Multi-arm multi-stage designs (MAMS)

CEN2023 pre-conference course on ”Advanced group-sequential and adaptive confirmatory clinical trial designs, with R practicals using rpact”

Marcel Wolbers & Kaspar Rufibach
Statistical Methods, Collaboration & Outreach (MCO)
Data & Statistical Sciences Department, Roche Pharma Development, Basel
Sunday, September 3, 2023

Good outcome for this session:

1) Not all MAMS are created equal.

2) Understand the MAMS landscape.

3) Understand the theoretical basis of pre-defined and flexible adaptive MAMS.

4) Awareness of available R software and rpact functionality.

General considerations for confirmatory multi-arm trials

Multi-arm trials

Comparison of $G > 1$ experimental treatment arms versus a **shared control arm**:

- Different molecules or combination therapies in same indication.
- Multiple doses of same molecule.

Features:

- Lower probability of being randomized to control: popular with patients.
- Efficiency gains.
- Shared trial infrastructure.
- Allows for randomized comparison between intervention arms.
- Treatment arm selection at interim analyses.
- With master protocols, treatment arms may also be added.
- Combine development phases in seamless designs.
  - Caution: Planning a phase III trial without phase II data is risky!

Pair-wise (PWER) or family-wise error rate (FWER) control?

PWER: Probability that a specific true null hypothesis H_0^g is falsely rejected.
FWER: Probability that at least one of (up to G) true null hypotheses is falsely rejected.

FWER of unadjusted comparisons to control in a multi-arm trial vs G independent two-arm trials:

- Positive correlation between test statistics in multi-arm trial due to shared control.
- This correlation **reduces** FWER!

FWER adjustment:

- **Not recommended**: solely due to shared control. Example: **Several drugs with different mechanisms of action**.
- **Recommended**: if there is increased chance of making **single claim of effectiveness** by testing multiple hypotheses. Example: **Several doses of same drug**.

For more details: Howard et al. (2018).

How to control the FWER? ⇒ Apply closed testing!

Set-up: For a set of G null hypotheses, define all associated **intersection hypotheses** and corresponding tests with significance level α.

Example: Closed testing for a 4-arm trial with **3 comparisons versus control**.

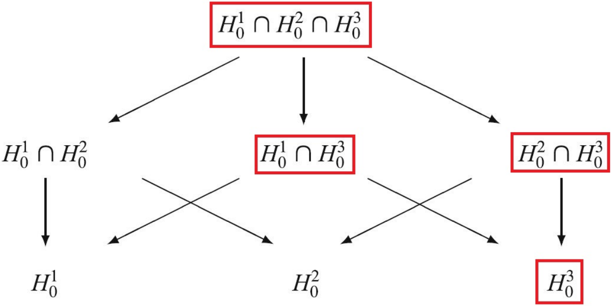
- Elementary null hypotheses: $H_0^g : \mu_g = \mu_C$ $(g = 1, \ldots, 3)$.
- Pair-wise intersection hypotheses: $H_0^{12} = H_0^1 \cap H_0^2$: $\mu_1 = \mu_2 = \mu_C$, H_0^{13}, H_0^{23}.
- Overall rejection hypothesis: $H_0^{123} = H_0^1 \cap H_0^2 \cap H_0^3$: $\mu_1 = \mu_2 = \mu_3 = \mu_C$.

Closed testing principle: An elemental null hypothesis H_0^g can be rejected while maintaining strong control of the FWER at level α if one can reject H_0^g plus all intersection hypothesis that imply it, each at level α (Marcus et al. (1976)).

Example: In order to reject H_0^3 at the one-sided family-wise 2.5% level, one needs to reject H_0^3 as well as H_0^{13}, H_0^{23}, and H_0^{123} at the 2.5% level.

Note: More intersection hypotheses would need to be tested if one wanted to control the FWER across **all pair-wise comparisons** (i.e. not only the G comparisons versus control). Exception: $G = 2$ (see Asikanius et al. (2016) for an example).

