Bayesian Non-Inferiority Trials for Dose Finding.

Connor Fitchett
MRC Biostatistics Unit
University of Cambridge

14th May, 2026

MRC
Biostatistics
Unit

UNIVERSITY OF
CAMBRIDGE

Overview

1. Introduction
2. RAR Framework
3. Results
4. Discussion

Table of Contents

Dose Finding

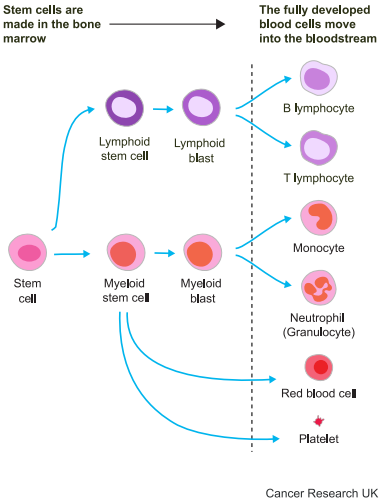
- Dose finding is the challenge of evaluating the **best dose** of a drug from a range of options.
- **Project Optimus** is an FDA initiative currently promoting **innovative and rigorous** early phase dose finding, in the hope that selected doses have an optimal safety and efficacy profile (FDA, 2024).
- However, classical dose finding methods could fail to consider safety and efficacy simultaneously, leading to **poorly characterised dose response profiles**.
- Thus, there are many drugs currently on the market whose doses may be able be to optimised more so than they currently are.

Late Phase Dose Finding

- This is why we consider late phase dose finding. As we have already established efficacy with drugs on the market, typically we want to show that the efficacy of a new dose is **not worse** whilst having other better properties.
- We establish a 'not worse' efficacy by implementing a **non-inferiority trial** (NI). Rather than attempting to show superiority, we attempt to show that the new dose is not worse than the standard of care by some **NI margin**.
- Determining the NI margin is hugely **complex** however. What is the best trade off between expected efficacy losses and alternative gains? It is hugely dependent on context, which is why we focus on a case study in this work.

Acute Lymphoblastic Leukaemia (ALL)

- ALL is a cancer that effects the blood: the cancerous cells take away the resources needed to produce blood cells. (Cancer Research UK, 2024).
- Studies show that **reducing** some of the older drugs, or **replacing** them with additional doses of more novel drugs, can lead to **improved outcomes**. (Gupta et al., 2025).
- This motivates a hypothetical ALL '**de-intensification**' study.



Three Questions

What hypothesis testing strategy is best?	**What is the best way to randomise?**	**What benefits does time to event data bring over binary?**
- Benefits of early stopping. - Bayesian vs. Frequentist tests.	- Response-adaptive or equal? - How do we account for safety and efficacy data in the randomisation.	- Different ways to analyse time to event data - Comparing against binary data.

Three Questions - Focus

	What is the best way to randomise?	
	- Response-adaptive or equal? - How do we account for safety and efficacy data in the randomisation.	

Table of Contents

Types of RAR - BRAR and ER

- **Equal Randomisation** (ER) is simply allocating with **equal probability** to both arms, ensuring an even split.

$$a_t = 0.5$$

- **Response Adaptive Randomisation** (RAR) adapts the allocation probability based on how well the arms are performing relative to one another. Bayesian RAR (BRAR) is an example this, where the allocation probability is proportional to the **probability one treatment is better than another** (Thompson, 1933):

$$a_t = P(\text{The next patient is allocated to the experimental treatment}) \propto P(\theta_1 > \theta_0 | D)^c$$

Types of RAR - RITS

- **Risk Inclusive Thompson Sampling** (RITS) applies BRAR on both the safety and efficacy data and takes a **weighted average** of the two (Kanrar et al., 2025). If $a_{t,\text{eff}}$ is the result of BRAR applied on the efficacy data, and $a_{t,\text{safe}}$ is the result of of BRAR applied on the safety data, then our allocation with RITS becomes:

$$a_t = w a_{t,\text{safe}} + (1 - w) a_{t,\text{eff}}$$

- So this scheme attempts to **balance safety and efficacy** performance in order to prioritise the treatment that performs well in both.
- w, the weight, can be specified a priori to prioritise the safety or efficacy results based on the trial specifications.

Types of RAR - Gaussian RITS

- **Gaussian RITS** is our novel contribution to RITS which allows it to **adaptively update its weighting**: preferring efficacy when there is evidence of superiority and preferring safety when there is evidence of non-inferiority.

$$a_t = w_\sigma a_{t,\text{safe}} + (1 - w_\sigma) a_{t,\text{eff}}$$

$$w_\sigma = e^{-\frac{(P(\text{Efficacy BRAR allocation}) - 0.5)^2}{\sigma}}$$

- This allows us to favour efficacy in situations where the safety and efficacy data are in conflict, but also allows us to not weaken the safety allocation when there is non-inferiority.

