Basket Trials

Berry Adaptive Design and FACTS Webinar
Kert Viele
X/Twitter @KertViele, LinkedIn

Outline

The Basics of Basket Trials

- What problem are we solving?
- How do Basket trials work?
- What advantages might occur in practice?

FACTS implementation

- Changing decision rules
- Changing final analysis

Advanced Topics

- What is type 1 error control?
- Borrowing information

Regulatory

- Uncertain landscape in phase 3

What problem are we solving?

- Trials enroll a prespecified population

- We may not expect a therapy to work in everyone.
  - Every indication where there are first line, second line, etc.
  - Stroke – time since last seen well, stroke severity
  - Oncology - targeted therapies
  - Epilepsy – different syndromes
  - etc.

- May wish to identify a population in phase 2 to go to phase 3
- May wish to gain approval for subgroups

What problem are we solving?

- How should we enroll?
  - enroll broadly….might dilute treatment effects
  - enroll narrowly….miss people who may benefit
  - in either, performing a single analysis may result in giving a therapy to everything that only benefits some.
  - anything better?

- Basket trials attempt to focus on subgroups where the therapy is effective, while eliminating subgroups where therapy is ineffective

- We are likely never sufficiently powered for all subgroups

Basket trials

- Start by enrolling a specific (usually broad) population
  - Population divided into prespecified, disjoint subgroups
  - (not considering continuous thresholds or ordered subgroups here)

- Perform periodic interim analysis
  - Analyse each prespecified subgroup
  - Drop any subgroups that are performing poorly
    - (can also open new subgroups…not considered here)
  - May declare success on subgroups performing well (optional!)

- Final analysis, several options
  - Analyse remaining groups in study, typically either pooled or with some sort of information borrowing across groups
  - Separate analyses reduces you to separate trials in each subgroup

Accrual

- Multi-arm trials differ from multiple subgroup trials
- If an arm stops mid-trial, trial accrual rate unchanged
  - we can assign more future patient to continuing arms
- When a subpopulation is stopped, we lose part of our accrual
  - Trial duration/feasibility may be affected
  - Cannot always increase allocation to other subpopulations, even if budget exists. The patients may not be available.
- Thus operational and inferential issues must both be considered

Simple example

- Oncology example with 10 rare subgroups, phase 2
- Want to select a population to continue into phase 3
- Single arm in each group, would like to be superior to 10%
  - 15 pts/grp
  - analyzed separately, can achieve 5.6% type 1 error, 82.7% power
  - declare group effective if 4/15 or more responses

- Most important question….what is “good performance”?
  - Limited agreement on this
  - Of course, getting every group correct is ideal, but difficult
  - Typically a tradeoff exists, with different designs performing well depending on the underlying truth.

Sample scenarios and Metrics

Group	1	2	3	4	5	6	7	8	9	10
All 10	10	10	10	10	10	10	10	10	10	10
Two 35	35	35	10	10	10	10	10	10	10	10
Half 35	35	35	35	35	35	10	10	10	10	10
Eight 35	35	35	35	35	35	35	35	35	10	10
All 35	35	35	35	35	35	35	35	35	35	35

Useful to consider inferential metrics such as

Number of groups correctly identified (broken down by effective/ineffective groups)

Can be reformulated as percentage of total population treated identified correctly

Average sample size (total and within each group)

Average duration of trial (noting issues with accrual)

Separate and pooled analyses

Group	Avg nulls successful	Avg alts successful	Avg total success (% alt)
All 10	0.556 / 10	NA	0.556 (0%)
Two 35	0.444 / 8	1.655 / 2	2.099 (78.8%)
Half 35	0.278 / 5	4.137 / 5	4.414 (93.7%)
Eight 35	0.111 / 2	6.618 / 8	6.730 (98.3%)
All 35	NA	8.273 / 10	8.273 (100%)

Group	Avg nulls successful	Avg alts successful	Avg total success (% alt)
All 10	0.440 / 10	NA	0.440 (0%)
Two 35	4.683 / 8	1.171 / 2	5.853 (20%)
Half 35	4.977 / 5	4.977 / 5	9.954 (50%)
Eight 35	2.000 / 2	8.000 / 8	10.0 (80%)
All 35	NA	10.0 / 10	10.0 (100%)

Separate analyses

15/grp, need 4/15 to claim efficacy
4.4% type 1 error, 82.7% power
in each group

Averages are over the 2^10 possible conclusions of (yes/no) for each of the 10 groups

Pooled analysis

15/grp, pooled over all 150 pts
need 22/150 to claim efficacy
in all groups

averages simply average over the 2 possible trial conclusions
yes in all, no in all

Separate and pooled analyses

Group	Avg nulls successful	Avg alts successful	Avg total success (% alt)
All 10	0.556 / 10	NA	0.556 (0%)
Two 35	0.444 / 8	1.655 / 2	2.099 (78.8%)
Half 35	0.278 / 5	4.137 / 5	4.414 (93.7%)
Eight 35	0.111 / 2	6.618 / 8	6.730 (98.3%)
All 35	NA	8.273 / 10	8.273 (100%)

