

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Frequentist and Bayesian Sequential
Experiments for Pandemic Disrupted
Clinical Trials: Can Novel Methodologies

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Novelli, M., Forster, M., Welch, C. (2025). Frequentist and Bayesian Sequential Experiments for Pandemic Disrupted Clinical Trials: Can Novel Methodologies Deliver Greater Value for Health Care Systems?. Cham : Springer [10.1007/978-3-031-96736-8_4].

Availability:

This version is available at: <https://hdl.handle.net/11585/1017923> since: 2025-06-18

Published:

DOI: http://doi.org/10.1007/978-3-031-96736-8_4

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

Frequentist and Bayesian Sequential Experiments for Pandemic Disrupted Clinical Trials: Can Novel Methodologies Deliver Greater Value for Health Care Systems?

Marco Novelli¹, Martin Forster¹, and Charlie Welch²

¹ University of Bologna, Bologna, Italy

m.novelli@unibo.it

<https://www.unibo.it/sitoweb/m.novelli/en>

² University of York, Department of Health Sciences, UK

Abstract. We apply innovations in frequentist and Bayesian sequential experimental design to clinical and cost-effectiveness data from the DISC trial, a UK National Institute for Health and Care Research-funded clinical trial that was disrupted during the COVID-19 pandemic. Analysis of the cost-effectiveness data suggests that stopping recruitment at the pandemic’s outbreak was indicated by some models, but not others. Analysis of the clinical effectiveness data suggests that stopping recruitment was never indicated. Reflecting recent contributions [5, 6], we illustrate how trial management decisions during a pandemic might benefit from insights provided by a range of statistical and economic analyses of the same experimental data.

Keywords: Group sequential experiments; Bayesian decision theory; Health economics; Pandemic preparedness

1 Introduction

There is growing interest in how adaptive experimental designs might be used to handle so-called ‘disrupted’ clinical trials, in which extenuating circumstances such as a pandemic force modifications to the trial’s original protocol [7]. [5] consider how such designs might be used to handle disruption to non-COVID-19 trials during the COVID-19 pandemic. [6] study how to carry out an unplanned interim analysis with an early-stopping or sample size alteration option, as well as how to handle potential changes to the study population. This evolving and innovative literature is accompanied by ongoing developments in the design of sequential clinical trials, from both frequentist [3] and Bayesian [1] perspectives.

In this paper, we apply a number of these innovations to analyse clinical and cost-effectiveness data from the DISC trial, a UK National Institute for Health and Care Research (NIHR)-funded clinical trial that faced substantial disruption during the COVID-19 pandemic. We study a hypothetical scenario in which an unplanned interim analysis was carried out when the NIHR paused recruitment

to the trial in March 2020, comparing frequentist group sequential (GS) designs and a Bayesian decision-theoretic design. We ask whether, given the accumulated clinical and cost-effectiveness evidence at the time the trial paused, recruitment to the trial should have stopped. Analysis of the cost-effectiveness data shows that stopping recruitment was indicated by some models, but not others. In contrast, analysis of the clinical effectiveness data suggests that stopping recruitment was never indicated. Our aim is not to draw an overall conclusion about whether stopping recruitment was the best decision (this is a matter that is perspective-dependent). Rather, in line with recent research [5, 6], we illustrate how trial management during a pandemic might benefit from a range of statistical and economic analyses of the same experimental data.

The DISC trial was a two-armed, fixed sample size randomised clinical trial. It investigated whether collagenase injection and manipulation of a hand joint affected by Dupuytren’s contracture was not inferior to limited fasciectomy surgery in adults. The primary clinical effectiveness endpoint was the Patient Evaluation Measure (PEM) Hand Health Questionnaire score at 12 months post-treatment. The trial also considered the cost-effectiveness of the treatments. Recruitment started in July 2017 and paused on 20th March 2020 as a result of the COVID-19 pandemic, by which time 482 patients had been treated, 162 of whom had one-year outcomes observed. Recruitment and treatment delivery restarted in June 2020. Full details are provided in [2].

2 Methods

We consider a two-armed randomised clinical trial comparing a new health technology, N (collagenase), with an existing/standard technology, S (surgery). There is balanced randomisation to the arms and delay in observing the clinical and cost-effectiveness outcomes. We consider the hypothetical scenario in which a single interim analysis is carried out at the point at which recruitment to the trial paused as a result of the COVID-19 pandemic. The frequentist models [3, 4, 8] are applied to both clinical and cost-effectiveness outcomes. The Bayesian decision-theoretic model [1] is applied to the cost-effectiveness outcome. The models are designed to handle the presence of ‘pipeline patients’, patients who have been treated, but whose outcomes have yet to be observed.