Types of RAR - SAFER

- **Safety-Aware Flexible Elastic Randomization** (SAFER) is another type of RAR that **varies** between **equal allocation** and the **safety Neyman allocation** (RAR algorithm that optimises power).
- When we have **equivalence** of efficacy (when $\hat{\Phi} < 0.5$), we have **equal randomisation**
- When we have one arm demonstrating **superiority** over the other ($\hat{\Phi} \approx 1$), we have **Neymann allocation** ($\hat{\pi}_E$).

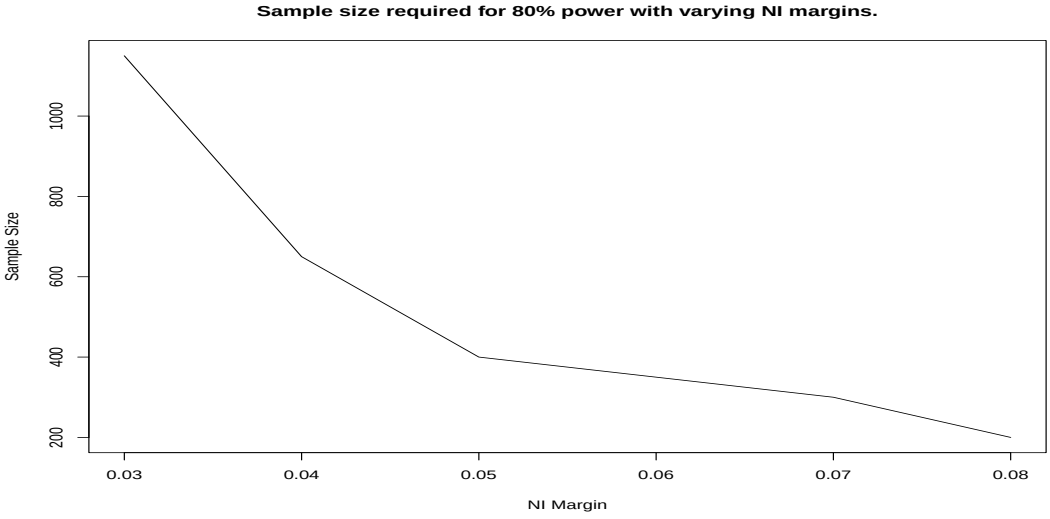
$$\text{SAFER}(\hat{\pi}_E) = \begin{cases} 0.5 & \text{if } \hat{\Phi} \leq 0.5 \\ 0.5 + (\hat{\pi}_E - 0.5) \cdot \left(1 - \left(1 - \dfrac{\hat{\Phi} - 0.5}{0.5}\right)^{\eta}\right) & \text{if } 0.5 < \hat{\Phi} < 1 \\ \hat{\pi}_E & \text{if } \hat{\Phi} = 1 \end{cases}$$

Table of Contents

Simulation Details

Control success rate	0.96
Low toxicity rate	0.4
High toxicity rate	0.5
Number of interims	4
Target type one error	5%
Target power	80%
Number of simulations	$10,000$
Test	Confidence interval testing
Non-inferiority margin	...
Sample size	...

Choosing the NI margin

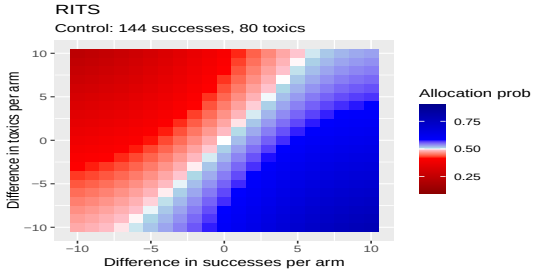


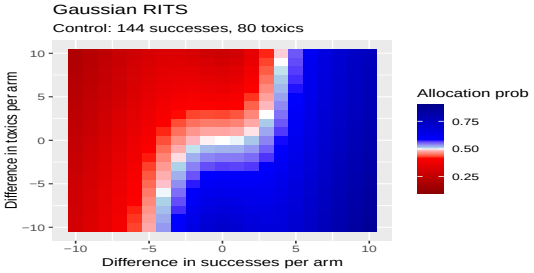
Choosing The NI Margin - Continued

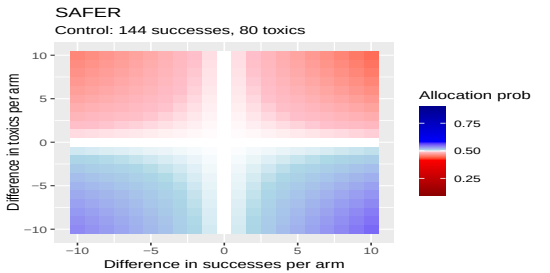
- It is a hugely **complex and complicated** decision to choose the NI margin. Not only does it effect sample size, but it quantifies how much of a drop in efficacy you're willing to accept. In our case that directly translates to a loss of survival.
- We choose an **NI margin of** 0.06 and a corresponding sample size of 320. This is not a clinical recommendation, and instead just convenient for our simulations. A proper selection would look at expert opinions and evaluate the health economics of the decision.
- Then, noting that $\theta_{0,1}$ represents the survival rate and $\phi_{0,1}$ represents the toxicity rate for the control and experimental arm receptively, the table below gives us the following parameters for our null (H0) and alternative (H1) hypothesis:

