Constructing a comprehensive simulation study

Pavel Mozgunov

MRC Biostatistics Unit, University of Cambridge, UK,

22 October 2025

Selection of simulation scenarios

In this course, we have looked into how to conduct simulations studies under several scenarios and how to analyse and interpret these results.

The good news is that with the current computational capacities, we are not really restricted to the number of settings that we can explore - we can explore many!

However, how do we know that we have explore "enough" simulation settings?

Good practice

- Define your "working" simulation scenario - the best guess assumptions (from previous trials or RWE) for each of the parameters of the trial
- Justify them!
- Define several values around the "working" assumptions
- Consider all possible combinations of these (not one at a time)
- If computationally expensive, be smarter!
  - fix the things that YOU choose in the real trial
  - consider more "extreme" cases to narrow down which parameters should be scrutinised more.
  - Start with fewer simulations but in all (plausible) settings
- Define "**edge cases**" regardless of how implausible they might be

Example: MAMS

Things to explore that **we** choose:

- Shapes of the decision boundaries
- Timing of the interim analyses
- Number of interim analyses
- Different adaptations (futility only? with efficacy? combination of thereof)

Things to explore that **we cannot control**:

- Standard deviation of the outcome (on different arms)
- Treatment effect on each arm (with 1, 2, ... arms having promising, uninteresting and in-between treatment effect)
- Distribution of the outcome
- Missing data (start with MCAR)
- Different recruitment rates

Principle of a comprehensive simulation study

- Include more most plausible but also less plausible scenarios to explore the deviations from the assumptions;
- Include the "edge cases"
- Study the literature on the sub-category of methods to identify such problematic cases
- Explore the parameter space extensively
- Explore both the assumed and misspecified data generation algorithm
- Always include a **reference method** or methods (!)

Reference methods/Benchmarks

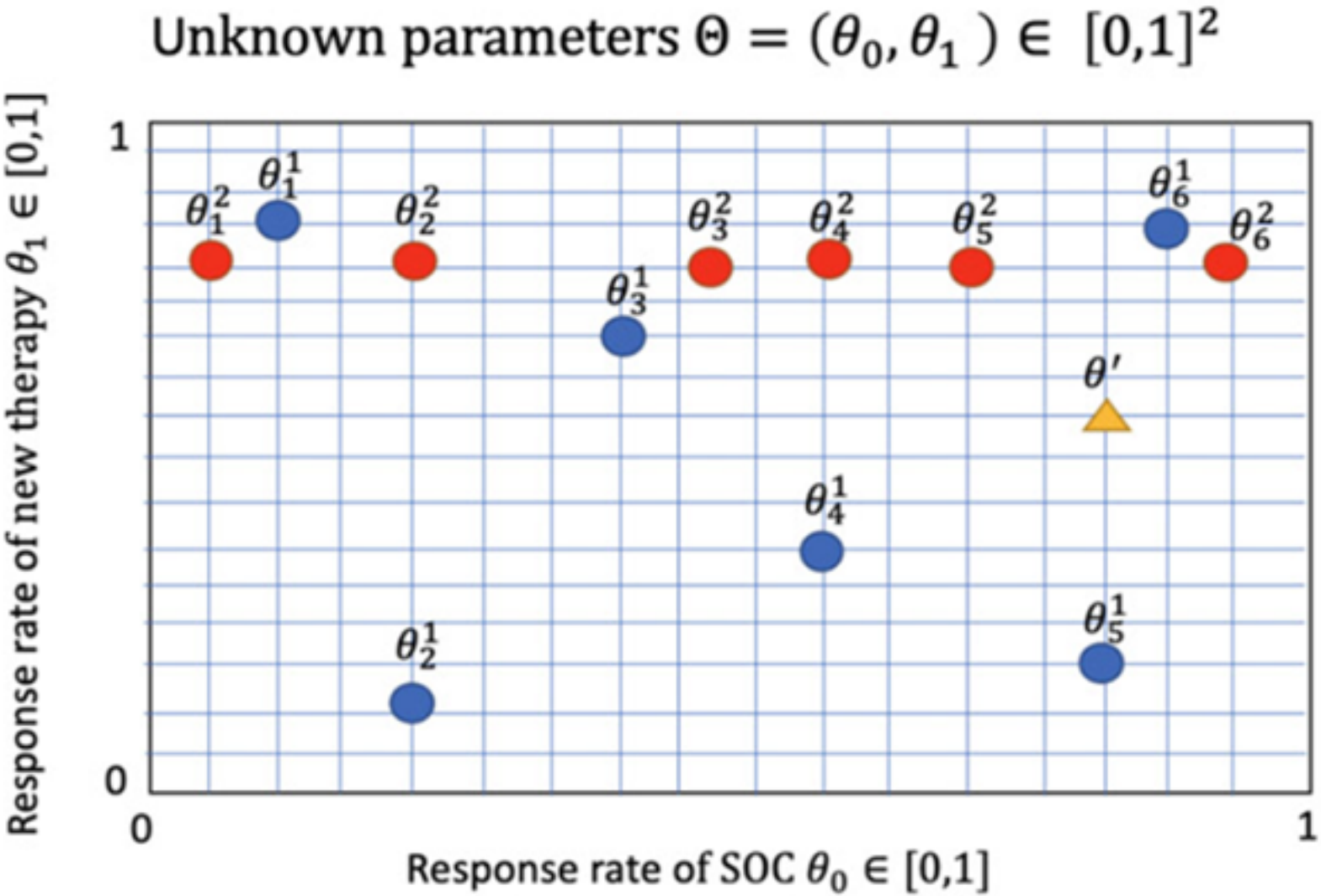
Reference methods (also referred to as benchmarks) helps us attribute whether the observed performance is due to the