2.1 Frequentist Group Sequential Designs

Patient outcomes Y_j , $j \in \{N, S\}$, are assumed to have normal distributions with a common, known variance, σ_Y^2 , and unknown expected values, μ_j . Define $\theta = \mu_N - \mu_S$ as the treatment effect of interest. The designs test $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$, with significance level α and a power $1 - \beta$ to detect a predefined clinically relevant difference, δ , where β is the type II error probability. For the analysis of cost-effectiveness, interest is in testing the equivalence of the expected values of the net monetary benefits of the two treatments, $NB_j = \lambda E_j - C_j$, where random variables E and C are, respectively, effectiveness (the Quality Adjusted

Life Year (QALY) in the DISC trial) and treatment cost. λ is the maximum willingness to pay for a unit of E .

[3] extend the classical GS model of a clinical trial to account for delay in observing the primary outcome. With a single interim analysis, the trial is divided into two stages: during *stage 1*, patients are recruited, randomised, treated and their outcomes are followed up, with an interim analysis taking place after $\tilde{n}_1$ patients per arm have been treated (n_1 of whom will have had their outcomes observed, the balance being ‘in the pipeline’); during *stage 2*, recruitment continues to the trial’s planned sample size, n_2 . At the interim analysis, two actions are available: 1. recruitment to the trial terminates, the pipeline patients are followed up and H_0 is tested using the outcomes for only $\tilde{n}_1$ patients; 2. recruitment continues to n_2 , all patients are followed up and H_0 is tested.

The decision at the interim analysis is informed by comparing the standard normal test statistic, $Z_1 = \sqrt{n_1}\hat{\theta}_1/\sqrt{2\sigma_Y^2}$ (we use ‘ $\hat{\cdot}$ ’ to denote a MLE of θ) with the lower and upper thresholds of a ‘continuation region’, (l_1, u_1) . If $Z_1 \in (l_1, u_1)$, the trial moves to stage 2 and H_0 is tested by comparing $Z_2 = \sqrt{n_2}\hat{\theta}_2/\sqrt{2\sigma_Y^2}$ with a critical value, c_2 . Otherwise, recruitment stops, all $\tilde{n}_1$ patients recruited during stage 1 are followed-up and H_0 is tested by comparing $\tilde{Z}_1 = \sqrt{\tilde{n}_1}\tilde{\theta}_1/\sqrt{2\sigma_Y^2}$ with the stage 1 critical value, $\tilde{c}_1$. Values for l_1, u_1 and c_2 are obtained by defining how errors α and β should be ‘spent’ across stages 1 and 2. We adopt the Lan and DeMets spending function approach, using the ρ -family spending function (we assume $\rho = 2$), together with one that mimics the original O’Brien-Fleming approach (‘OBF-like’) [8, 3, section 4.1].

The presence of the critical value $\tilde{c}_1$ represents a novelty for GS designs with delayed responses, because it is used only in the event that recruitment terminates at the end of stage 1; with no delayed responses, $\tilde{c}_1$ does not exist. [3] solve their model with a ‘binding’ futility boundary, in which recruitment to the trial stops at the interim analysis in the event that $Z_1 < l_1$. They obtain $\tilde{c}_1$ by balancing ‘reversal probabilities’ under the null hypothesis, choosing its value to equate the probability of rejecting H_0 after stopping recruitment for futility ($Z_1 < l_1$) with the probability of failing to reject it having stopped recruitment following promising interim results ($Z_1 > u_1$). In contrast, [4] employ a non-binding futility boundary, setting $\tilde{c}_1 = \Phi^{-1}(1 - \alpha)$. Finally, [8] allow recruitment to restart having observed outcomes for all $\tilde{n}_1$ patients treated during stage 1. Choice of $\tilde{c}_1$ exploits the notion of ‘conditional power’, pre-specifying the probabilities of achieving efficacy after pausing the trial for promising results and of concluding futility after observing non-promising results at the interim analysis. The option to restart recruitment using the $\tilde{n}_1$ patients results in a second continuation region, $(\tilde{l}_1, \tilde{c}_1)$, such that $\tilde{c}_1$ tests for the superiority of N and $\tilde{l}_1$ is used to decide whether to terminate the trial for futility. If $\tilde{Z}_1$ lies between $\tilde{l}_1$ and u_1 , recruitment should restart.

2.2 Bayesian Decision-Theoretic Sequential Design

In the Bayesian decision-theoretic sequential design of [1], the measurement of cost-effectiveness is Incremental Net Monetary Benefit, $\text{INMB} = \text{NB}_N - \text{NB}_S$.