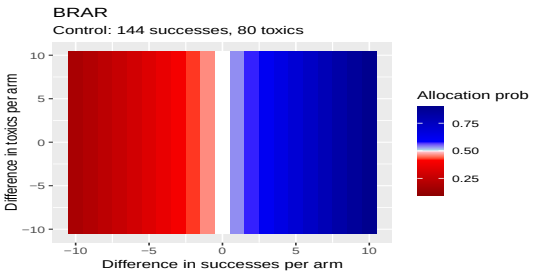
Hypothesis	θ_0	θ_1	ϕ_0	ϕ_1
H0	0.96	0.9	0.5	0.5
H1	0.96	0.96	0.5	0.4

Theoretical Performance - Allocation Probability









Results - H0

Survival rate in control: 96%. Survival rate in experimental: 90%.

Toxicity rate in control: 50%. Toxicity rate in experimental: 50%.

Randomisation	Alpha	ESS	Proportion to control	Number of deaths
ER	0.051	164.0 (79.9)	0.500 (0)	11.4 (5.6)
RITS	0.051	163.9 (79.8)	0.506 (0.013)	11.4 (5.6)
Gaussian RITS	0.047	164.1 (79.7)	0.510 (0.022)	11.4 (5.6)
SAFER	0.051	163.4 (79.4)	0.495 (0.013)	11.4 (5.6)
BRAR	0.048	164.6 (80.1)	0.529 (0.026)	11.2 (5.4)

Results - H1

Survival rate in control: 96%. Survival rate in experimental: 96%.

Toxicity rate in control: 50%. Toxicity rate in experimental: 40%.

Randomisation	Power	ESS	Proportion to experimental	Number of toxics
ER	0.819	212.0 (80.9)	0.500 (0)	95.4 (37.2)
RITS	0.820	209.9 (80.2)	0.513 (0.017)	94.2 (36.7)
Gaussian RITS	0.817	216.6 (74.1)	0.521 (0.037)	97.0 (34.0)
SAFER	0.820	210.8 (80.2)	0.514 (0.015)	94.6 (36.6)
BRAR	0.800	217.2 (74.4)	0.480 (0.037)	98.1 (34.9)

Table of Contents

Summary

- RITS and Gaussian RITS consistently allocate **more patients to the superior treatments** in both null and alternative settings. RITS also has **fewer toxicity events** than any other scheme whilst not suffering in any other statistic.
- RAR that **balances both safety and efficacy** perform better in more scenarios than simply considering the efficacy.
- RAR that offer flexibility on the balance between safety and efficacy allow for NI trials to meaningfully have a **trade off between the two endpoints**. This acts on the frequent critique that NI trials prioritise efficacy over any potential secondary endpoint gains (Tannock et al., 2024).

Further questions

- How does this scale to multi-arm?
- Are we considering the most appropiate posterior probability? I.e. $P(\theta_1 > \theta_0 - 0.06|D)$ instead of $P(\theta_1 > \theta_0|D)$
- How would this randomisation extend to other endpoints/considerations? For example, if one treatment was significantly cheaper, or if we wanted to use Patient Reported Outcomes.

References

Cancer Research UK (2024). Phases of treatment for acute lymphoblastic leukaemia (ALL). https://www.cancerresearchuk.org/about-cancer/acute-lymphoblastic-leukaemia-all/treatment/phases. Last accessed on 2026-04-06.

FDA (2024). Project Optimus. https://www.fda.gov/about-fda/oncology-center-excellence/project-optimus. Last accessed on 2025-08-20.

Gupta, S., Rau, R. E., Kairalla, J. A., Rabin, K. R., Wang, C., Angiolillo, A. L., Alexander, S., Carroll, A. J., Conway, S., Gore, L., et al. (2025). Blinatumomab in standard-risk b-cell acute lymphoblastic leukemia in children. *New England Journal of Medicine*, 392(9):875–891.

Kanrar, R., Li, C., Ghodsi, Z., and Gamalo, M. (2025). Risk-inclusive contextual bandits for early phase clinical trials.

Tannock, I. F., Buyse, M., De Backer, M., Earl, H., Goldstein, D. A., Ratain, M. J., Saltz, L. B., Sonke, G. S., and Strohbehn, G. W. (2024). The tyranny of non-inferiority trials. *The Lancet Oncology*, 25(10):e520–e525.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

The End

Feel free to ask any questions

connor.fitchett@mrc-bsu.cam.ac.uk