A practical introduction to simulating complex trial designs

Pavel Mozgunov Dominique-Laurent Couturier

MRC Biostatistics Unit, University of Cambridge, UK,

https://github.com/dose-finding/simulations-2025

22 October 2025

Acknowledgements

- Prof Thomas Jaki (MRC Biostatistics Unit)
- Dr Tom Burnett (University of Bath)

Introduction to simulation studies

Pavel Mozgunov

MRC Biostatistics Unit, University of Cambridge, UK,

https://github.com/dose-finding/simulations-2025

22 October 2025

Motivation

Simulation studies are the **universally used tool** (and, in many instances, the only tool) for evaluations of novel statistical methods.

Simulations have been proven particularly useful when evaluating

- statistical methods for which **no closed-form analytical solutions** are available
- the properties of statistical methods with closed-form solutions **under range of conditions**

An assessment via simulations brings **a great deal of subjectivity**

- as it is based on the simulation conditions that are chosen by researchers, and
- the interpretation of findings can be equivocal due to methods' complexity.

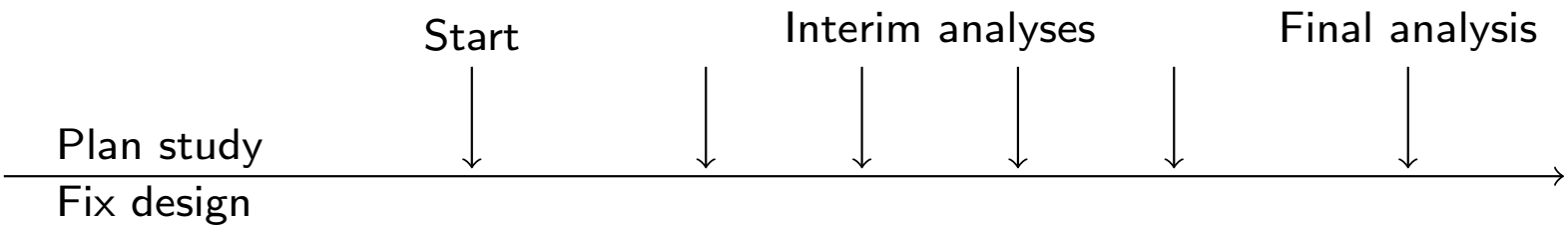
Motivation for simulations in clinical trials

There is an increasing uptake of **more flexible and more efficient** (and hence more complicated) clinical trial designs.

There are two main classes:

- Adaptive designs
- Bayesian designs

Adaptive design



At each interim:

- decide whether or not to stop
- change dose/regimen
- change sample size
- change population
- drop or add a treatment
- ...

Adaptive Designs[1]

- Benefits: "better" and/or faster decisions
- Challenges:
  - Operationally more complex
  - Care needed around false decision risk
  - Risk of bias
- For (complex) adaptations, such designs **typically require simulations** as closed-form solutions for type I error and power might not be available

[1] Pallmann et al. (2018) Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*. 16:29

Bayesian Designs

- Determining a prior
- Updating the prior with data
- Make conclusion based on full distribution of effect

\+ Can use all information available for decision making

\+ Provide information on a scale that stakeholders are often more familiar with

– Sample size estimation can be complex

– Prior input always present at final result

- Simulations are needed as there are **no closed-form solutions** for complex (realistic) models;
- Simulations are needed to understand **the impact of the priors**

Definition of a simulation study

- A **simulation study** consists of generated data sets (simulated trials)
- Each **simulated trial** consists of outcomes for hypothetical subjects given the set of conditions (referred to as a "simulation scenarios")
  - distributions of outcomes;
  - characteristics of study participants;
  - and forms of dependencies between them;
  - sample size;
  - mechanism of missingness;
  - timing of adaptations;
  - recruitment rates
  - ...
- **Numerous realisations** of simulated trials are analysed and operating characteristics (OCs) are computed to describe the trial's properties
- This is **repeated under other scenarios** and the overall conclusion of method's performance is made

Role of simulations for design of clinical trials

The **main objectives** of a simulation study are to

- calculate sample size, type I error and power (and more)
- understand the behaviour under various design and scenario parameters and misspecification(s)

However, we also use the simulations to

- Check that method (and the code) does the intended analysis
- Better understand statistical concepts
- Confirm mathematical results
- Compare methods head-to-head
- Check robustness of the code

(Un)popular opinion

Phase I trial design: Is 3+3 the best? — Hansen *et al.* (2014)

> *The evidence from this review suggests that the 3+3 design identifies the recommended phase 2 dose and toxicities with an acceptable level of precision in some circumstances*

> *Novel trial designs demonstrating superiority over the 3+3 method in statistical simulations without corroborating clinical evidence are of theoretical value alone*

What comes first the simulations (chicken) or the practice (egg)?