- selection of simulation scenarios
- features of the design/model
- specific choices within this design/model

There are generally two approaches to selected a benchmark

- **Theoretical** (but not achievable in practice) methods
  - Theoretically optimal allocation of patients given the true model
  - Methods based on the "complete information" (i.e. one knows how each patient would respond to each individual treatment)
- **Simple and well-understood** (practical) methods
  - Simple non-adaptive options
  - Correct model specification
  - ...

Exploration of the parameter space



Principle of a comprehensive simulation study

It is not always about the parameters directly (!) but it is always about the **objective of the trial**!

For example, in a MAMS trial, if this is about selecting the best treatment then one should explore:

- Different number of promising arms
- Different treatment effects for uninteresting arms

There is (yet) **no unified framework** to construct a simulation study regardless of the research question.

We are going to provide three further examples of the structured simulation studies and emphasise the **overarching approach** of constructing the simulation study with respect to the research objective:

- Response-adaptive randomisation
- Dose-finding in combinations
- Basket trial designs

Response adaptive randomisation (RAR)

Idea: change the allocation probabilities during the conduct of the trial towards more promising methods

Methods: different "rules" for changing the allocation ratio

Expected challenges: underpowered (deviates from the optimal allocation), type I error inflation, "locking in" on one treatment arm

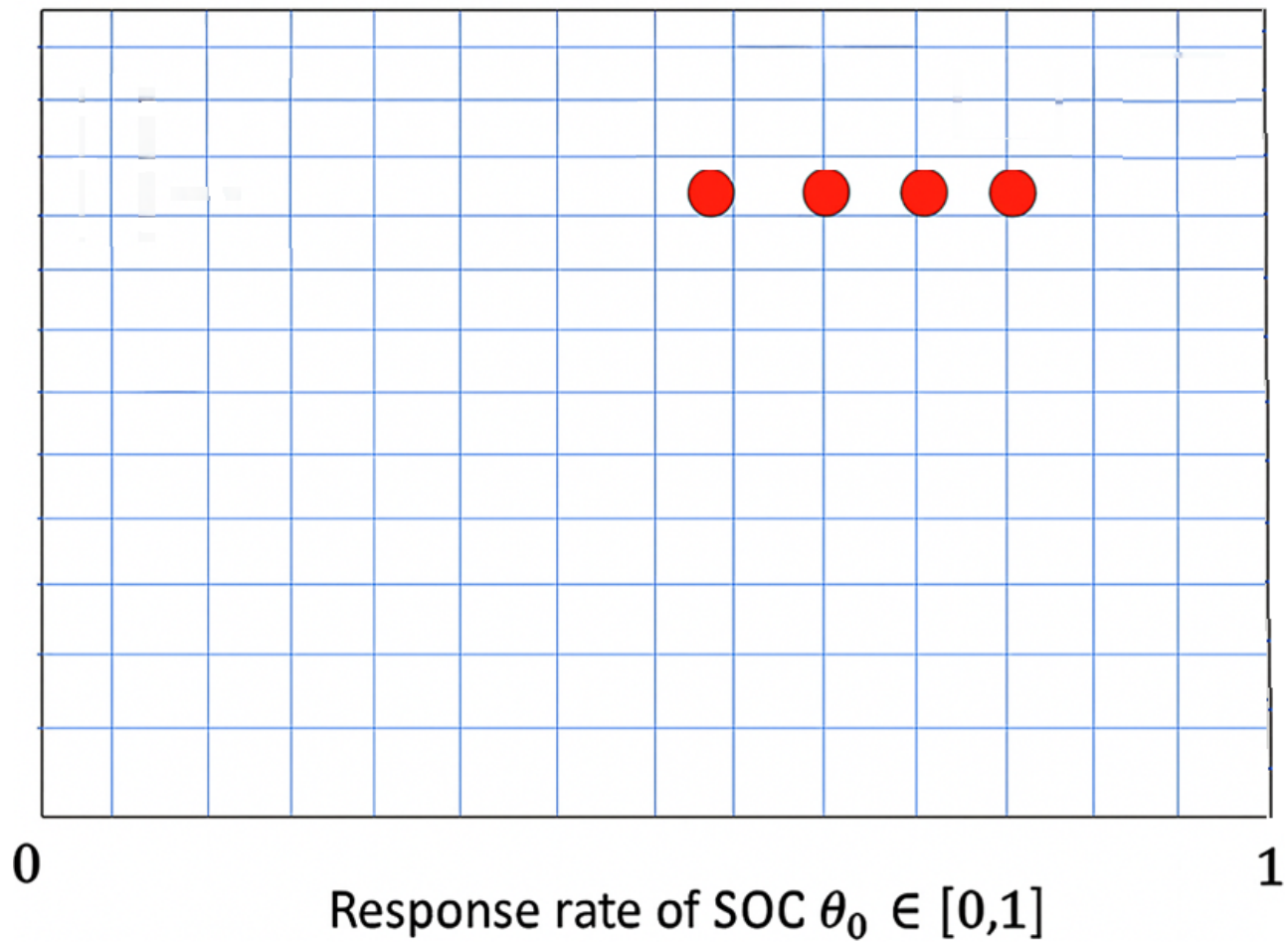
Key metrics of interest:

- Type I error
- Power
- Expected benefit of the patient
- Standard errors of the proportion of correct allocations

