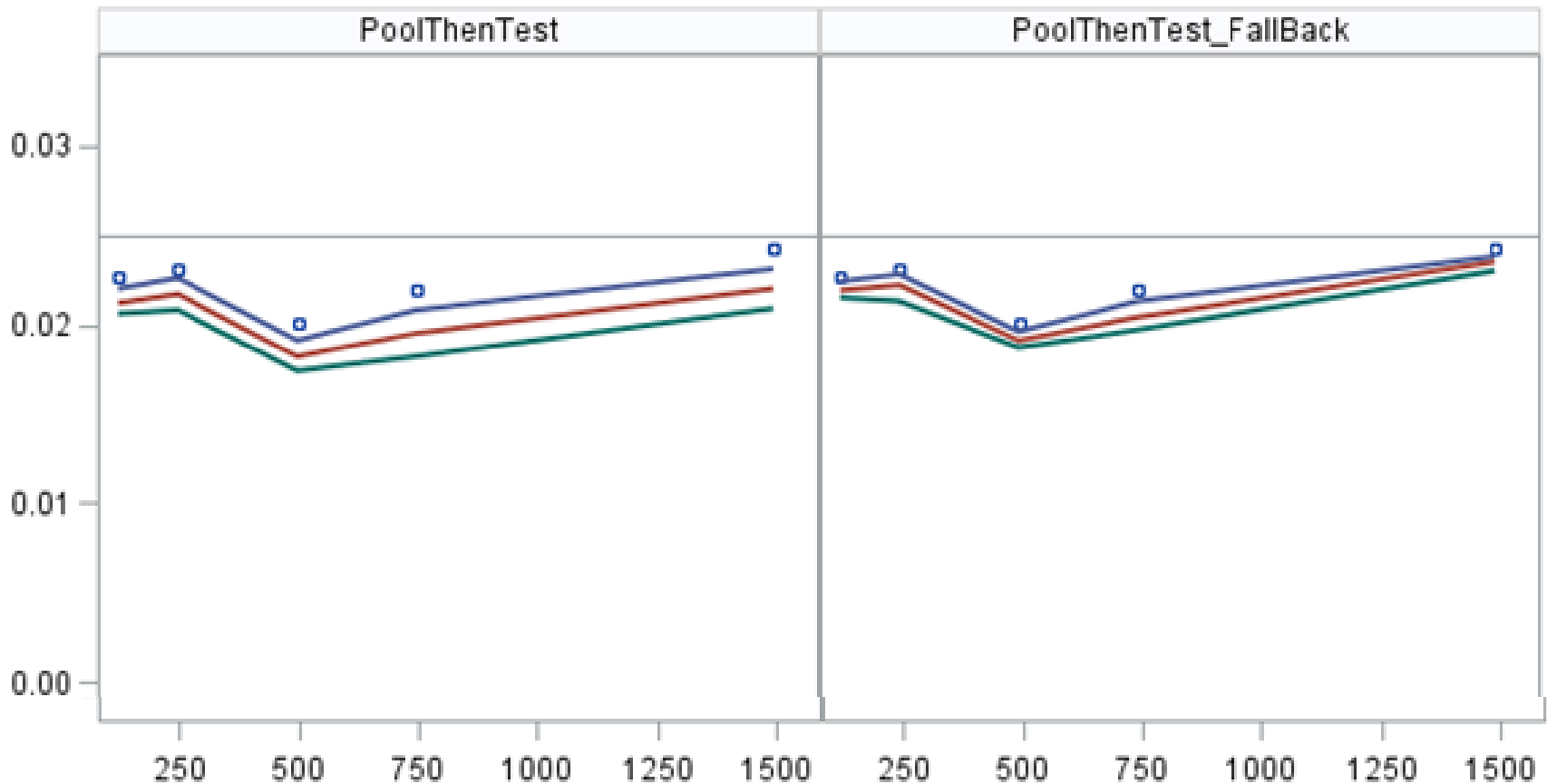
- A little worse than (naive) pooled test (see dots)
- More conservative if being more strict on 'similarity' (i.e. increase α_{pool} from 0.05 to 0.15)

- Always 'type I error' control as it is gatekept by the 'type I error' control of the pooled test that comes first.



Can pool-then-test be improved?

- If the (naive) pooled test is successful, but the pooling criterion fails, then no efficacy is claimed.
- What if we 'fall back' on the RCT test in that case?