Illustration of closed testing



$H_0^1 \cap H_0^2 \cap H_0^3$

$H_0^1 \cap H_0^2$ $H_0^1 \cap H_0^3$ $H_0^2 \cap H_0^3$

H_0^1 H_0^2 H_0^3

Wassmer and Brannath (2016).

How to test intersection hypotheses?

Null hypotheses $H_0^g : \mu_g = \mu_C$ $(g = 1, \ldots, G)$; observed Z-scores z_g and p-values p_g.

Test for intersection hypothesis $H_0^{\mathcal{I}} = \cap_{g \in \mathcal{I}} H_0^g$ for $\mathcal{I} \subseteq \{1, \ldots, G\}$:

- **Dunnett test**: Let $z_{max} = \max\{z_g : g \in \mathcal{I}\}$. Then $p_{\mathcal{I}}^{adj} = 1 - \Phi(z_{max}, \ldots, z_{max})$ where Φ is the Dunnett distribution, i.e. the joint multivariate t- (or approximate normal) distribution of the Z-statistics under $H_0^{\mathcal{I}}$ (with known positive correlation due to the shared control group).
- **Bonferroni test**: $p_{\mathcal{I}}^{adj} = |\mathcal{I}| \cdot \min_{g \in \mathcal{I}}\{p_g\}$.
- **Simes test**: Let $p_{[1]} \leq \ldots \leq p_{[|\mathcal{I}|]}$ be the ordered p-values p_g $(g \subset \mathcal{I})$. Then $p_{\mathcal{I}}^{adj} = \min\{|\mathcal{I}| \cdot p_{[1]}, \frac{|\mathcal{I}|}{2} \cdot p_{[2]}, \frac{|\mathcal{I}|}{3} \cdot p_{[3]}, \ldots, p_{[|\mathcal{I}|]}\}$.
- **A priori hierarchical test**: $p_{\mathcal{I}}^{adj} = p_{\max\{g \in \mathcal{I}\}}$ where $\max\{g \in \mathcal{I}\}$ refers to the hypothesis of highest importance.

More details: Wassmer and Brannath (2016), Section 11.1.2.

Optimal randomization ratio

If comparing multiple treatments to control **but not to each other** (in superiority trial) $\Rightarrow$ equal randomization inefficient.

Dunnett (1955), Wassmer (2011), Wason and Jaki (2012): each of G treatment groups gets $1/\sqrt{G}\times$ sample size of control.

Chandereng et al. (2020): Shows that the above randomization ratio minimizes $\sum_{g=1}^{G}\mathrm{Var}(\bar{X}_g-\bar{X}_c)$ for normal endpoints with known variance.

Application:

- $G=2$: 1.41 : 1 : 1.
- $G=3$: 1.73 : 1 : 1 : 1.

Caveat: The optimal allocation ratio is likely closer to equal randomization if treatments can be dropped at interim analyses. Wason and Jaki (2012)

Examples

“MAMS” used very broadly.

RECOVERY:

- Landmark UK COVID-19 trial: `https://www.recoverytrial.net`, link to SAP.
- Design:
  - Platform trial of pairwise RCTs.
  - **No type 1 error correction** ⇒ shared control “only”.
- Status (as of 07August2023):
  - 48’569 participants from 190 active sites.
  - Results for 12 interventions so far, 4 of them with proven efficacy.
  - 5 interventions currently tested in the ongoing trial (2 for COVID-19, 3 for influenza).

Examples - continued

STAMPEDE:

- Since 2005 in UK, high-risk prostate cancer, `http://www.stampedetrial.org`.
- Initial design: 5 treatment groups vs control, randomized 1:1:1:1:1:2.
- 4 stages with pairwise comparisons to control
  - 3 futility interims to drop groups based on failure-free survival (FFS).
  - Final efficacy analysis based on primary outcome overall survival (OS).
- Pair-wise comparisons to control at unadjusted one-sided $\alpha = 0.025$. $\Rightarrow$ Maximum FWER of 0.103. Bratton et al. (2016).
- Power of pair-wise comparisons 90% ($\approx$ 83% after accounting for futility interims).

