Randomization in clinical trials with small sample sizes using group sequential designs

D. Bodden*[1], R.-D. Hilgers[1], Franz König[2]
[1] Institute of Medical Statistics, RWTH Aachen University, Germany
*Presenting author: dbodden@ukaachen.de
[2]Center for Medical Data Science, Medical University of Vienna, Vienna, Austria

Background and Objective

Group sequential designs (GSDs) allow for interim analyses to potentially stop trials early for efficacy or futility, making them especially valuable in small population trials, such as those for rare diseases. However, these designs require careful control of the overall type I error rate (T1E) due to multiple testing. Randomization, a critical component of trial design [1, 2], can influence GSD performance, particularly in small samples where imbalances may have pronounced effects. This study evaluates how different randomization procedures (RPs) affect type I error rate and power in small sample GSDs, focusing on Pocock, O'Brien & Fleming, Lan-DeMets, and inverse normal combination test designs [3, 4]. The aim is to provide guidance on selecting appropriate GSD and RP combinations for small population trials.

Group Sequential Design

Simulations of randomized controlled trials were conducted using a two-arm parallel group design with a group sequential approach and equidistant interim analysis intervals. A two-sided z-test was employed to assess the difference in the mean response for normal responses with a variance of $\mathbf{1}$. The GSDs employed were:

- **Pocock:** A type of efficacy boundaries in a GSD that adjusts the significance level equally for each stage to control the overall significance level α. Originally proposed for GSD with equally sized stage-wise sample sizes resulting in the same boundaries across all stages.
- **O'Brien & Fleming (OBF):** A type of efficacy boundaries in a GSD that adjusts the significance level for each stage to control the overall significance level α. The design starts with a low significance level at early stages, gradually increasing it with each successive test.
- **Lan-DeMets (LDM):** A more flexible way to derive boundaries in a GSD using a pre-defined alpha spending function based on the observed information for each stage. This allows for unequal stage-wise sample sizes and adjustments in the timing and number of analyses. The alpha spending function can be chosen to resemble OBF and Pocock type boundaries.
- **Inverse Normal Combination Test (INCT):** The INCT combines stage-wise test statistics using a predefined combination function, rather than accumulating data as in standard GSD. INCT has the advantage of allowing standard GSD boundaries, as described above, while also permitting data-dependent adaptations [4], such as adjustments to the allocation ratio and sample size, all while strictly controlling the type I error rate.

The illustration below demonstrates how deviations from an intended $\mathbf{1:1}$ allocation ratio at interim stages can impact test statistics and decision boundaries in GSDs.

