Balancing Power and Interpretation: Responder Analysis vs Mean Difference for PRO Endpoints in oncology clinical trials

MSR44

Hannah Fairbanks, Emma C. Martin, Rachael Lawrance.
Adelphi Values Patient-Centered Outcomes, Bollington, United Kingdom.

Objective

- In the analysis of patient reported outcome (PRO) endpoints in clinical trials, it is common to evaluate the effect of treatment by comparing differences in mean change from baseline (CfB) scores, evaluating if differences between the treatment group means are meaningful using Minimally Important Differences (MIDs).
- Analysis of CfB score, using the underlying continuous nature of the PRO score is a statistically robust approach to evaluate the treatment effect. However, there is concern that differences in mean scores can be difficult to interpret by patients and clinicians.[1]
- Responder analysis, where CfB is dichotomised using meaningful within-person change (MWPC) thresholds, is recommended as being more intuitive for interpretation (or a useful aid to interpretation).[1,2] However, in most situations the analysis of dichotomised data, rather than in its continuous form, results in a reduction in power, a key criticism of the responder analysis approach.[3,4]
- Our primary aim was to explore the expected reduction in power, comparing the power of a continuous CfB endpoint to that of a binary responder endpoint, in a phase III oncology PRO setting.

Driving Rationale

In a 'typical' PRO scenario - is the loss of power enough to discount a responder analysis, given the results are more interpretable for a wider audience?

Defining a typical PRO endpoint

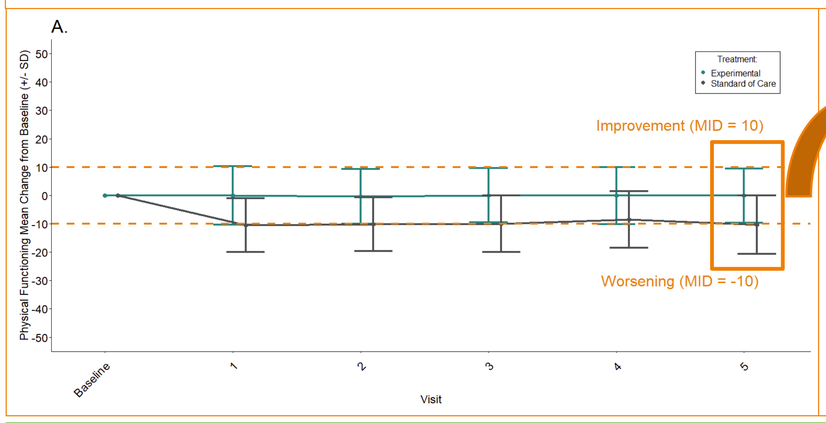
- We based this on a confirmatory, parallel group, phase III clinical trial scenario, comparing an experimental treatment (TRT) against standard of care (SOC) for superiority.
- The focus is on differences in treatment mean scores that are at least 'Small' (or non-trivial), and therefore meaningful to patients. The choice of the size of between-group difference in means will directly influence the power of the CfB continuous analysis, as well as the difference in the proportion of responders between the treatment groups, and therefore the power of the responder analysis.
- Our selected PRO subscale of interest was EORTC QLQ-C30 Physical Functioning (PF).
  - PF is a domain score comprised of 5 items, range: 0 – 100, higher values = higher level of functioning. It is not truly continuous in nature – scores can only change in increments of 6.67.
  - *Continuous, between-group:* Cocks et al. (2011) published between-group differences in treatment mean scores of 5-14 points as 'Small' (0-5 = 'Trivial', 14-22 = 'Medium', >22 'Large'). A single score 10-point MID for between-group comparison is also commonly applied.[6,7]
  - *Dichotomised, within-patient:* The MWPC threshold ≥ -6.67 points was used as the responder threshold for deterioration based on the smallest possible change, ≥ -13.34 (smallest possible change +1, Cocks & Buchanon, 2023) was also used to align with a corresponding sensitivity analysis.