At the start of the trial, a normal prior probability distribution for $\mathbb{E}[\text{INMB}]$ is specified and the INMB data are assumed to arrive as pairwise, i.i.d. draws from a normal distribution with unknown mean, θ_{INMB} , and known variance, σ_{INMB}^2 . Bayes' rule is used to update the prior distribution sequentially, yielding a sequence of normal posterior distributions for $\mathbb{E}[\text{INMB}]$.

Given the sequential arrival of data and a defined marginal cost of making a single pairwise allocation, c , the objective in the model is to choose a sample size at which to stop recruitment so as to maximise the expected net benefit of the trial and its technology adoption decision, assuming that a treatment population of P patients will benefit from the trial's results. Accounting for the defined delay in observing outcomes, an optimal policy comprises a continuous 'stopping boundary', defined as a function of the number of patients treated. Once observations on patients start arriving, for any given sample size, the posterior mean for θ_{INMB} is compared with the lower and upper limits of the stopping boundary to inform a decision about whether recruitment should stop or whether another pairwise allocation should be made. We consider this comparison for the sample size recruited to the trial on 20th March 2020, when recruitment paused. [1] provides a full statement of the model and its solution.

3 Results and Discussion

The DISC trial planned to obtain an effective sample size of $n_2 = 284$ patients to each arm. When recruitment paused on 20th March 2020, outcomes for approximately 28% of the target sample size had been observed and about 42% of the target sample size was in the pipeline, $\tilde{n}_1 = 201$. The rate of treatment delivery, and subsequent follow-up was not balanced (91 (collagenase) vs 71 (surgery)), so we use the average, $n_1 = 81$, for the number of observations at the interim analysis. For all three frequentist models, we set $\alpha = 0.025$ and $\beta = 0.1$. For [8], the probabilities of achieving efficacy after pausing the trial for promising results and the probability of concluding futility after observing non-promising results are 0.9, with the futility threshold for the interim decision set to 0. For the Bayesian model, we used appropriate sources to set $c = \text{£}2000$, $P = 65000$, $\sigma_{\text{INMB}} = \text{£}6395$, the prior mean equal to 0 and the prior standard deviation equal to $\text{£}4522$ (which fixes the effective sample size of the prior to approximately 2). *Frequentist GS results:* **Table 1** presents the outputs of the frequentist GS models. The model outputs comprise the continuation region thresholds (l_1, u_1) (all models) and $(\tilde{l}_1, \tilde{u}_1)$ (SKK24 only), together with the critical values for the hypothesis tests. The futility thresholds for [4] (HJ22) are highlighted, to indicate that they are non-binding. **Table 2** presents the values of the test statistics $Z_1, \tilde{Z}_1$ and Z_2 , together with the decisions that apply at each point in **Table 1**, for the clinical (upper panel) and cost effectiveness (lower panel) analysis. For clinical effectiveness, results are consistent across designs: the recommendation at the hypothetical interim analysis of 20th March 2020 would have been to continue recruitment, because Z_1 lies within the continuation regions. So, had any of these models been deployed in such an interim analysis, the DISC trial

Table 1: Outputs of the frequentist GS models: upper and lower boundaries at the interim analysis of 20th March 2020, together with the critical values for the hypothesis tests: cost-effectiveness analysis. (Notes: HJ13 is [3], JH22 is [4] and SKK24 is [8]).

Spending function	Design	Interim analysis (20/3/20) (l_1, u_1)	If recruitment at the interim analysis:			
			stops (20/03/21) $\tilde{c}_1$	continues (Final date T) ($\tilde{l}_1, \tilde{u}_1$) c_2	restarts ($T+1$ year) c_2	
OBF-like	HJ13	(-1.181, 4.07)	2.1	-	1.964	-
	JH22	(-1.24, 4.023)	1.96	-	1.96	-
	SKK24	(-1.574, 1.873)	-	(0, 1.972)	2.028	2.028
$\rho = 2$	HJ13	(-0.68, 2.882)	1.526	-	1.956	-
	JH22	(-0.807, 2.616)	1.96	-	1.97	-
	SKK24	(-1.574, 1.742)	-	(0, 1.765)	2.116	2.116

Table 2: Test statistics and decisions. (Note: the decisions are obtained by comparing the relevant test statistics, Z , in the first row of each panel, with the appropriate boundary/critical value in **Table 1**. For the clinical effectiveness comparison, the relevant boundary/critical value is multiplied by -1 to reflect the fact that a lower PEM score indicates a better outcome.)

