Cytel

Statistical Design and Planning of an Adaptive Trial w/ Hierarchical Composite Outcomes: A Practical Example

Krishna Padmanabhan, Cyrus Mehta
Cytel Inc.

ASA NJ Chapter Meeting, June 2025

Summary

- Clinical trials using composite outcomes in a hierarchy are gaining popularity, especially in the CV space.
- The Finkelstein and Schoenfeld (FS) test is a generalized pairwise comparison approach to analyze prioritized composite endpoints
- Using the FS-method as the primary statistical analysis and quantifying the magnitude of effect using Win ratio (WR) is a strategy many sponsors are employing in this situation.
- However, adaptive trial desings using this analytic approach are not as well-known.
- This presentation describes the successful design and implementation of an adaptive clinical trial with composite hierarchical endpoints, using an SSR at an Interim Analysis.

Finkelstein and Schoenfeld (FS) Method

- Used frequently in cardiovascular trials that have hierarchical composite outcomes.

- Pairwise comparisons based on a clearly pre-specified clinical hierarchy:
  - Mortality (highest priority)
  - Cardiovascular hospitalizations (CVH)
  - Functional improvement (6MWT, NT-proBNP, PRO instrument etc.)

- Method is based on pairwise comparison of all patients in the trial across each of the 3 endpoints in order to determine a non-parametric estimate of the number of “wins”.

FS-Method

Steps for the Finkelstein–Schoenfeld Test

1. **Create Pairwise Comparisons:**
   - Randomize n_1 patients to placebo and n_2 to active treatment.
   - Form all $n_1 \times n_2$ pairwise comparisons (patient i in placebo vs. patient j in active).
2. **Compare the Highest-Priority Endpoint (Mortality):**
   - If one patient died and the other did not, the group without death gets a "win."
   - If both died or both survived, proceed to the next endpoint.
3. **Compare of Cardiovascular Hospitalizations:**
   - If one patient had more CV hospitalizations than the other in the shorter of the follow up times, then they "win"
   - If tie cannot be broken, proceed to the next endpoint.
4. **Compare Functional Improvement:**
   - If one patient's 6MWT improvement is greater, assign a "win" to that treatment arm.
   - If both are equal, this pair is considered a "tie."
   - Sum up all wins/losses ($+1$ or -1) across all pairs to obtain the overall Finkelstein–Schoenfeld statistic.

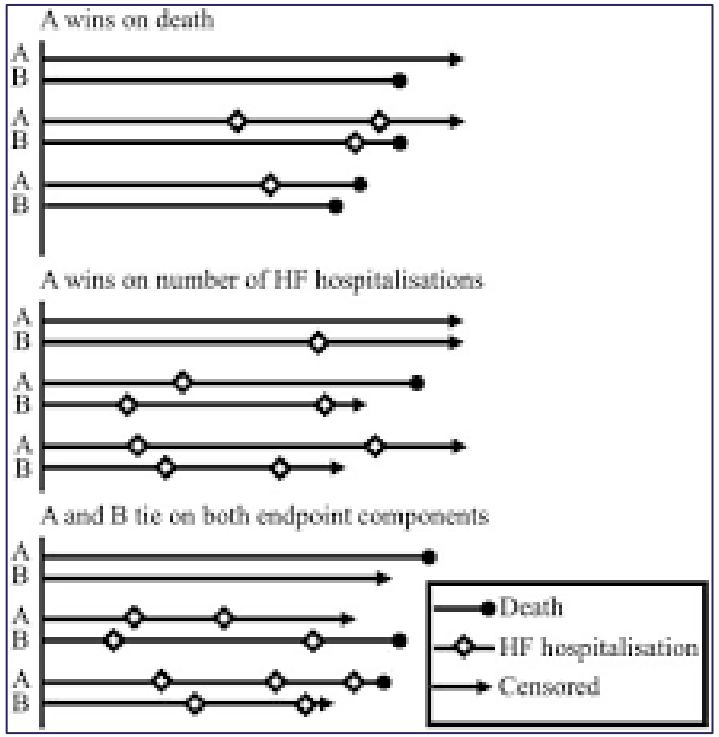


Image from - Redfors et. Al: European Heart Journal (2020) 41, 4391–4399

Large Sample distribution

- Test statistic:

$$T = \sum_{i=1}^{2n} Z_i U_i$$

- Mean:

$$E(T) = 2n^2(\theta - 1/2)$$

- Variance:

$$Var(T) = \frac{n^2}{2n(2n-1)} \sum_{i=1}^{2n} U_i^2$$

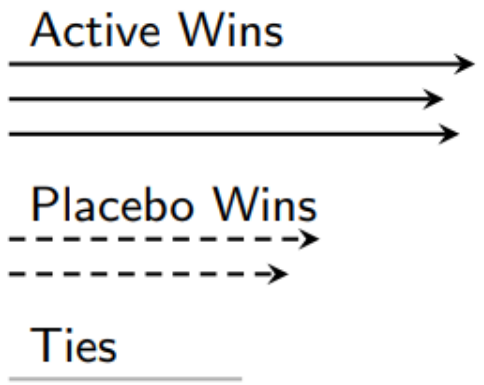
- U_i is the score obtained by summing all the wins (+1) and losses (-1) for an individual patient i.
- T is the test statistic calculated by summing the total wins and losses across all patients for the active treatment.
- θ is the probability that a random subject i in the treatment group has a better outcome than a random subject j in the control group.
- Therefore, H_0 is: θ = 1/2

Win Ratio: Quantifying Treatment Benefit

1. **Count Wins & Losses**
   - Record $\text{Wins}_{\text{active}}$ and $\text{Losses}_{\text{active}}$
   - Unresolved pairs $\rightarrow$ *Ties* (ignored)
2. **Compute the Win Ratio**

$$\text{WR} = \frac{\text{Wins}_{\text{active}}}{\text{Losses}_{\text{active}}}$$

3. **Confidence Interval & p-Value**
   - Delta-method or bootstrap variance on pair scores
   - 95 % CI (log-scale back-transformed)
4. **Clinical Interpretation**
   - WR = 1.40 $\rightarrow$ "A 40% higher likelihood of a better outcome with the treatment compared to control"



Example: Wins = 180, Losses = 128, Ties = 62

WR = 180/128 = 1.40

Trial Design for new CV investigational agent

- Hierarchical Endpoints:
  - Mortality
  - # of CV Hospitalizations
  - Functional response (0/1)
- Possible Adaptations at IA
  - Non-binding futility
  - SSR

- Sample size: $N = 400$, interim at $N = 200$.
- Historical rates:
  - Mortality: 30% (Active) vs. 40% (Control).
  - CVH Hazard: 0.25 (Active) vs. 0.4 (Control).
  - Functional response: 50% (Active) vs. 25% (Control).

- Predictive Probability (PP) determines SSR decision.
- Decision zones:
  - $PP < 10\%$: Stop for futility.
  - $10\% \leq PP < 30\%$ or $PP \geq 90\%$: Continue with $N = 400$.
  - $30\% \leq PP < 75\%$: Increase to $N = 600$.
  - $75\% \leq PP < 90\%$: Increase to $N = 500$.

Optional Initial Sample Size Determination: Using Win Ratio An Alternative to Simulation based approach

(via Yu, Ganju 2021)

Win Ratio

- The **win ratio (WR)** extends the FS statistic by expressing the results as a ratio.
- The **win ratio = (number of favorable comparisons) / (number of unfavorable comparisons)**.
- It summarizes how often the treatment group "wins" (has a better outcome) compared to the control.
- A win ratio >1 suggests that the treatment is beneficial.

- Simulate patient data per the desired HA.
- Estimate empirical WR
- Estimate p_tie
- Use asymptotic distribution to calculate N_Init.
- N_Max can be varied based on WR under MCID.

The approximate variance of the log of the win ratio is

$$\mathrm{Var}(\ln(WR)) \approx \frac{1}{N} \times \left\{ \frac{4(1+p_{tie})}{3k(1-k)(1-p_{tie})} \right\} \quad (1)$$