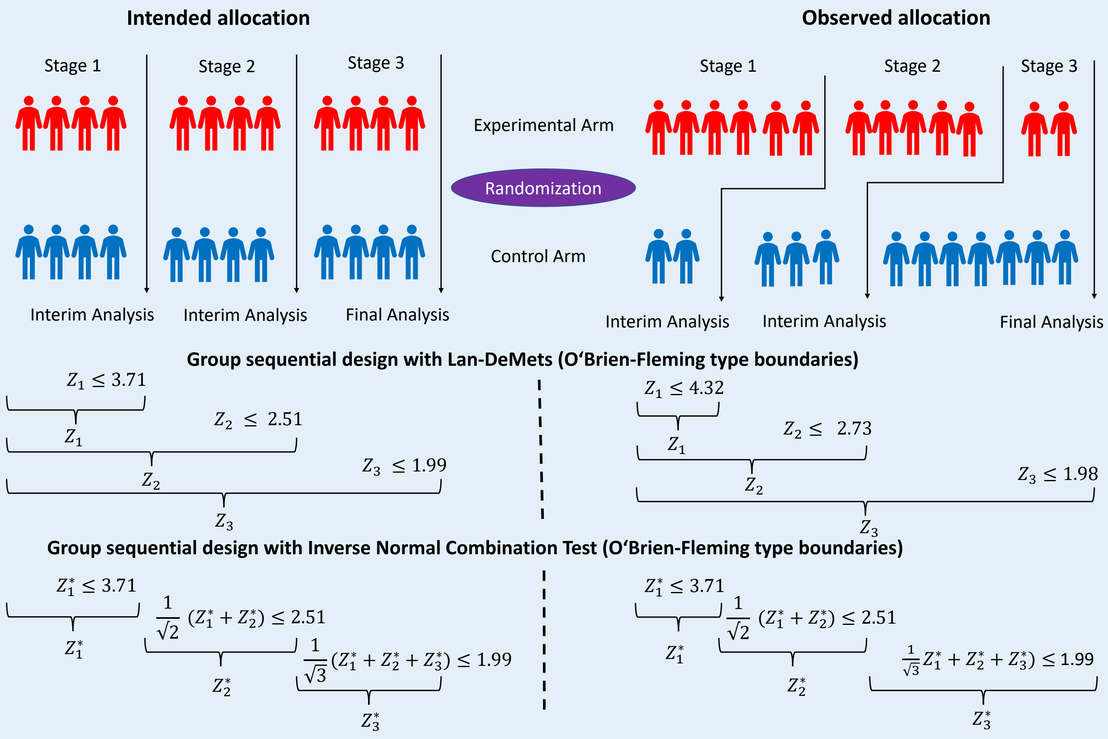

Figure: Visualization of allocation imbalances in group sequential designs.

Randomization Procedures

The following randomization procedures were evaluated:

- *Complete Randomization* (CR): Randomization achieved by flipping a fair coin. Sometimes this method is also referred to as simple or full randomization.
- *Random Allocation Rule* (RAR): Randomization assigning the same proportion of patients to each treatment.
- *Permuted Block Randomization* (PBR($\boldsymbol{l}$)): Allocation in blocks of length $\boldsymbol{l}$, with randomization within each block according to RAR.
- *Efron's biased coin* (EBC($\boldsymbol{p}$)): Randomization using a biased coin with probability $\boldsymbol{p}$ in favor of the treatment with fewer allocations and a fair coin in case of equal allocations.
- *Big stick design* (BSD($\boldsymbol{a}$)): Complete randomization with deterministic assignment when a maximum tolerated imbalance $\boldsymbol{m}$ is reached.
- *Chen's design* (Chen($\boldsymbol{p, a}$)): EBC($\boldsymbol{p}$) with deterministic assignment when a maximum tolerated imbalance $\boldsymbol{m}$ is reached.

Results

- T1E inflates when standard Pocock or OBF boundaries are used without adjusting the boundaries for the observed allocation ratio during the trial.
- LDM designs are robust to imbalances, maintaining power across various RPs.
- INCT designs require strict stage-wise allocation to avoid power loss, but always controls T1E.
- PBR performs best when stage size is divisible by the block size.