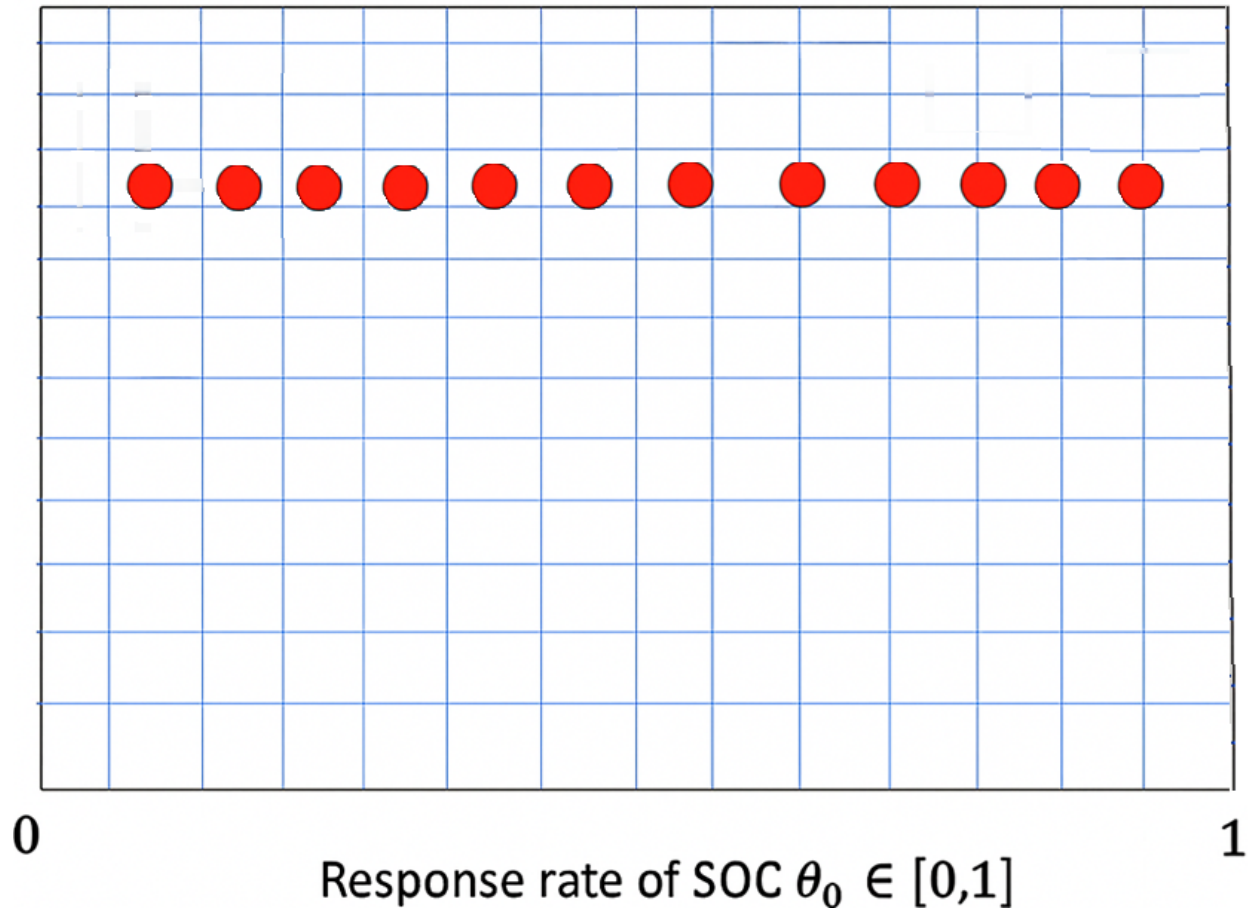
RAR: reference methods

- **Fixed Equal Allocation** – good (best) performance in terms of the power and type I error control is expected
- **Gittins Index** – the best possible performance in terms of the patient benefit
- "**Oracle" allocation** knowing the truth – all patients are allocated to the true best-performing arm