Spending function		Recruitment decision at the interim analysis (20/3/20)	Test of H_0 if recruitment at the interim analysis:		
	Design		stops (20/03/21) $\tilde{Z}_1$	continues (Final date T) Z_2	restarts ($T+1$ year)
<i>Clinical effectiveness</i>		$Z_1 = 0.488$	$\tilde{Z}_1 = -0.697$	$Z_2 = -1.015$	
OBF-like	HJ13	Continue	-	Do not reject	-
	JH22	Continue	-	Do not reject	-
	SKK24	Continue	-	Do not reject	-
$\rho = 2$	HJ13	Continue	-	Do not reject	-
	JH22	Continue	-	Do not reject	-
	SKK24	Continue	-	Do not reject	-
<i>Cost effectiveness</i>		$Z_1 = 2.988$	$\tilde{Z}_1 = 4.826$	$Z_2 = 5.824$	
OBF-like	HJ13	Continue	-	Reject	-
	JH22	Continue	-	Reject	-
	SKK24	Stop	Reject	-	-
$\rho = 2$	HJ13	Stop	Reject	-	-
	JH22	Stop	Reject	-	-
	SKK24	Stop	Reject	-	-

166 should have moved to stage 2 as soon as possible. The hypothesis tests at the
 167 target sample size all reach the same conclusion, which is not to reject the null
 168 hypothesis that collagenase is inferior to surgery, a result which is consistent
 169 with the actual result of the DISC trial.

170 For the cost-effectiveness outcome, the OBF-like spending function designs
 171 [3] (HJ13) and [4] (JH22) are consistent with the clinical effectiveness results at
 172 the interim analysis, in that they recommend moving to stage 2. However, the
 173 null hypothesis of no difference between the expected net benefit of the two treat-

ments is rejected in favour of collagenase, as a result of trial data suggesting that collagenase is cheaper than surgery, albeit less effective. In contrast, SKK24 [8] recommends stopping recruitment, as do all of the $\rho = 2$ designs. All six designs ultimately reject the null of no difference between the expected net benefits, in favour of collagenase. The difference lies in the timing of the decisions: for HJ13 and JH22 (OBF-like), the decision occurs at the planned sample size of n_2 , in July 2022; for HJ13 and JH22 ($\rho = 2$) and SKK24 (both versions), the decision is made earlier, in March 2021, albeit with a lower sample size, ($\tilde{n}_1$).

Bayesian Decision-Theoretic Results: Comparison of the posterior mean for $\mathbb{E}[\text{INMB}]$ with the limits of the stopping boundary shows that the Bayesian model recommends not stopping recruitment. The final posterior mean (n_2 patients) indicates that collagenase is cost-effective. This result is consistent with the hypothesis tests for cost-effectiveness for the HJ13 and JH22 (OBF-like) models in the lower panel of **Table 2**, but not the other four.

Conclusion. Our results show how different perspectives to the analysis of a single experimental data set can better inform clinical trial management.

Acknowledgements

MN and MF were supported by EU NextGenerationEU PRIN2022 funds, ‘Optimal and adaptive designs for modern medical experimentation’ (2022TRB44L).

References

- [1] Chick, S. E., Forster, M., and Pertile, P. (2017). A Bayesian decision-theoretic model of sequential experimentation with delayed response. *J R STAT SOC B*, 79(5):1439–1462.
- [2] Dias, J., Tharmanathan, P., Arundel, C., et al. (2024). Collagenase injection versus limited fasciectomy surgery to treat Dupuytren’s contracture in adult patients in the UK: DISC, a non-inferiority RCT and economic evaluation. *HEALTH TECHNOL ASSESS*, 28(78).
- [3] Hampson, L. and Jennison, C. (2013). Group sequential tests for delayed responses. *J R STAT SOC B*, 75:3–54.
- [4] Jennison, C. and Hampson, L. (2022). Analysing over-run data after a group sequential trial. In *Adaptive Designs and Multiple Testing Procedures Workshop. Heidelberg: Joint Working Group of the ROeS and the German Region of the IBS*.
- [5] Kunz, C., Jörgens, S., Bretz, F., et al. (2020). Clinical trials impacted by the COVID-19 pandemic: adaptive designs to the rescue? *STAT BIOPHARM RES*, 12(4):461–477.
- [6] Kunz, C., Tarima, S., Rosner, G., et al. (2024). Bayesian and frequentist approaches to rescuing disrupted trials: A report from the NISS Ingram Olkin forum series on unplanned clinical trial disruptions. *STAT BIOPHARM RES*, 16(4):383–402.
- [7] Orkin, A. M., Gill, P. J., Ghera, D., et al. (2021). Guidelines for reporting trial protocols and completed trials modified due to the COVID-19 pandemic and other extenuating circumstances: The CONSERVE 2021 statement. *JAMA*, 326:257–265.
- [8] Schüürhuis, S., Konietzke, F., and Kunz, C. U. (2024). A two-stage group-sequential design for delayed treatment responses with the possibility of trial restart. *STAT MED*, 43(12):2368–2388.