Simulations should define what we will **actually do in practice**!

Regulatory perspective - FDA

FDA sources:

- Adaptive Designs for Clinical Trials of Drugs and Biologics
- Interacting with the FDA on CIDs

  For many Complex Innovative Designs (CIDs) the properties (e.g. type I error, power, sample size, etc.) cannot be derived analytically. Simulation studies are often required.

Sponsors are encouraged to **engage with FDA early** to discuss the proposed complex design, **including the simulation plan**.

Special attention to: type I error control (false positives), power, bias in estimation, confidence interval coverage, and mean squared error.

Simulation is **one of the inputs** into that decision, not the only one.

Regulatory perspective - ICH E20

ICH E20 adaptive designs for clinical trials

> *If simulations are critical to understand operating characteristics of an adaptive design, the simulation study should be carefully planned, conducted, and reported. All relevant details pertinent to the planning of an adaptive trial design should be documented*

When: where trial features cannot be analytically derived.

Simulations should

- help ensure that after adaptations, estimates remain reliable
- show that under the planned design, Type I error is controlled
- show risks of false negatives under reasonable scenarios

The statistical analysis plan (and regulatory submission) should include a **description of the simulation plan**. The report should allow regulators to see how the design was evaluated by simulations.

Lack of guidance

Although there is increasing recognition of the important of well-designed and well-reported simulation studies, there is still **lack of guidance** even for specific sub-categories of clinical trial designs.

In many instance, you have to be an **expert in that particular area of clinical trial designs** to know what to look out for.

Things to be improved:

- Reporting of simulation studies
- Reproducibility of simulation studies
- Selection of simulation scenarios
- Justification for simulation scenarios

ADEMP reporting framework

T Morris et al proposed ADEMP reporting framework which stands for

- **A**ims
  Set out what question(s) the simulation study is designed to address
- **D**ata Generation
  Set out how the simulated data will be created
- **E**stimands
  What quantity the analysis is trying to estimate
- **M**ethods
  Set out how each simulated data set will be analysed to give an estimate of the estimand
- **P**erformance measures
  What is a good result?

ADEMP can enhance the validity and transparency of simulations

Useful for the analysis but of limited guidance in *trial design* settings

Good practices

Similar to the guidelines for RCTs, statistical simulations should meet several criteria, and address important design/reporting issues:

- How to simulate data in a **realistic way**?
- How to simulate data to full **unveil the properties** of the design?
- How to ensure the **reproducibility** and transparency of the methods used for data generation and analyses?
- What **parameters and assumptions should be varied** across simulated scenarios?
- Which "**competing methods**" should be considered?
- How can the results be **interpreted**, without the risk of over-interpretation?

Good practices

It is often useful to **separate code** for the three tasks:

1. Function(s) to **generate data** sets
2. Function(s) to **analyse** the generated/observed data
3. Function(s) that calls 1. and 2 to **store/presents the results** in a well-structured outputs

Start small and build up slowly! When you build something on, check it works as expected.

Good practices

In many instances we need to **vary several parameters at the same time** (not one at a time) to understand how interaction of various assumptions will affect the trial's properties.

As a result, we will be dealing with huge (!) simulation studies across many different scenarios. It might be challenging to navigate across them but we can

- Plot the distribution across simulations
- Average across a design/model parameter to understand the patterns;
- Study "outliers"
- Explore individual datasets to understand the behaviour;
- Always include a comparator (reference method/benchmark)

Common operating characteristics

- Type I error
- Power
- Expected sample size
- Distribution of the realised sample sizes
- Properties of the treatment effect estimator
  - Bias
  - SE
  - MSE
- Width and coverage of the confidence/credible intervals
- Convergence
- Computational speed

Simulation (Monte Carlo) error

TABLE 6 Performance measures: definitions, estimates and Monte Carlo standard errors