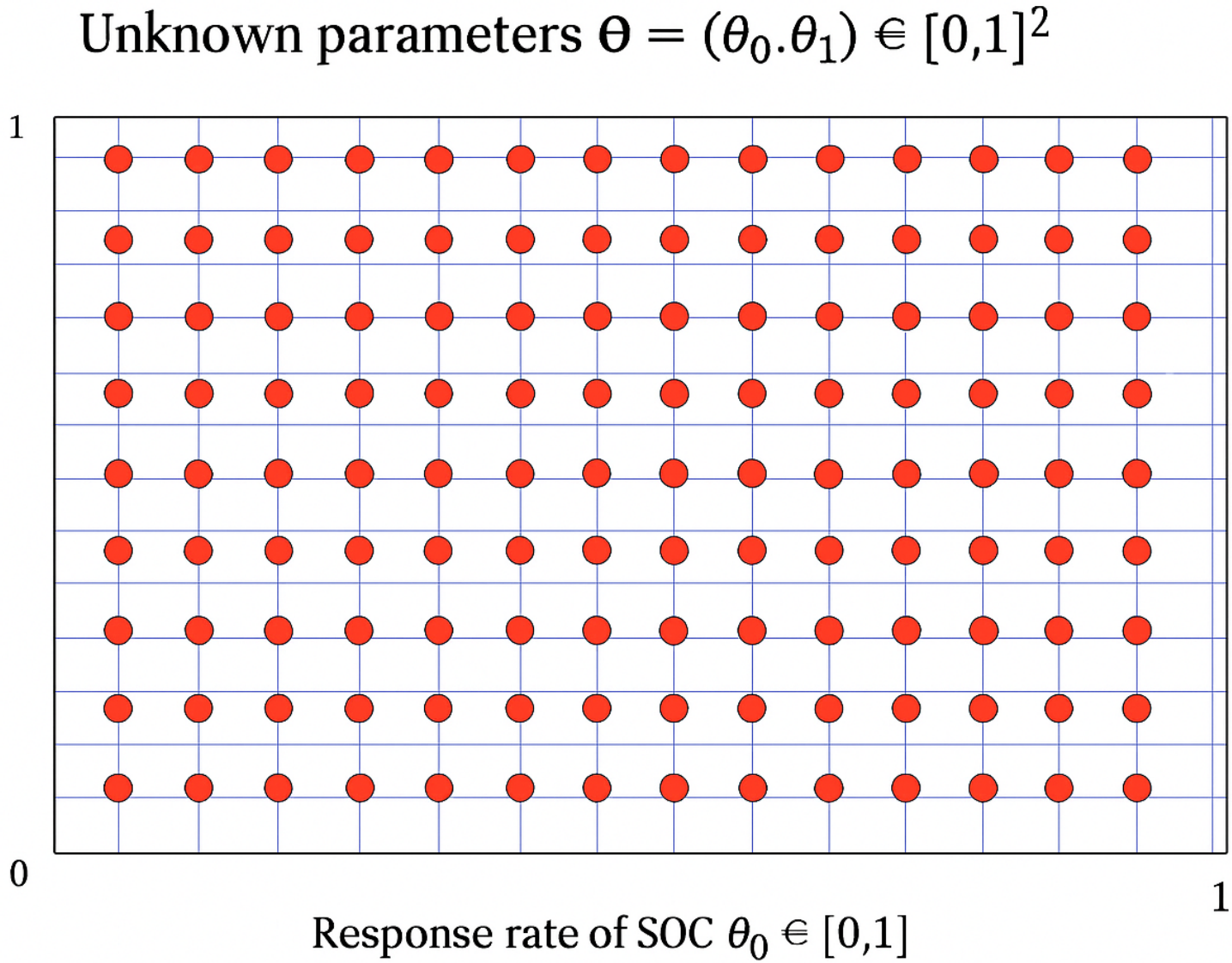
RAR: exploration of the parameter space



RAR: exploration of the parameter space



RAR: exploration of the parameter space



Basket trial designs

Idea: study the same treatment in different patient populations (but with the same marker expression) and borrowing "strength" across baskets

Methods: models based on the borrowing of information between treatment arm, e.g. Bayesian Hierarchical Model

Expected challenges: type I error inflation, biased estimates

Key metrics of interest:

- Type I error
- Power
- Bias

Basket designs: reference methods

- **Individual (stand-alone) analysis** – no type I error is expected, the lowest possible power
- **Bayesian Hierarchical Model** – high type I error inflation but a high (best?) possible power. Well studied in the literature now.