Stage	Target HR	Outcome	Continuation prob.: HR=1	Continuation prob.: HR=0.75	Required control group events
1	0.75	FFS	0.500	0.95	113
2	0.75	FFS	0.250	0.95	216
3	0.75	FFS	0.100	0.95	334
4	0.75	OS	Sig. level: 0.025	Power: 0.90	403

Pre-planned MAMS designs with FWER control / cumulative MAMS

Pre-planned MAMS

Pre-planned MAMS:

- **Extend group-sequential designs** to "multiple groups to control" comparison.
- Interim analyses:
  - **Futility**: select promising treatment(s) to be compared with control in subsequent stages ⇒ drop ineffective groups. Some publications suggest **binding** rules for dropping arms (see later).
  - **Efficacy**: potential to demonstrate superiority of a treatment group over control early. The trial may then stop altogether or continue with the remaining arms. (Note that the global intersection null hypothesis is rejected at this stage, i.e. it does not need to be re-tested for the remaining comparisons.)

Once trial started ⇒ type I error protection only guaranteed if efficacy and binding futility interim decisions follow pre-specified criteria.

Follmann et al. (1994), Wason and Jaki (2012), Magirr et al. (2012), Magirr et al. (2014), Jaki et al. (2019), Ghosh et al. (2017), Ghosh et al. (2020), many more.

Setup (template case)

Normally distributed outcomes with known variance.

G groups vs common control.
$H_0^g : \mu_g - \mu_C \leq 0$ $(g = 1, \ldots, G)$ vs $H_A^g : \mu_g - \mu_C > 0$ $(g = 1, \ldots, G)$.

J stages.
At interim j, compute standardized test statistics Z_j^g of group g vs control based on the **cumulative data** from stage 1 until stage j.

The Z-scores Z_j^g $(g = 1, \ldots, G, j = 1, \ldots, J)$ follow a **multivariate normal distribution** with known correlation matrix (Anderson et al. (2022)).

Group-sequential case with efficacy interim analyses only

Denote the maximum Z-score among all comparisons at stage j by $Z_j^{max} = \max_{g \in \{1,\ldots,G\}}\{Z_j^g\}$.

If one wants to spend α_j of the total type I error at stage j with $\sum_{j=1}^J \alpha_j = \alpha$, then associated **efficacy boundaries** b_j for the Z-scores can be calculated via the equations:

$$P_0(Z_1^{max} > b_1) = \alpha_1 \text{ and } P_0(\cap_{l=1}^{j-1}\{Z_l^{max} \leq b_l\} \cap \{Z_j^{max} > b_j\}) = \alpha_j \text{ (j>1).}$$

The critical values b_j are calculated under the global null hypothesis. However, it can be demonstrated that for these b_j, the probability to reject any true null hypothesis is $\leq \alpha$ regardless which null hypotheses are true, i.e. that the test **strongly controls the FWER** (Theorem 1 in Magirr et al. (2012)).

Calculations of multivariate normal probabilities are **computationally intensive**.

Massively reduced computation time: Ghosh et al. (2017). Implemented (binary, continuous) in East MAMS module.

Adding pre-planned binding treatment selection rules

Critical values depend on **selection rule** for **binding** rules (as per the cited articles below)!

Select the best:

- Treatment with largest test statistic only continues with control beyond first interim.
- Stallard and Todd (2003).

Keep all promising:

- Add **binding** futility boundaries for treatments to proceed from stage j to $j+1$.
- Magirr et al. (2012).

What happens if we do not follow binding selection rules?