Performance Measure	Definition	Estimate	Monte Carlo SE of Estimate
Bias	$E[\hat{\theta}] - \theta$	$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\hat{\theta}_i - \theta$	$\sqrt{\frac{1}{n_{sim}(n_{sim}-1)}\sum_{i=1}^{n_{sim}}(\hat{\theta}_i - \bar{\theta})^2}$
EmpSE	$\sqrt{Var(\hat{\theta})}$	$\sqrt{\frac{1}{n_{sim}-1}\sum_{i=1}^{n_{sim}}(\hat{\theta}_i - \bar{\theta})^2}$	$\frac{\widehat{EmpSE}}{\sqrt{2(n_{sim}-1)}}$
Relative % increase in precision (B vs A) [a]	$100\left(\frac{Var(\hat{\theta}_A)}{Var(\hat{\theta}_B)} - 1\right)$	$100\left(\left(\frac{\widehat{EmpSE}_A}{\widehat{EmpSE}_B}\right)^2 - 1\right)$	$200\left(\frac{\widehat{EmpSE}_A}{\widehat{EmpSE}_B}\right)^2\sqrt{\frac{1-Corr(\hat{\theta}_A,\hat{\theta}_B)^2}{n_{sim}-1}}$
MSE	$E[(\hat{\theta} - \theta)^2]$	$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}(\hat{\theta}_i - \theta)^2$	$\sqrt{\frac{\sum_{i=1}^{n_{sim}}\left[(\hat{\theta}_i-\theta)^2-\widehat{MSE}\right]^2}{n_{sim}(n_{sim}-1)}}$
Average ModSE[a]	$\sqrt{E[\widehat{Var}(\hat{\theta})]}$	$\sqrt{\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\widehat{Var}(\hat{\theta}_i)}$	$\sqrt{\frac{\widehat{Var}[\widehat{Var}(\hat{\theta})]}{4n_{sim}\times\widehat{ModSE}^2}}$ [b]
Relative % error in ModSE[a]	$100\left(\frac{ModSE}{EmpSE} - 1\right)$	$100\left(\frac{\widehat{ModSE}}{\widehat{EmpSE}} - 1\right)$	$100\left(\frac{\widehat{ModSE}}{\widehat{EmpSE}}\right)\sqrt{\frac{\widehat{Var}[\widehat{Var}(\hat{\theta})]}{4n_{sim}\times\widehat{ModSE}^4} + \frac{1}{2(n-1)}}$ [b]
Coverage	$Pr(\hat{\theta}_{low} \le \theta \le \hat{\theta}_{upp})$	$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}1(\hat{\theta}_{low,i} \le \theta \le \hat{\theta}_{upp,i})$	$\sqrt{\frac{\widehat{Cover.}\times(1-\widehat{Cover.})}{n_{sim}}}$
Bias-eliminated coverage	$Pr(\hat{\theta}_{low} \le \bar{\theta} \le \hat{\theta}_{upp})$	$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}1(\hat{\theta}_{low,i} \le \bar{\theta} \le \hat{\theta}_{upp,i})$	$\sqrt{\frac{\widehat{B\text{-}E\ Cover.}\times(1-\widehat{B\text{-}E\ Cover.})}{n_{sim}}}$
Rejection % (power or type I error)	$Pr(p_i \le \alpha)$	$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}1(p_i \le \alpha)$	$\sqrt{\frac{\widehat{Power}\times(1-\widehat{Power})}{n_{sim}}}$

[a] Monte Carlo SEs are approximate for *Relative % increase in precision*, *Average ModSE*, and *Relative % error in ModSE*.

[b] $\widehat{Var}[\widehat{Var}(\hat{\theta})] = \frac{1}{n_{sim}-1}\sum_{i=1}^{n_{sim}}\{\widehat{Var}(\hat{\theta}_i) - \frac{1}{n_{sim}}\sum_{j=1}^{n_{sim}}\widehat{Var}(\hat{\theta}_j)\}^2$.

ICH E20 explicitly mentions $\geq 100,000$ for the estimation of type I error.

Coding errors

Coding errors are inevitable but we can try to isolate them with simple techniques

- Check of the results that look "off" - do we have a good explanation for them?
- Compare to the benchmark/reference method (e.g. no adaptations)
- Plot the distribution of the estimates under a given scenario
- Explore the individual datasets (possibly with a LARGE sample size)
- Isolate different parts of the "complexity" (e.g. remove some of the adaptations)
- Sticking to the 3-part code (as per above) helps a lot
- Double-programming

Discussion

Simulation is a powerful (and accessible) tool to better understand properties of clinical trial designs and to make an informed decision.

Unfortunately, many simulations studies are not reproducible or/and the choices for them are not reproducible.

This is due to a limited guidelines/advice even for the specific sub-categories of trial designs

In this course, we will try to tackle this!