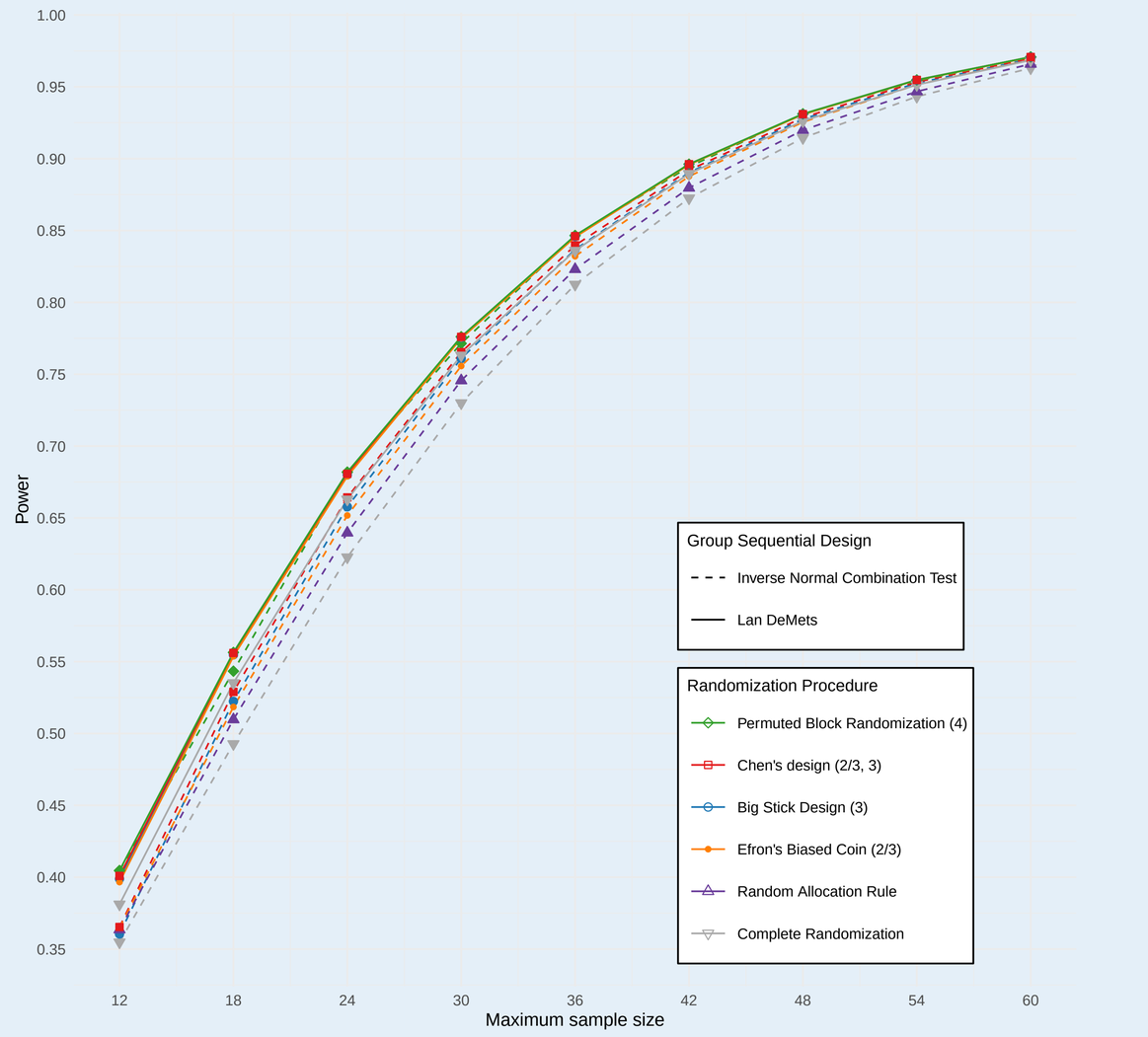

Figure: **Power as a function of the maximum sample size for $\delta = 1.0$ and $K = 3$ equidistant stages.** The power was computed for sample sizes ranging from $\mathbf{12}$ to $\mathbf{60}$ in steps of $\mathbf{6}$. Different RPs are shown with solid lines, and adjustments for repeated significance testing were made using LDM or INCT with OBF boundaries.

Recommendations for small samples

- Avoid RPs prone to imbalances (e.g., CR, RAR).
- Ensure stage-wise balance when using INCT designs.
- Use PBR only with appropriately chosen block sizes.

Table: **Framework for selecting the randomization procedure in group sequential designs for an intended $\mathbf{1:1}$ allocation ratio with equal stage-wise sample sizes.**

Randomization Procedure	Lan-DeMets	Inverse Normal	Standard Boundaries*
Complete Randomization	Low power; risk of allocations to single group	Low power; risk of allocations to single group	Type I error rate (T1E) not controlled
Random Allocation Rule	Low power; risk of allocations to single group	Low power; risk of allocations to single group	T1E not controlled
Big Stick Design (m)	For $m \leq n/K$	For $m \leq 2n/K$; power reduction relative to $\mathbf{1:1}$ allocation per stage	T1E not controlled
Permuted Block Randomization (l)		l should divide n/K for optimal power	l must divide n/K to control T1E
Efron's Biased Coin (p)	Moderate power; risk of allocations to single group	Moderate power; risk of allocations to single group	T1E not controlled
Chen's Design (p, m)	For $m \leq n/K$	For $m \leq 2n/K$; power reduction relative to $\mathbf{1:1}$ allocation per stage	T1E not controlled

The maximum sample size is n and K is the number of pre-planned stages. Therefore, the stage-wise sample size is given by n/K. The framework is intended for cases with small stage-wise sample sizes (say $n/K \leq 20$). For the inverse normal combination test, weights proportional to the pre-planned stage-wise sample sizes were used. * With standard boundaries we refer to the case where the boundaries are calculated assuming an $\mathbf{1:1}$ allocation ratio.

References

[1] ICH E9: Statistical principles for clinical trials (1998).

[2] Hilgers, R.-D., Uschner, D., Rosenberger, W.F., Heussen, N.: ERDO - a framework to select an appropriate randomization procedure for clinical trials. BMC Medical Research Methodology 17(1), 159 (2017).

[3] Jennison, C., Turnbull, B.W.: Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, New York (1999).

[4] Wassmer, G., Brannath, W.: Group Sequential and Confirmatory Adaptive Designs in Clinical Trials. Springer Series in Pharmaceutical Statistics. Springer, Cham (2016).

This paper has been submitted and can be requested by the authors. The research is funded by the European Rare Disease Research Coordination and Support Action consortium (ERICA) funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement. no. 964908.