Group	Avg nulls successful	Avg alts successful	Avg total success (% alt)
All 10	0.440 / 10	NA	0.440 (0%)
Two 35	4.683 / 8	1.171 / 2	5.853 (20%)
Half 35	4.977 / 5	4.977 / 5	9.954 (50%)
Eight 35	2.000 / 2	8.000 / 8	10.0 (80%)
All 35	NA	10.0 / 10	10.0 (100%)

Pooled analyses are decently powered even when only two groups effective. Carry a lot of null groups forward

Pooled absolutely superior in “all 35” and equal to separate in “all 10”.

Separate performs better at screening out null groups. While not all alternative groups are successful, most of the successful groups are effective

Can we do better?

(open question, which performs better at “eight 35”?)

Common themes

- Periodic interims
  - Based on overall sample size or time
  - Very difficult to base on group specific requirements
    - differential enrollment across groups is likely
  - Allow for groups to stop for futility or success

- Modeling across groups
  - Bayesians often employ hierarchical models
  - Frequentist often pool groups that are still enrolling at trial end

An adaptive design

- Interims at 50 and 100 patients (final analysis still at 150)
- Stop group at interim if
  - Pr(beat 10%) < 0.60 (for futility)
  - Pr(beat 10%) > 0.95 (for success)

- Final analysis at N=150
  - Group declared successful if Pr(beat 10%) > 0.95

- Analyses performed with a hierarchical model
  - Each of the 10 rates (logit transformed) assume to arise from a normal distribution, with priors on the mean and variance.
  - Allows dynamic borrowing between groups

Adding adaptive

Separate with adaptive	Avg nulls successful	Avg alts successful	%alt continuing	E[N]
All 10	0.556 / 0.085	NA	0.000 / 0.000	150 / 88.8
Two 35	0.444 / 0.402	1.655 / 1.122	0.788 / 0.736	150 / 125.8
Half 35	0.278 / 0.879	4.137 / 3.863	0.937 / 0.815	150 / 136.7
Eight 35	0.111 / 0.972	6.618 / 7.126	0.983 / 0.880	150 / 112.4
All 35	NA	8.273 / 9.842	1.000 / 1.000	150 / 73.6

Pooled	Avg nulls successful	Avg alts successful	Avg total success (% alt)
All 10	0.440 / 10	NA	0.440 (0%)
Two 35	4.683 / 8	1.171 / 2	5.853 (20%)
Half 35	4.977 / 5	4.977 / 5	9.954 (50%)
Eight 35	2.000 / 2	8.000 / 8	10.0 (80%)
All 35	NA	10.0 / 10	10.0 (100%)

No method "dominates"
Depends on scenario

In the joint null, the adaptive trial is vastly superior

1) E[N]=88.8 vs 150
2) group type 1 error 0.0085
3) Pr(go to phase) =
   4-5% for pooled,adaptive
   44% for separate!

Separate and pooled analyses

Separate	Avg nulls successful	Avg alts successful	%alt continuing	E[N]
All 10	0.556 / 0.085	NA	0.000 / 0.000	150 / 88.8
Two 35	0.444 / 0.402	1.655 / 1.122	0.788 / 0.736	150 / 125.8
Half 35	0.278 / 0.879	4.137 / 3.863	0.937 / 0.815	150 / 136.7
Eight 35	0.111 / 0.972	6.618 / 7.126	0.983 / 0.880	150 / 112.4
All 35	NA	8.273 / 9.842	1.000 / 1.000	150 / 73.6

Pooled	Avg nulls successful	Avg alts successful	Avg total success (% alt)
All 10	0.440 / 10	NA	0.440 (0%)
Two 35	4.683 / 8	1.171 / 2	5.853 (20%)
Half 35	4.977 / 5	4.977 / 5	9.954 (50%)
Eight 35	2.000 / 2	8.000 / 8	10.0 (80%)
All 35	NA	10.0 / 10	10.0 (100%)

For two 35

Separate and adaptive screen out null groups better than pooled
Separate modestly better

Separate and pooled analyses

Separate	Avg nulls successful	Avg alts successful	%alt continuing	E[N]
All 10	0.556 / 0.085	NA	0.000 / 0.000	150 / 88.8
Two 35	0.444 / 0.402	1.655 / 1.122	0.788 / 0.736	150 / 125.8
Half 35	0.278 / 0.879	4.137 / 3.863	0.937 / 0.815	150 / 136.7
Eight 35	0.111 / 0.972	6.618 / 7.126	0.983 / 0.880	150 / 112.4
All 35	NA	8.273 / 9.842	1.000 / 1.000	150 / 73.6

