Separating a treatment effect's **precision** from its **magnitude** provides a different, more **clinically relevant**, way to stop trials at interim.

Considering 'Non-Promising' Treatment Effects at Interim Analyses: Futility of the Treatment, Rather than Futility of the Trial.

Authors: Matthew J Parkes[1 2 3], Mark Lunt[1], Philip Pallmann[4], David T Felson[3 5]

Affiliations:

1. Oxford Clinical Trials Research Unit (OCTRU), Nuffield Department of Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.
2. Centre for Biostatistics, Division of Population Health, Health Services Research & Primary Care, University of Manchester, UK.
3. NIHR Manchester Musculoskeletal Biomedical Research Unit, Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK.
4. Centre for Trials Research, Cardiff University, Cardiff, UK.
5. Section of Rheumatology, Boston University School of Medicine, Boston, MA, USA.

Introduction

Futility analyses are used to a allow a trial showing small treatment effects to **stop early**.
Current methods of assessing futility focus on whether the trial's **final analysis** will likely demonstrate a **statistically significant** effect.
We posit that this is an assessment of whether the **trial,** rather than the **treatment, is futile**.
Current methods may also allow trials which have **excluded clinically meaningful effects** at interim **to continue, due to the use of test statistics** in their derivation.

We propose an alternative stopping rule that stops trials when the interim estimate excludes treatment effects deemed potentially clinically useful, considering the treatment under assessment therefore '**non-promising**'.

We contend that this approach has more desirable operating characteristics:

- It selects for treatments which may show **clinically meaningful treatment effects directly**.
- This results in either trials that stop at interim with useful interval estimates, or they continue to final analysis.

Methods

- Simulation study.
- 8 different scenarios: fictional parallel groups clinical, trial active treatment vs placebo
    - 4 null at final analysis
    - 4 with clinically meaningful treatment effects at final analysis
- One interim analysis at 1/3 recruitment
- Compare 7 interim futility analysis methods:
    1. Group sequential, O'Brien-Fleming stopping behaviour
    2. Group sequential Pocock stopping behaviour
    3. Conditional power approach
    4. Frequentist 'non-promising stopping', O'Brien-Fleming-based simultaneous confidence intervals
    5. Frequentist 'non-promising stopping', Pocock-based simultaneous confidence intervals
    6. Bayesian approach using Region of Practical Equivalence (ROPE)
    7. Bayesian implementation of the 'non-promising region' approach
- Ran 250 iterations of each approach/scenario
- Compared the number of trial iterations stopped at interim, and mean interval estimate at interim analysis.

Results

Table 1: Summary of Treatment Estimates from Stopped Interim Trial Iterations

Approach	Scenario	Interim Sample Size	Number of Iterations Stopped at Interim (out of 250)	Mean Test Statistic	Mean Point Estimate of Between-Groups Difference	Mean Lower Bound of Estimate	Mean Upper Bound of Estimate	Mean Width of Estimate Interval
GSD, O'Brien-Fleming behaviour	A	98	2	-0.14	-0.59	-9.17	7.99	17.15
	B		139	0.61	2.8	-6.39	11.99	18.38
	C		89	0.39	2.14	-8.84	13.12	21.96
	D		146	0.6	3.26	-7.69	14.21	21.9
	F		114	0.48	1.29	-4.13	6.71	10.84
	G		20	0.21	0.59	-4.9	6.07	10.97
	H		118	0.5	2.8	-8.36	13.95	22.3
GSD, Pocock behaviour	B	108	82	0.98	4.32	-4.41	13.06	17.47
	C		35	0.97	5.05	-5.38	15.48	20.86
	D		84	1	5.22	-5.19	15.63	20.81
	F		58	0.96	2.49	-2.67	7.64	10.31
	G		7	0.7	1.81	-3.34	6.96	10.3
	H		63	0.92	4.93	-5.65	15.52	21.17
Conditional power	B	96	97	0.43	4.06	-5.17	13.3	18.46
	C		44	0.44	4.54	-6.62	15.7	22.31
	D		97	0.42	5.06	-5.98	16.09	22.07
	F		71	0.45	2.19	-3.28	7.65	10.93
	G		9	0.52	1.42	-4.13	6.97	11.09
	H		70	0.42	5.01	-6.26	16.28	22.55
Non-promising stopping, O'Brien-Fleming simultaneous CIs	B	96	5	2.02	8.76	-7.94	25.47	33.41
	D		4	2.42	13.27	-7.93	34.47	42.4
	F		39	1.07	2.86	-7.45	13.17	20.62
	G		3	0.76	2.04	-8.14	12.21	20.36
	H		1	2.62	15.53	-7.41	38.48	45.89
Non-promising stopping, Pocock-based simultaneous CIs	B	106	87	0.94	4.16	-6.08	14.4	20.48
	C		25	1.19	6.24	-5.92	18.4	24.32
	D		58	1.25	6.54	-5.55	18.63	24.18
	F		185	0.05	0.14	-5.97	6.24	12.21
	G		66	-0.53	-1.35	-7.34	4.65	11.99
	H		42	1.14	6.1	-6.27	18.47	24.74
Bayesian ROPE	A	96	1	0.98	-0.45	-7.58	7.87	15.44
	B		10	0.96	0.53	-7.88	8.73	16.6
	F		206	0.99	-0.6	-6.15	4.89	11.04
	G		116	0.98	-2.24	-7.76	3.22	10.98
Bayesian Non-promising stopping	A	96	2	0.96	-0.24	-7.86	9.15	17.01
	B		104	0.82	3.88	-5.43	13.23	18.66
	C		35	0.71	5.31	-6.02	16.59	22.61
	D		74	0.66	6.16	-4.79	17.31	22.1
	F		201	0.99	-0.15	-5.69	5.33	11.02
	G		86	0.99	-1.72	-7.15	3.72	10.87
	H		58	0.68	5.73	-5.62	17.14	22.76