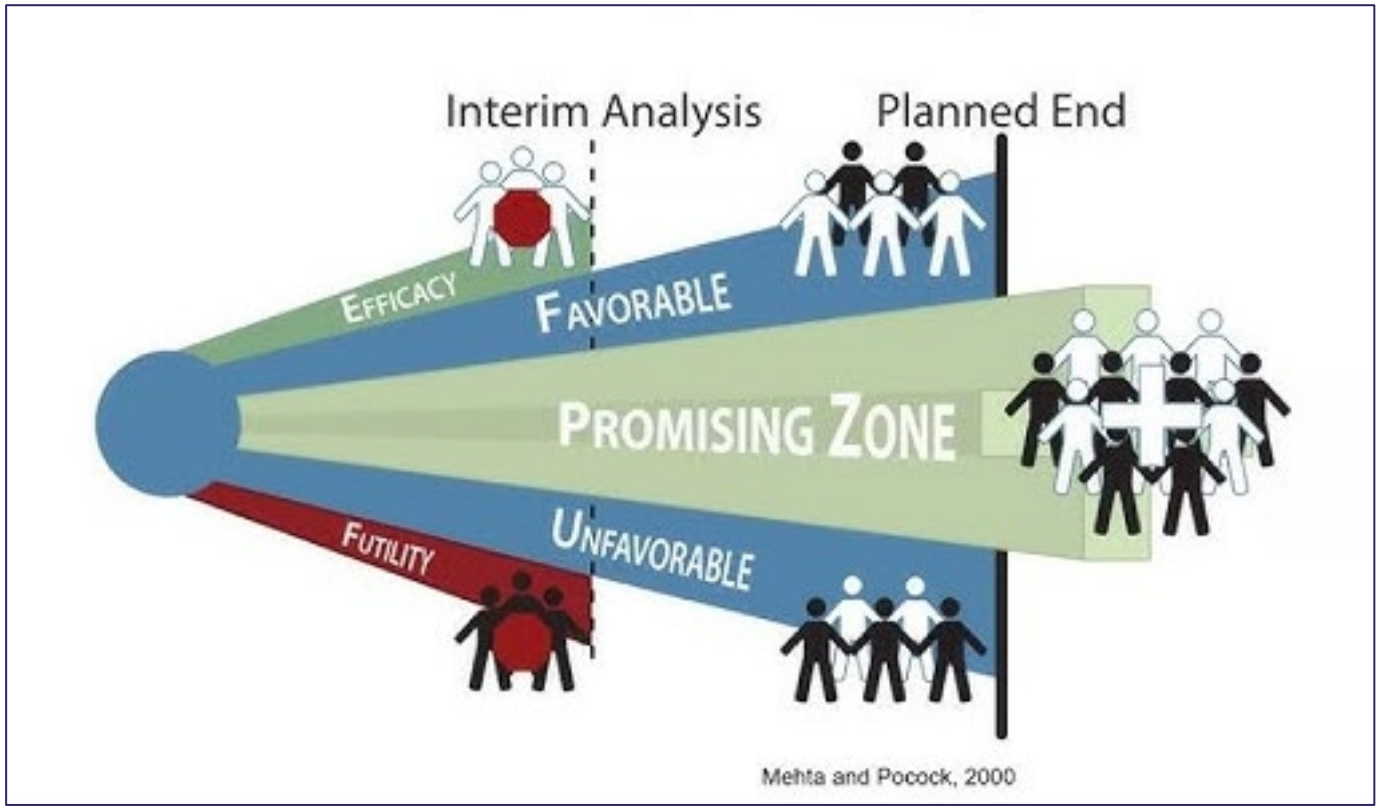
Using conventional notation, we write this as $\frac{1}{N} \times \sigma^2$, where σ^2 represents the term within {}.

The required sample size is

$$N \approx \frac{\sigma^2 \times \left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\ln^2(WR_{true})} \quad (2)$$

where WR_{true} refers to the assumed or true value of the win ratio, and $\ln(WR_{true})$ refers to its natural log-transformed value. As usual, α and β refer, respectively, to type I and II error rates

Adaptive Design with unblinded SSR



Operating Principles:

- Ability to increase SS at a pre-specified IA to re-size the study appropriately
- Typically operationalized using conditional Power
- Pre-specified IA timing, Max SS, T1E methodology and Final analysis method are required
- Frequently accepted by agencies as a well-understood adaptive design.

CP vs Approx. PP: Both result in similar OCs

CP provides finer control, but approx. Approx. PP is easier to implement

The first 200 participants (from both arms combined) will be assigned to cohort 1 while the subsequent 200 planned participants (from both arms combined) will form cohort 2

Let

$$w_i = \sqrt{\frac{n_i}{n}}, \quad i = 1, 2,$$

be the weights that we will use to combine the F-S statistic T_1, obtained from the cohort 1 data, and the F-S statistic T_2, obtained from the cohort 2 data, such that the standardized test statistic for testing the null hypothesis at the final analysis is

$$w_1 \frac{T_1}{s_1} + w_2 \frac{T_2}{s_2},$$

where s_j is the standard error of T_j. Then the CP, given $T_1 = t_1$, is

$$CP(t_1, n_2) = P\left(w_1 \frac{t_1}{s_1} + w_2 \frac{T_2}{s_2} > z_\alpha \mid T_1 = t_1 \right).$$