Select the best:

- Select experimental treatment other than that with largest $Z_j^g \Rightarrow$ conservative.
- Select > 1 experimental treatment to go beyond 1st stage $\Rightarrow$ T1E not controlled.

Keep all promising:

- Dropping experimental treatment(s) although not formally futile $\Rightarrow$ conservative.
- Keep experimental treatment although declared futile $\Rightarrow$ T1E not controlled.

Rescue to maintain T1E control:

- Apply **Conditional Rejection Principle (CRP) and closed testing** after deviations from pre-planned selection rule (Magirr et al. (2014),Ghosh et al. (2020)).
- Note: If the variance is unknown, the conditional error rate is difficult to calculate and relies on additional assumptions (Wassmer and Brannath (2016), Section 11.1.5).

Power and sample size

With $G > 1$ treatments, definition of power **not obvious**.

- δ: effect that, if present, we would like to detect with high probability.
- δ_0: effect that, if present, would not be of interest. ($\delta_0 = 0$ implies that any effect would be worth detecting.)
- Dunnett (1984): **least favorable configuration**:

$$P(\text{reject } H_0^1 \text{ assuming } \mu_1 - \mu_0 = \delta \text{ and } \mu_g - \mu_0 = \delta_0, g = 2, \ldots, G).$$

- Minimizes

$$P(\text{reject } H_0^1 \text{ over all choices of } \mu_1, \ldots, \mu_G \text{ s.t. } \mu_1 - \mu_0 \geq \delta \\ \text{and } \mu_g - \mu_0 \leq \delta_0, g = 2, \ldots, G).$$

Expected sample size: mean number of patients recruited before trial stops.

Analytical expressions: Magirr et al. (2012). Does not mean closed form - integrals!

Example: Boundaries and sample size using R package MAMS

```
> library(MAMS)
> # Two interventions (K = 2) vs control, 2 stages (J = 2) with equal sample size per group
> # Allocation ratios:
> # r0 refers to relative cumulative allocation across stages in control; r refers to treatment
> # O'Brien-Fleming boundary shape for efficacy and a binding futility boundary at Z = 0
>
> r0 <- c(1, 2)
> mams22 <- mams(K = 2, J = 2, alpha = 0.025, power = 0.8, r = r0, r0 = r0,
+               ushape = "obf", lshape = "fixed", lfix = 0,
+               delta = 10, delta0 = 4, sd = 24, p = NULL, p0 = NULL)
> mams22

Design parameters for a 2 stage trial with 2 treatments


                                            Stage 1 Stage 2
Cumulative sample size per stage (control):      57     114
Cumulative sample size per stage (active):       57     114


Maximum total sample size:  342


             Stage 1 Stage 2
Upper bound:   3.139    2.22
Lower bound:   0.000    2.22
```

Summary: Pre-planned MAMS

- Generalization of group-sequential designs.
- Rely on **joint distribution of cumulative test statistics**.
- Type I error protection:
    - Original design: Only if conduct compliant with pre-defined interim futility / efficacy boundaries.
    - Deviations from pre-defined rules: Rescue with Conditional Rejection Principle (CRP) and closed testing (Magirr et al. (2014),Ghosh et al. (2020)) .
- Design may be more efficient than adaptive designs using stage-wise *p*-value combination (Ghosh et al. (2020)) but application of CRP principle (required for full adaptivity) assumes known variances.
- Numerically challenging, but feasible (for reasonable number of stages).
- R package **MAMS**. Gives sample size, critical values, allows trial simulation.
- Time-to-event endpoints: timing needs more work, e.g. via **rpact**.

Flexible adaptive (stage-wise) MAMS

p-value combination across stages

$+$

closed testing

Setup (template case)

Normally distributed outcomes.

G groups vs common control.
$H_0^g : \mu_g - \mu_C \leq 0$ $(g = 1, \ldots, G)$ vs $H_A^g : \mu_g - \mu_C > 0$ $(g = 1, \ldots, G)$.