Basket designs: exploration of the parameter space

For the binary endpoint with an equal sample size for each basket:

	p_1	p_2	p_3	p_4	p_5
Scenario 1	0.15	0.15	0.15	0.15	0.15
Scenario 2	0.45	0.15	0.15	0.15	0.15
Scenario 3	0.45	0.45	0.45	0.45	0.15
Scenario 4	0.45	0.45	0.45	0.45	0.45

If the sample size is not equal, then all possible permutations of these for different number of the promising baskets

+ in-between scenarios

Dose-finding designs

Idea: escalate in small cohorts of patients to find the maximum tolerated dose (MTD) and optimal biologic dose (OBD)

Methods: designs that "prescribe" how the dose levels should be explored based on accumulate data (CRM, BLRM, BOIN)

Expected challenges: small sample sizes, complex models in combination setting

Key metrics of interest:

- Proportion of the correct dose selection
- Proportion of the overly toxic selection
- Proportion of patients receiving toxic doses

Dose-finding designs: reference methods

- **3+3 design** – the most commonly used in practice but easy to beat
- **Non-parametric optimal benchmark** (based on the complete information) – the best optimal performance under given scenario (without a strong prior)
- **Continual Reassessment Method** (CRM) – known for the high accuracy

Dose-finding designs: exploration of the parameter space

Scenarios with 3 possible doses of two drugs. Target toxicity rate = 30%.

Scenario 1				Scenario 2				Scenario 3				Scenario 4		
0.50	0.60	0.70		0.50	0.60	0.70		0.45	0.50	0.60		0.45	0.60	0.70
0.45	0.50	0.60		0.45	0.50	0.60		0.20	0.45	0.50		**0.30**	0.50	0.60
0.30	0.45	0.50		0.20	**0.30**	0.45		0.10	0.20	**0.30**		0.20	0.45	0.50
Scenario 5				Scenario 6				Scenario 7				Scenario 8		
0.45	0.50	0.60		0.20	0.45	0.50		**0.30**	0.50	0.60		0.20	**0.30**	0.50
0.20	**0.30**	0.50		0.10	0.20	**0.30**		0.20	0.45	0.50		0.10	0.20	0.45
0.10	0.20	0.45		0.00	0.10	0.20		0.10	0.20	0.45		0.00	0.10	0.20
Scenario 9				Scenario 10				Scenario 11				Scenario 12		
0.10	0.20	**0.30**		0.45	0.50	0.60		**0.30**	0.50	0.60		0.45	0.50	0.60
0.00	0.10	0.20		**0.30**	0.45	0.50		0.20	0.45	0.50		**0.30**	0.45	0.50
0.00	0.00	0.10		0.20	**0.30**	0.45		0.10	**0.30**	0.45		0.10	0.20	**0.30**
Scenario 13				Scenario 14				Scenario 15				Scenario 16		
0.30	0.50	0.60		0.45	0.50	0.60		0.20	**0.30**	0.50		**0.30**	0.45	0.60
0.20	0.45	0.50		0.20	**0.30**	0.45		0.10	0.20	0.45		0.20	**0.30**	0.50
0.10	0.20	**0.30**		0.10	0.20	**0.30**		0.00	0.10	**0.30**		0.10	0.20	0.45
Scenario 17				Scenario 18				Scenario 19						
0.30	0.45	0.50		0.20	**0.30**	0.45		**0.30**	0.45	0.50				
0.10	0.20	**0.30**		0.10	0.20	**0.30**		0.20	**0.30**	0.45				
0.00	0.10	0.20		0.00	0.10	0.20		0.10	0.20	**0.30**				

One more trick...

Although the above examples have a **clear structure** and **justifiable approach** on how the simulation scenarios are constructed, they still can result in missing some challenging scenarios unintentionally in which the method will struggle to answer the research question.

A useful tool to mitigate this is to **randomly generate scenarios** (!) and then generate the data under this scenario.

There are two main options:

- Under each simulation scenario **generate one** simulated trials – allows to estimate the performance on average across scenarios;
- Under each simulation scenario **generate multiple** simulated trials – allows to identify the "edge" cases for the design.

Conclusions

- There is currently no prescriptive guidance on how to choose simulations scenarios for the majority of classes of clinical trial designs - nor from regulators nor in scientific community;
- Exploring the plausible scenarios is only one dimension as these are subjectively chosen;
- It is recommended to work with experts in the specific category of methods that you are exploring to apply;
- There is a growing recognition of doing simulation studies well and more and more good practices are shared.
- Start small, make mistakes, learn, repeat