Pooled	Avg nulls successful	Avg alts successful	Avg total success (% alt)
All 10	0.440 / 10	NA	0.440 (0%)
Two 35	4.683 / 8	1.171 / 2	5.853 (20%)
Half 35	4.977 / 5	4.977 / 5	9.954 (50%)
Eight 35	2.000 / 2	8.000 / 8	10.0 (80%)
All 35	NA	10.0 / 10	10.0 (100%)

For half 35

similar to two 35, separate and adaptive screen out null groups

separate modestly better

Separate and pooled analyses

Separate	Avg nulls successful	Avg alts successful	%alt continuing	E[N]
All 10	0.556 / 0.085	NA	0.000 / 0.000	150 / 88.8
Two 35	0.444 / 0.402	1.655 / 1.122	0.788 / 0.736	150 / 125.8
Half 35	0.278 / 0.879	4.137 / 3.863	0.937 / 0.815	150 / 136.7
Eight 35	0.111 / 0.972	6.618 / 7.126	0.983 / 0.880	150 / 112.4
All 35	NA	8.273 / 9.842	1.000 / 1.000	150 / 73.6

Pooled	Avg nulls successful	Avg alts successful	Avg total success (% alt)
All 10	0.440 / 10	NA	0.440 (0%)
Two 35	4.683 / 8	1.171 / 2	5.853 (20%)
Half 35	4.977 / 5	4.977 / 5	9.954 (50%)
Eight 35	2.000 / 2	8.000 / 8	10.0 (80%)
All 35	NA	10.0 / 10	10.0 (100%)

For eight 35

Adaptive avg sample size 112.4

Adaptive has higher power and type 1 error compared to separate

Pooled might be viable depending on utilities

Separate and pooled analyses

Separate	Avg nulls successful	Avg alts successful	%alt continuing	E[N]
All 10	0.556 / 0.085	NA	0.000 / 0.000	150 / 88.8
Two 35	0.444 / 0.402	1.655 / 1.122	0.788 / 0.736	150 / 125.8
Half 35	0.278 / 0.879	4.137 / 3.863	0.937 / 0.815	150 / 136.7
Eight 35	0.111 / 0.972	6.618 / 7.126	0.983 / 0.880	150 / 112.4
All 35	NA	8.273 / 9.842	1.000 / 1.000	150 / 73.6

Pooled	Avg nulls successful	Avg alts successful	Avg total success (% alt)
All 10	0.440 / 10	NA	0.440 (0%)
Two 35	4.683 / 8	1.171 / 2	5.853 (20%)
Half 35	4.977 / 5	4.977 / 5	9.954 (50%)
Eight 35	2.000 / 2	8.000 / 8	10.0 (80%)
All 35	NA	10.0 / 10	10.0 (100%)

For all 35

Adaptive avg N is 73.6!

Adaptive power nearly matches pooled.

Summary for example

Scenario	Pooled	Separate	Adaptive	Overall
All 10	Good performance	BAD! Multiplicities greatly increase change of mistakenly going to phase 3 (44% chance compared to 5% for others)	Good performance overall, great performance per group. Reduced N	Adaptive clear winner
Two 35	Tend to run phase 3 with very diluted effect	Good performance	Screens closer to separate, but worse than separate	Separate likely winner? Adaptive better than pooled
Half 35	Phase 3 always run with diluted effect	Good performance	Screens modestly worse than separate	Separate winner? Adaptive closer to separate, but not quite there
Eight 35	Might be good?	Good performance	More power than separate, but more type 1 errors	???? quite subjective
All 35	Great! 100% power	Power only 83.2% per group	Great! High power, expected N reduced	Adaptive clear winner

What to pick?

- Adaptive is the clear winner for all null, all alternative
  - Separate is particularly bad for these scenarios
- Separate is the likely winner for two 35, five 35
  - Adaptive would likely be considered worse, but closer to ballpark of separate than pooled.
  - Pooled does particularly bad in these scenarios, bringing forward lots of null groups
- Eight 35 is a tough call, depends on sponsor utilities
- Summary – adaptive represents a reasonable compromise, clearly winning some scenarios, obtaining much of the advantage of separate in mixed scenario, can avoid mistakenly going to phase 3

Type 1 error and Regulatory

- The regulatory pathway for basket trials remains unclear
- Differs between divisions
  - Best established in oncology

- A key issue is type 1 error control
  - Do we require type 1 error control in the joint null, OR
  - Do we required familywise type 1 error

- Familywise type 1 error control is not practically possible without separate analyses, including alpha sharing
  - Often better to never acknowledge patient heterogeneity, and perform post hoc subgroup analysis (this is not ideal!)

Thank you

- Thank You for attending
- Link to Recording will be sent out tomorrow
- Slides will be available via our website at the end of the series
- Any questions please contact us:
  - tom@berryconsultants.com
  - kert@berryconsultants.com
  - facts@berryconsultants.com
  - demo and/or a free evaluation copy of FACTS
- Berry regularly produces blogs and social media posts on adaptive designs
  - @KertViele, Kert Viele on LinkedIn