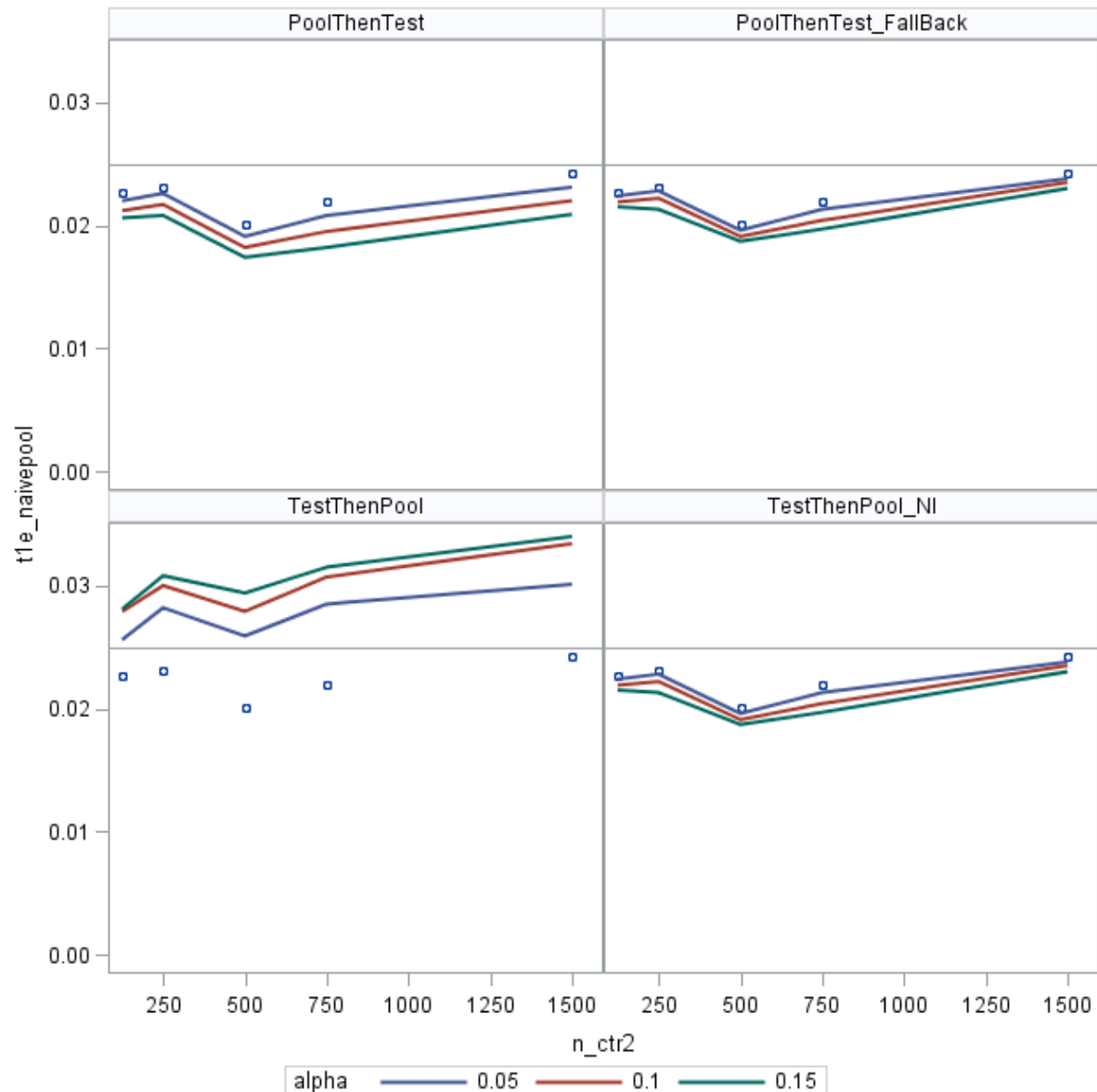
Type I error control of all approaches

- Pool-then-test with 'fall back'

similar to

Test-then-pool with "not worse" pool criterion

- Both rather close to naive pooling



Conclusion

- Simple frequentist methods for borrowing are possible
 - easy in implementation
 - protect ‘type I error’ on the same level as RCT
 - Not too far from ‘type I error’ of naive pooled test
(the most powerful test when $p_{exp} = p_{hist.ctl} = p_{rand.ctl}$)
- When there is need to use external data, better interpretation when explicitly assessing and justifying whether to pool or not:
 - pooling criteria allow an explicit decision whether to pool or not
 - Pooling criteria with clinically motivated margins possible
 - e.g. “not stat.different” => “equivalent by a clinical margin”
=> better clinical interpretation of
when data are pooled and when not