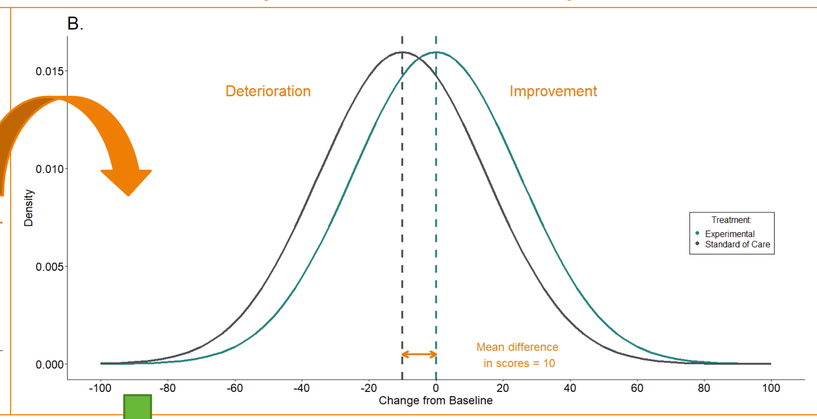
Methods

- We assumed CfB scores were normally distributed, around a mean of 0 (no change) in the TRT group with all treatment effect evidenced as a deterioration in the SOC group (negative CfB mean). T-test methodology was utilised to conduct power and sample size calculations for this continuous endpoint.
- Using the CfB score distributions we applied the MWPC threshold (responder definition) to obtain the proportion of patients who deteriorated vs improved/remained stable. The proportion of responders in each treatment group was then used to perform power and sample size calculations for the binary endpoint.
- The above was repeated for a range of SOC CfB means representing 'Small' differences in treatment means, two MWPC thresholds, and a variety of sample sizes, the significance level of was fixed at 0.05 (5%).
- Having performed the power and sample size calculations, for both the continuous and responder endpoints, we visually explored the relationships.

Illustration: Continuous endpoint to responder analysis

Group level - Continuous endpoint





Example utilising a difference in treatment means of 10-points

A. Typical descriptive presentation of group level **continuous** summary of PF (e.g. from a Mixed Model for Repeated Measures). TRT experiencing stability with a mean CfB of 0, SOC a deterioration in the mean of 10-points from the first post-baseline visit which remains stable thereafter.

B. Shows the corresponding theoretical **probability density function** (pdf) curves for each treatment for a **single post-baseline** timepoint, the difference in means being 10-points.

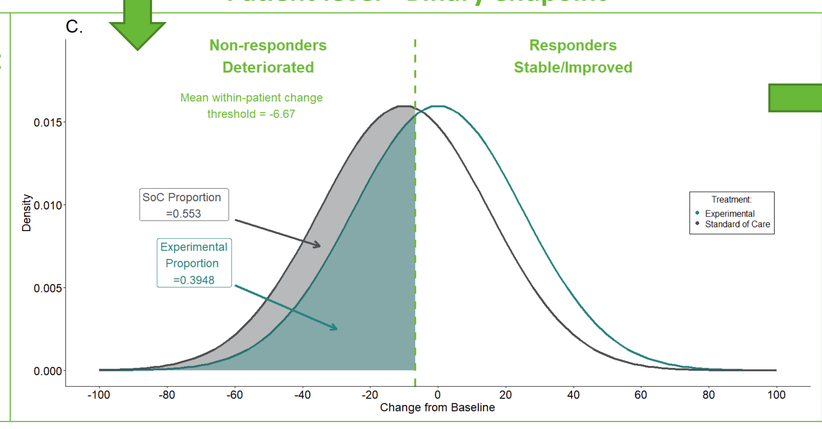
The treatment effect observed in the continuous data directly influences the difference in proportions on the dichotomised scale.

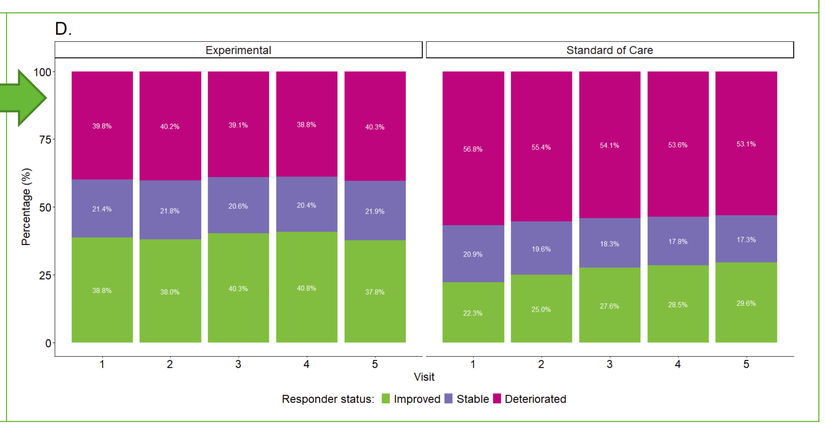
Patient level - Binary endpoint

Example utilising a difference in treatment means of 10-points and MWPC of -6.67 (deterioration)

C. Same pdf curves as shown in **B**, with the inclusion of the MWPC responder threshold of -6.67 (minimal one level deterioration) to dichotomise the continuous endpoint, with the proportions of patients in each treatment group reaching this threshold.

D. Typical responder analysis presentation, although responder analysis is a binary endpoint, for PROs in particular, descriptive summaries often present the **trichotomisation** with the inclusion of a 'stable' category.