J stages (i.e. $J - 1$ interim analyses plus final analysis).

After each stage j, analyse data and based on these data make a decision:

- **Stop for efficacy** of one or multiple treatment groups.
- **Stop for futility** for all treatment groups.
- **Proceed to stage $j + 1$** but may drop treatment groups for futility or re-assess sample size.

Methodology to control the FWER

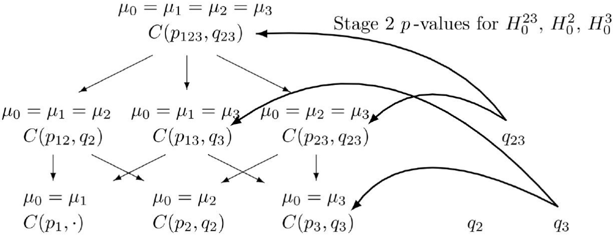
After each stage j, **calculate p-values for the elementary null hypotheses H_0^g and all intersection null hypotheses $H_0^{\mathcal{I}} = \cap_{g \in \mathcal{I}} H_0^g$** for $\mathcal{I} \subset \{1, \ldots, G\}$ **based on data from stage j only** (i.e. not cumulative data).

- If treatment groups have been dropped prior to stage j, then the p-value for testing $H_0^{\mathcal{I}}$ is obtained by testing $H_0^{\mathcal{I} \setminus \mathcal{E}}$ where $\mathcal{E}$ denotes the set of excluded groups.

To make an interim test decision after stage j, **combine each of the stage-wise p-values across stages** $1, \ldots, j$ using a combination test.

Reject H_0^g after stage j **if all combination p-values** for H_0^g and for all intersection hypotheses $H_0^{\mathcal{I}}$ with $g \in \mathcal{I}$ **are below the local significance level** of the combination test for stage j.

Illustration of p-value combination and closed testing



Combination tests to be performed for the closed system of hypotheses ($G = 3$) for testing hypothesis H_0^3 if treatment groups 2 and 3 are selected for the second stage.

Design choices for adaptive MAMS

Design choices (including planned adaptations) should be **pre-defined** in the protocol and SAP.

Number of stages J and **sample size** in the control and each (remaining) treatment group per stage.

- Typically chosen based on trial simulations.

***p*-value combination** test across stages.

- E.g. inverse normal combination test with pre-defined α-spending (for efficacy interims) and weights aligned with planned sample sizes.

Intersection test

- E.g. Dunnett test.
- Caution: Bonferroni tests may lead to intersection p-values of 1 which imply an implicit futility stop (because inverse normal combination tests cannot lead to rejection if one of the involved p-values is 1).

Design choices for adaptive MAMS - continued

Futility stopping rules for treatment groups.

- Can be based on conditional power.
- Alternatively, **rpact**'s simulation tool allows treatment selection options:
  - Select best or r best treatment groups (`"best", "rbest"`)
  - Select treatment groups not worse than ϵ compared to the best (`"epsilon"`).
  - User-defined (`"userDefined"`).

Sample size re-assessment rules (if any).

- Can be based on conditional power.
- Also specify minimum and maximum allowed sample size.

Design and analyses of MAMS using rpact

Key **functions**:

- Specify *p*-value combination test: `getDesignInverseNormal`.
- Trial simulation:`getSimulationMultiArm[Means,Rates,Survival]`.
- Trial analysis: `getDataset, getAnalysisResults`.

Useful **vignettes** (`https://www.rpact.com/vignettes`):

- Simulating Multi-Arm Designs with a Continuous Endpoint.
- Analysis of a Multi-Arm Design with a Binary Endpoint.