Since T_1, T_2 are independent and asymptotically normal, the CP can be evaluated as

$$CP(t_1, n_2) = 1 - \Phi\left(z_\alpha - \frac{w_1 t_1}{s_1} - w_2 \frac{E(T_2/s_2)}{\sqrt{n_2}} \right),$$

where $E(T_2/s_2)$ is estimated by repeatedly generating simulated responses for the n_2 subjects in cohort 2.

$$PP(p_n, r, \alpha) = \Phi\left(\frac{\Phi^{-1}(1 - p_n) - \Phi^{-1}(1 - \alpha)\sqrt{r}}{\sqrt{1 - r}} \right)$$

Image from – Marion et. al. 2024

Final analysis

- Predictive Power & Sample Size Re-evaluation: After estimating predictive power, stage 2 data is collected, and the final analysis occurs once all subjects complete 12 mo. or are terminated early.
- Hypothesis Testing: The null hypothesis is rejected based on the weighted combination of test statistics from both stages.
- Rejection Criteria:
  - No SSR at IA: $w_1 \frac{t_1}{s_1} + w_2 \frac{t_2}{s_2} \geq z_\alpha$
  - SSR at IA: $w_1 \frac{t_1}{s_1} + w_2 \frac{t_2^*}{s_2^*} \geq z_\alpha$

Equivalent to CHW method, that controls Type I error.

Weights $w_1 = \sqrt{n_1/n}$ and $w_2 = \sqrt{n_2/n}$ remain fixed

Simulation Study: Joint Frailty Model and Scenarios

For person i, let $T_{i0} = 0$ and $T_{i1}, T_{i2}, \ldots, T_{iN_i}$ represent recurrent event times. Here, N_i counts events before $X_i = \min(C_i, D_i)$, where C_i is censoring time and D_i is CV death time. The hazard functions define the JFM:

$$r_i(t \mid \omega_i) = \omega_i \exp\{\beta_1 z_i\} r_0(t) = \omega_i r_i^*(t)$$
$$\lambda_i(t \mid \omega_i) = \omega_i^\alpha \exp\{\beta_2 z_i\} \lambda_0(t) = \omega_i^\alpha \lambda_i^*(t).$$

For patient i, r_i gives the hospitalization hazard, scaling with baseline intensity r_0 and frailty ω_i. The CV death hazard λ_i builds on baseline λ_0, with β_1, β_2 capturing treatment effects (z_i). Random effects ω_i follow a lognormal distribution (mean 0, variance θ).

Table 2: Hypothetical Scenarios to be Explored

Five Hypothetical Scenarios (Active vs. Control)					
Event Rates	Null	Alternative	Alternative2	Middling	Middling2
Mortality	40% vs 40%	30% vs 40%	40% vs 40%	32.5% vs 40%	35% vs 40%
CVH Rate	0.4 vs 0.4	0.25 vs 0.375	0.10 vs 0.40	0.275 vs 0.40	0.28 vs 0.4
Func. Resp.	25% vs 25%	50% vs 25%	60% vs 25%	45% vs 30%	40% vs 25%

Results (10K runs for each scenario)

Table 3: Trial Operating Characteristics Without Sample Size Re-estimation

Scenario	Power (%)	Average N	Futility Stop (%)
Null	2.3	261.5	68.2
Alternative	90.9	378.6	2.8
Alternative2	73.4	384.3	7.9
Middling	63.6	377.0	11.4
Middling2	47.0	348.6	19.2

Table 4: Trial Operating Characteristics With Sample Size Re-estimation

Scenario	Power (%)	Average N	Futility Stop (%)	Max SSR (%)
Null	1.9	286.8	69.1	12.5
Alternative	94.5	437.0	2.5	21.3
Alternative2	81.5	473.4	8.4	45.1
Middling	72.2	471.7	11.7	48.1
Middling2	53.3	425.6	20.0	32.8

Conclusions

Fixed Design Without SSR:

- Strong Type I error control (2.3%) with high futility stopping under the null (68.2%).
- High power (82.9% - 90.9%) in alternative scenarios with moderate sample sizes (373 - 379).
- Power drops significantly in middling scenarios (47 - 52%), highlighting risk when true effects are marginal.

Adaptive Design with SSR:

- Maintains Type I error control (1.9%) while preserving high futility stopping (69.1%).
- Increases power in alternative scenarios (88.7-94.5%) at the cost of larger avg. N (437 - 444).
- Middling scenarios show modest power improvements (53 - 59%), but SSR often requires maximum sample size (600 patients in 33% of trials).
- Suggests SSR helps address power deficits for smaller treatment effects but may demand substantial increases in sample size.