Results/Discussion

- Figure 1 shows, as expected, the power associated with the responder endpoint is lower than that of the CfB continuous endpoint, but even for small differences between treatment means the power of the responder analysis shows this is a statistically viable endpoint, with the benefit of interpretability.
- In studies with over 200 patients per treatment group (400 total), the power of both the responder and continuous CfB analysis was over 80% for a 'Small' difference in treatment means (≥ 10 points) and a one increment MWPC responder definition threshold; 300 patients per treatment group would be required for power ≥ 95% for both analyses.
- The expected power is an important consideration when selecting secondary PRO outcomes, we recommend looking at both CfB and responder rates as part of the exploratory outcomes in a study, as considering both gives a more holistic picture of patients' experiences.

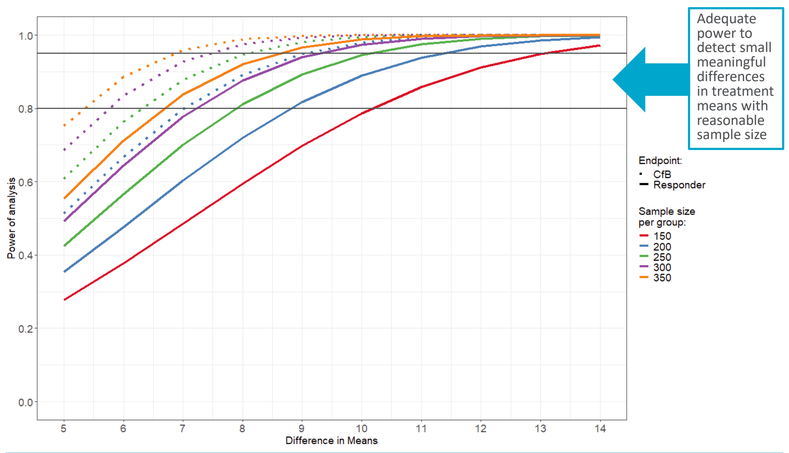


Figure 1: Power of EORTC QLQ-C30 PF CfB and corresponding responder analysis, for a range of 'Small' differences in treatment means and sample sizes utilising a MWPC responder definition of -6.67 (one increment deterioration)

Table 1: Summary of responder and continuous CfB analysis powers for EORTC QLQ-C30 PF with 'Small' difference in treatment means and two MWPC responder definition thresholds

Sample size per group	EORTC QLQ-C30 Physical function MWPC (one and two increment change)	Power	
		Difference in treatment means = 10	
		Responder Analysis	Continuous change from baseline analysis
200	6.67	88.9%	97.9%
	13.34	87.7%	
300	6.67	97.4%	99.8%
	13.34	96.9%	
350	6.67	98.8%	99.9%
	13.34	98.5%	

Conclusions

In phase III oncology clinical trials the sample size of the study will be determined based on the primary endpoints (e.g. Progression Free Survival, Overall Survival). Where the study sample size is sufficiently large (e.g. >400 patients), responder analysis for PRO endpoints will have >80% statistical power to detect small meaningful differences between the treatments providing an additional interpretable endpoint to complement CfB analysis.

[1] Tannock, Ian F et al. Importance of responder criteria for reporting health-related quality-of-life data in clinical trials for advanced cancer: recommendations of Common Sense Oncology and the European Organisation for Research and Treatment of Cancer The Lancet Oncology, Volume 26, Issue 9, e499 - e507. [2] HTA CG: Guidance on outcomes for joint clinical assessments. Document Adopted 10 June 2024: https://health.ec.europa.eu/document/download/a70a62c7-325c-401e-ba42-66174b656ab8_en?filename=hta_outcomes_jca_guidance_en.pdf. [3] Snapinn SM, Jiang Q. Responder analyses and the assessment of a clinically relevant treatment effect. Trials. 2007 Oct 25;8:31. doi: 10.1186/1745-6215-8-31. PMID: 17961249; PMCID: PMC2164942. [4] Uryniak, Tom & Chan, Ivan & Fedorov, Valerii & Jiang, Qi & Oppenheimer, Leonard & Snapinn, Steven & Teng, Chi-Hse & Zhang, John. (2011). Responder analyses- A PhRMA position paper. Statistics in Biopharmaceutical Research. 3. 476-487. 10.1198/sbr.2011.10070. [5] Cocks K, King M.T, Velikova G. et al. Evidence-Based Guidelines for Determination of Sample Size and Interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. J Clin Oncol 2011; 29(1) 89-96. [6] Musoro J.Z, Coens C, Sprangers A.G et al. Minimally important differences for interpreting EORTC QLQ-C30 change scores over time: A synthesis across 21 clinical trials involving nine different cancer types. European Journal of Cancer. 2023; 188: 171-182. [7] Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16(1):139–44. [8] Cocks K, Buchanan J. How scoring limits the usability of minimal important differences (MIDs) as responder definition (RD): an exemplary demonstration using EORTC QLQ-C30 subscales. Quality of Life Research. 2023;32(5):1247-1253.