Example: Adaptive design simulation using rpact

```
> # 2 stages of equal size, 2 treatment groups vs control
> # Normal outcomes, true mean diff: 10 (group 1), 4 (group 2); stDev: 24
> # For this example, use 56 subjects per group and stage
> # (as per getSampleSizeMeans(alternative=10,stDev = 24,alpha=0.025/2,beta=0.2)$nFixed1/2)
> library(rpact)
> designIN <- getDesignInverseNormal(kMax = 2, alpha = 0.025, sided=1, typeOfDesign = "OF",
+                                    informationRates = c(0.5, 1))
> flex_adap_sim <- getSimulationMultiArmMeans(design = designIN,
+                                           activeArms = 2,
+                                           typeOfShape = "userDefined",
+                                           effectMatrix = matrix(c(10,4), nrow = 1),
+                                           stDev = 24,
+                                           plannedSubjects = c(56,112),
+                                           intersectionTest = "Dunnett",
+                                           typeOfSelection = "best",
+                                           successCriterion = "atLeastOne",
+                                           maxNumberOfIterations = 1e5,
+                                           seed = 1234)
```

- typeOfShape: Models dose-response relationship ⇒ `effectMatrix`.
- typeOfSelection: Defines how treatment arm(s) selected at interim.
- successCriterion: Criterion to stop trial for efficacy at interim: `all` or `best`.

Example: Adaptive design simulation using rpact

```
> summary(flex_adap_sim)

Simulation of a continuous endpoint (multi-arm design)

Sequential analysis with a maximum of 2 looks
(inverse normal combination test design), overall significance level 2.5%
(one-sided).
The results were simulated for a multi-arm comparisons for means
(2 treatments vs. control), H0: mu(i) - mu(control) = 0, H1: mu_max = 10,
standard deviation = 24, planned cumulative sample size = c(56, 112),
effect shape = user defined, intersection test = Dunnett, selection = best,
effect measure based on effect estimate, success criterion: at least one,
simulation runs = 100000, seed = 1234.

...
```

Example: Adaptive design simulation using rpact

```
...

Stage                                  1      2
Fixed weight                       0.707  0.707
Efficacy boundary (z-value scale)  2.797  1.977
Reject at least one               0.8008
Rejected arms per stage
  Treatment arm 1                 0.2143 0.5549
  Treatment arm 2                 0.0236 0.0278
Success per stage                 0.2181 0.5827
Expected number of subjects        255.6
Overall exit probability          0.2181
Stagewise number of subjects
  Treatment arm 1                   56.0   49.9
  Treatment arm 2                   56.0    6.1
  Control arm                       56.0   56.0
Selected arms
  Treatment arm 1                 1.0000 0.6967
  Treatment arm 2                 1.0000 0.0852
Number of active arms              2.000  1.000
Conditional power (achieved)             0.3888

Legend:
  (i): treatment arm i
```

Prespecified vs. flexible adaptive MAMS

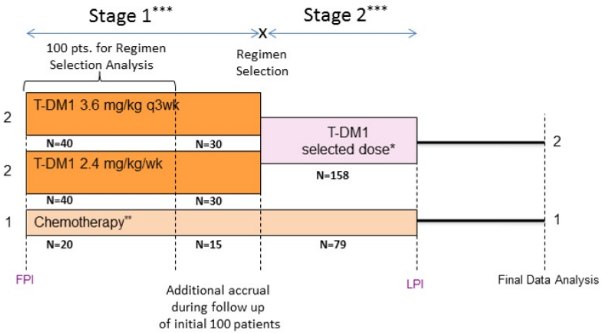
	pre-specified	flexible adaptive
Conceptually	joint distribution of cumulative test statistics	combine stagewise *p*-values
Control arm	Shared control arm	
Attractiveness	P(randomized to control) low ⇒ popular with patients	
Operational	More aligned than separate trials, shared infrastructure	
FWER control	Control FWER across all comparison, as opposed to separate trials	
Flexibility	Once trial started must be conducted as specified. Adaptive extension: Magirr et al. (2014), Ghosh et al. (2020).	Design changes (drop arm, change population, sample size re-estimation, ...) can be made at interim without pre-specification, while maintaining FWER.
R implementation	**MAMS**. Basic functionality (sample size, power, simulation) only. Only simulates test statistics (not patients) for T2E. No seed can be set for simulations.	**rpact**: Flexible simulation and analysis functions. Only simulates test statistics for T2E. Allowing interim decisions based on surrogate endpoints planned. **asd**: sample size for enrichment and arm selection, including surrogacy. Specification for arm selection for T2E unclear.