GSD = Group Sequential Design; CI = Confidence Interval; ROPE = Region of Practical Equivalence
Grey-highlighted rows indicate trials ideally should stop at interim.

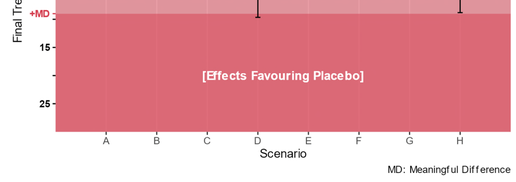
Additional Details

Test statistics

Group sequential designs: $Z = \frac{\Delta}{SE_{\Delta}}$

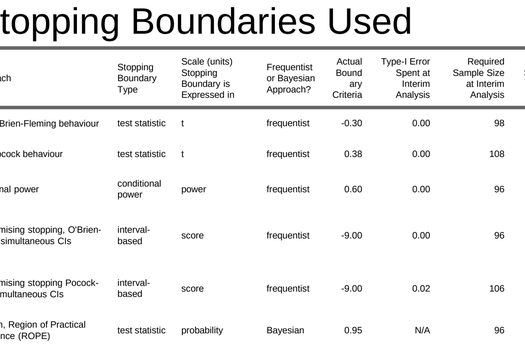
Conditional Power: $1 - \Phi[(Z_{\alpha/2} - E[B_{(1)}|B_{(\tau)}])/\sqrt{1-\tau}])$

'Non-promising approach'



MD: Meaningful Difference

Comparison of stopping rules



Group sequential stopping boundary (not plotted): Z = 0.25
Non-promising stopping boundary (MID): Δ = 0.4
y-axis relates to the point estimate, which is distinct from the Z-statistic

Simulation study scenarios

MD: Meaningful Difference

Stopping Boundaries Used

Approach	Stopping Boundary Type	Scale (units) Stopping Boundary is Expressed in	Frequentist or Bayesian Approach?	Actual Boundary Criteria	Type-I Error Spent at Interim Analysis	Required Sample Size at Interim Analysis	Required Sample Size at Final Analysis
GSD, O'Brien-Fleming behaviour	test statistic	t	frequentist	-0.30	0.00	98	294
GSD, Pocock behaviour	test statistic	t	frequentist	0.38	0.00	108	324
Conditional power	conditional power	power	frequentist	0.60	0.00	96	288
Non-promising stopping, O'Brien-Fleming simultaneous CIs	interval-based	score	frequentist	-9.00	0.00	96	288
Non-promising stopping Pocock-based simultaneous CIs	interval-based	score	frequentist	-9.00	0.02	106	317
Bayesian, Region of Practical Equivalence (ROPE)	test statistic	probability	Bayesian	0.95	N/A	96	288
Bayesian Non-promising stopping	interval-based	score	Bayesian	-9.00	N/A	96	288

Please scan the QR code for a .pdf version of this poster, plus the original abstract submission, or visit https://

Poster design based on Mike Morrison's 'Better Posters' design philosophy. See https://osf.io/ef53g/

Contact: matthew.parkes@ndorms.ox.ac.uk