Key Takeaways

- **Benefits of Adaptive SSR**: Improves trial efficiency, cost savings, flexibility, and statistical validity.
- **Application Challenges**: Common for simple endpoints (continuous, binary, survival) but less used for complex endpoints.
- **Proposed Approach**: Uses an Adaptive design with SSR for a CV trial with F-S statistic with multiple hierarchical endpoints.
- **Clinical Impact**: Enhances adaptive design applicability, particularly in cardiovascular trials using the F-S statistic.
- **Real World Application**: This is a case study inspired by (but does not show numbers and/or the exact design) a CV project on which the authors consulted with a sponsor on. Quantification of treatment effect is a simple extension using Win Ratio.
  - The trial read out in 2023-24 and was positive for the final analysis.
  - IA SSR decision was: No change to SS, since z-statistic was in the favorable zone.

Full Paper on arXiv

Cornell University

We gratefully acknowledge support from the Simons Foundation, m
institutions, and all contrib

arXiv > stat > arXiv:2504.14748

Search... All fields

Help | Advanced Search

Statistics > Methodology

[Submitted on 20 Apr 2025]

Statistical Design and Planning of an Adaptive Trial using Hierarchical Composite Outcomes: A Practical example

Krishna Padmanabhan, Cyrus Mehta

Hierarchical composite endpoints, such as those analyzed using the Finkelstein-Schoenfeld (FS) statistic, are increasingly used in clinical trials for their ability to incorporate clinically prioritized outcomes. However, adaptive design methods for these endpoints remain underdeveloped. This paper presents a practical framework for implementing sample size re-estimation (SSR) in trials using hierarchical composites, motivated by a cardiovascular trial with mortality, hospitalization, and a functional response as prioritized endpoints. We use a two-stage adaptive design with a single interim analysis for illustration. The interim analysis incorporates predictive probabilities to determine whether the trial should stop for futility, continue as planned, or increase the sample size to maintain power. The decision framework is based on predefined zones for predictive probability, with corresponding adjustments to the stage 2 sample size. Simulation studies across various treatment scenarios demonstrate strong type I error control and increased power compared to a fixed design, particularly for treatment effects that are clinically relevant but lower than the alternative hypothesis. We also explore an alternative conditional power approach for SSR, offering further sample size optimization. Our results support the use of SSR with hierarchical composite outcomes using an FS statistic, enhancing trial efficiency.

Access P

View PDF
HTML (expe
TeX Source
Other Forma

(cc) BY-NC-SA view

Current browse co
stat.ME
< prev | nex
new | recent | 202

Change to browse
stat

References &

NASA ADS
Google Schola
Semantic Scho

Export BibTeX C

Bookmark

https://arxiv.org/abs/2504.14748

Key References

(Not exhaustive)

1) Ajufo E, Nayak A, Mehra MR. Fallacies of Using the Win Ratio in Cardiovascular Trials: Challenges and Solutions. JACC Basic Transl Sci. 2023 Jun 26;8(6):720-727.

2) McCoy CE. Understanding the Use of Composite Endpoints in Clinical Trials. West J Emerg Med. 2018 Jul;19(4):631-634. doi: 10.5811/westjem.2018.4.38383. Epub 2018 Jun 4. PMID: 30013696; PMCID: PMC6040910.

3) Finkelstein D, Schoenfeld D. Combining mortality and longitudinal measures in clinical trials. Stat Med. 1999;18:1341–1354.

4) Marion, J et. al. (2024). Predictive Probabilities Made Simple: A Fast and Accurate Method for Clinical Trial Decision Making. arXiv preprint arXiv:2406.11406.

5) Yu RX, Ganju J. Sample size formula for a win ratio endpoint. Statistics in Medicine. 2022; 41(6): 950–963.

6) Cui L, Hung HJ, Wang S-J. Modification of sample size in group sequential clinical trials. Biometrics. 1999;55:853-857.

7) Gao P, Ware JH, Mehta C. Sample size re-estimation for adaptive sequential design in clinical trials. J Biopharm Stat. 008;18(6):1184-96

8) Rogers et. al. Analysis of recurrent events with an associated informative dropout time: Application of the joint frailty model. Stat Med. 2016 Jan 10;35(13):2195–2205.

9) Schoenfeld DA, Ramchandani R, Finkelstein DM. Designing a longitudinal clinical trial based on a composite endpoint: Sample size, monitoring, and adaptation. Stat Med. 2022 Oct 30;41(24):4745-4755.

10) Pocock SJ et. al. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J. 2012 Jan; 33(2):176-82.

Cytel