Example: Gatsby trial

Gatsby: Adaptive dose-selection trial

Population:
2L HER2+ AGC
Prior chemotx±prior
HER2-directed therapy
PS 0-1
Total n=412
(Stages 1 + 2)

Stratified by:
region, prior
gastrectomy, and
prior HER2-targeted tx



- **Objectives:**
  - **1º: OS**
  - **2º: PFS, ORR, DOR, PRO, safety**
- **Randomization:**
  - **Stage 1: 3 arm; 2:2:1 ratio**
  - **Stage 2: 2 arm; 2:1 ratio**

Total Sample size n=412
- **Selected T-DM1 arm: 228**
- **Control arm: 114**
- **Non-selected T-DM1 arm: 70**

* *Regimen selection based on efficacy, safety, and PK data available at timepoint of regimen selection analysis.*

**** Investigator's choice between paclitaxel 80 mg/m²/wk and docetaxel 75 mg/m² q3wk.**

*****Stage 1 (Stage 2) patients consist of all patients recruited before (after) the dosing decision.**

Gatsby: Study design features

Patient-wise staging:

- Final analysis data from stage 1: After 83% of stage 1 patients (all 3 groups) have died.
- Final analysis data from stage 2: After 63% of stage 2 patients (selected + control group) have died.
- Notes:
  - Requires that regimen selection does not affect study procedures. Especially, OS follow-up needs to continue until final analysis for all 3 groups.
  - Final analysis cut-off date for stage 1 and stage 2 data may not perfectly align.
  - Guarantees independence of stage 1 and stage 2 p-values under the null.

p-value combination: Inverse normal combination test, weights equal to square root of relative event number from each stage.

Intersection test: Simes test.

Gatsby: Study design features (continued)

Treatment regimen selection:

- Performed by an **IDMC** based on interim data from stage 1 patients.
- Design and selection criteria based on **extensive clinical trial simulations** using multivariate normal models for the correlation between cycle 1 AUC, treatment-related mortality (TRM), and OS data.

Positive Health Authority feedback.

Efficiency gains over two separate trials:

- No white space between dose selection and Phase 3.
- **Re-use dose selection data** for confirmatory analysis!

Gatsby was **negative**, because drug did not work sufficiently. Thuss-Patience et al. (2017)

Relevant references: Jennison (2023) (reflection talk on GATSBY), Magirr et al. (2016) (alternative stagings and approaches for adaptive survival trials), Jenkins et al. (2011), Carreras et al. (2015) (interim decisions based on surrogates).

Final comments

Final comments

Think of "MAMS" as of "platform": no clear definition, rather focus on **specific designs and their statistical properties**.

Flexible adaptive multiarm designs may offer an **efficient** way to develop drugs:

- Theory well established.
- Regulators accept it - if well planned and run.
- We have standard R tools to plan them: **MAMS**, **rpact**, **asd** (though additional fine-tuning may be required).
- May involve more work than "standard" approaches. But: upfront investment may pay off in shorter and more efficient trials. **Do not focus on date of first patient in, but on date of filing**!

Thank you for your attention.

marcel.wolbers@roche.com
kaspar.rufibach@roche.com

References I

- Anderson, K. M., Guo, Z., Zhao, J. and Sun, L. Z. (2022). A unified framework for weighted parametric group sequential design. *Biometrical journal. https://doi.org/10.1002/bimj.202100085*.

- Asikanius, E., Rufibach, K., Bahlo, J., Bieska, G. and Burger, H. U. (2016). Comparison of design strategies for a three-arm clinical trial with time-to-event endpoint: Power, time-to-analysis, and operational aspects. *Biometrical journal* **58** 1295–1310.

- Bratton, D. J., Parmar, M. K. B., Phillips, P. P. J. and Choodari-Oskooei, B. (2016). Type i error rates of multi-arm multi-stage clinical trials: strong control and impact of intermediate outcomes. *Trials* **17** 309.

- Carreras, M., Gutjahr, G. and Brannath, W. (2015). Adaptive seamless designs with interim treatment selection: a case study in oncology. *Statistics in Medicine* **34** 1317–33.

- Chandereng, T., Wei, X. and Chappell, R. (2020). Imbalanced randomization in clinical trials. *Statistics in Medicine* **39** 2185–2196. `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8539`

- Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50** 1096–1121. `http://www.jstor.org/stable/2281208`

- Dunnett, C. W. (1984). Selection of the best treatment in comparison to a control with an application to a medical trial. *Design of experiments: Ranking and selection* **47** 66.

References II

- Follmann, D. A., Proschan, M. A. and Geller, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* **50** 325–336.
- Ghosh, P., Liu, L., Senchaudhuri, P., Gao, P. and Mehta, C. (2017). Design and monitoring of multi-arm multi-stage clinical trials. *Biometrics* **73** 1289–1299.
- Ghosh, P., Liu, L. and Mehta, C. (2020). Adaptive multiarm multistage clinical trials. *Statistics in Medicine* **39** 1084–1102.
- Howard, D. R., Brown, J. M., Todd, S. and Gregory, W. M. (2018). Recommendations on multiple testing adjustment in multi-arm trials with a shared control group. *Statistical methods in medical research* **27** 1513–1530.
- Jaki, T., Pallmann, P. and Magirr, D. (2019). The r package mams for designing multi-arm multi-stage clinical trials. *Journal of Statistical Software, Articles* **88** 1–25. `https://www.jstatsoft.org/v088/i04`
- Jenkins, M., Stone, A. and Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* **10** 347–356.
- Jennison, C. (2023). Designing an adaptive trial with treatment selection and a survival endpoint: reflections on the GATSBY study. *Talk at Roche, Welwyn Garden City, June 2023* `https://people.bath.ac.uk/mascj/talks_2023/roche-june-2023.pdf`.

References III

- Magirr, D., Jaki, T., Koenig, F. and Posch, M. (2016). Sample size reassessment and hypothesis testing in adaptive survival trials. *PloS One* **11** e0146465.
- Magirr, D., Jaki, T. and Whitehead, J. (2012). A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* **99** 494–501.
- Magirr, D., Stallard, N. and Jaki, T. (2014). Flexible sequential designs for multi-arm clinical trials. *Statistics in Medicine* **33** 3269–3279.
- Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660.
- Stallard, N. and Todd, S. (2003). Sequential designs for phase iii clinical trials incorporating treatment selection. *Statistics in Medicine* **22** 689–703.
- Thuss-Patience, P. C., Shah, M. A., Ohtsu, A.,..., and Kang, Y. K. (2017). Trastuzumab emtansine versus taxane use for previously treated HER2-positive locally advanced or metastatic gastric or gastro-oesophageal junction adenocarcinoma (GATSBY): an international randomised, open-label, adaptive, phase 2/3 study. *The Lancet Oncology* **18.5** 640-653.
- Wason, J. M. S. and Jaki, T. (2012). Optimal design of multi-arm multi-stage trials. *Statistics in Medicine* **31** 4269–4279.
- Wassmer, G. (2011). On sample size determination in multi-armed confirmatory adaptive designs. *Journal of Biopharmaceutical Statistics* **21** 802–817.

References IV

- Wassmer, G. and Brannath, W. (2016). Adaptive group sequential tests. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials* . `http://dx.doi.org/10.1007/978-3-319-32562-0_6`