	Protocol Design Appendix	Sponsor: Imperial College London	V1.0
--	--------------------------	----------------------------------	------



Protocol Design Appendix


PANTHER:-

Precision medicine Adaptive Network platform Trial in Hypoxaemic acutE respiratory failuRe

Protocol design appendix Version V1.0 dated 19 FEB 2025

Property of Imperial Clinical Trials Unit (ICTU)

May not be used, divulged or published without the consent of ICTU

	Protocol Design Appendix	Sponsor: Imperial College London	V1.0
--	--------------------------	----------------------------------	------

Authors	Dr Ed Waddingham Dr Rachel Phillips Prof Victoria Cornelius
----------------	---

Table of Contents

1	BACKGROUND	4
2	TRIAL DESIGN	4
2.1	Overall design	4
2.2	Primary outcome	5
2.3	Number of sites and recruitment rate	5
2.4	Treatment eligibility and allocation	5
2.5	Sample size cap	6
2.6	Adaptive analyses	6
2.7	Selection of adaptive analysis schedule	6
2.7.1	Frequency of adaptive analyses	6
2.7.2	Timing of first adaptive analyses	7
3	AIM OF SIMULATIONS	8
4	SIMULATION METHODS	8
4.1	Evidence synthesis for primary outcome	8
4.2	Simulation assumptions	9
4.2.1	Recruitment and participant characteristics	9
4.2.2	Withdrawal/loss to follow-up	9
4.2.3	Treatment effects	9
4.2.4	Scope of platform	10
4.3	Simulation methods	10
4.3.1	Data generating mechanism	10
4.3.2	Platform estimand	10
4.3.3	Analysis model	11
4.3.4	Simulation outputs	11
4.3.5	Performance measures	11
4.3.6	Identification of optimal triggers	11
4.3.7	Estimation and software	12
4.3.8	Number of iterations	13
5	RESULTS	13
5.1	Probability distributions for primary outcome	13
5.2	Optimal statistical triggers	17
5.2.1	Efficacy triggers	17
5.2.2	Futility triggers	19
5.3	Operating characteristics under chosen triggers	21
5.3.1	Main assumptions	21
5.3.2	Sensitivity analyses	27
6	APPENDIX: FREQUENTIST SAMPLE SIZE CALCULATIONS	37
	REFERENCES	39
7	REVISION HISTORY	39

1 BACKGROUND

PANTHER is a precision medicine platform trial in Acute Respiratory Distress Syndrome (ARDS). Clinical observations and post hoc analyses of prior studies have suggested the patient population can be partitioned into two phenotypes in which the disease and treatment interact via different pathways, such that treatments which are efficacious in one phenotype may frequently be inefficacious in the other. These are the “hypoinflammatory” and “hyperinflammatory” subphenotypes, making up approximately 70% and 30% of the population respectively. This phenotyping can now be performed rapidly upon admission to ICU. The trial is designed to test promising new ARDS treatments in each phenotype separately.

The NIHR Efficacy & Mechanism Evaluation Programme is providing an initial 5 years’ funding for the platform, including 4 years of recruitment. This initial phase of the platform will evaluate simvastatin and baricitinib, the first two proposed treatments for investigation. The intention however is for the study to function as a perpetual platform in which regular adaptive analyses will identify any interventions already determined to be effective or futile in either subphenotype and replace them with new interventions identified by the Prioritisation Committee.

This appendix to the trial protocol sets out the design parameters for the trial, describes the methodology used to determine values for those design parameters, including extensive simulations, and summarises the trial’s operating characteristics under various assumed scenarios.

2 TRIAL DESIGN

2.1 OVERALL DESIGN

The trial will use a Bayesian adaptive multi-arm platform design, stratified by subphenotype, with regular adaptive analyses and the ability to drop and add additional treatment arms over time. The initial design will include two active interventions and usual care and is illustrated below. Patients are classed into either the hypoinflammatory or hyperinflammatory subphenotype. Then within each subphenotype, they are randomised to usual care (UC) or simvastatin or baricitinib in the first instance.

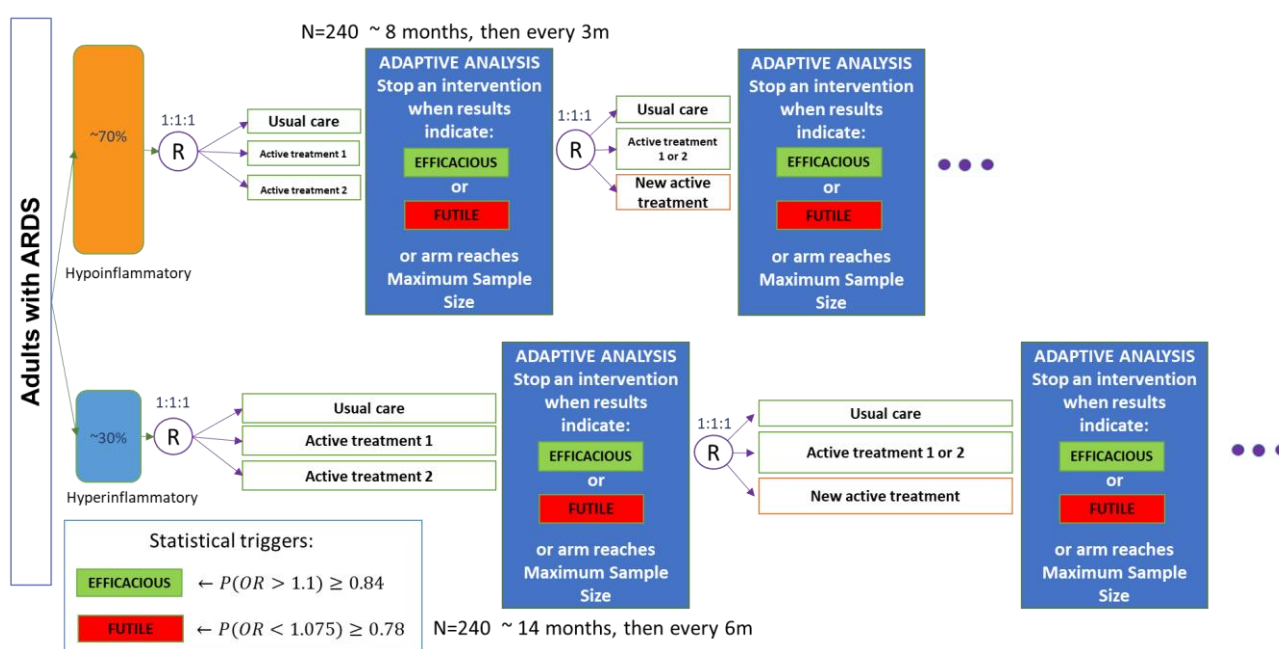


Figure 1 - Overall platform design. At each adaptive (interim) analysis, a treatment arm can be stopped for passing the efficacy/futility boundaries or exceeding the sample size cap within a phenotype. The interim results shown here are examples only.

New interventions will be added as the platform progresses, but initially the active interventions will be simvastatin and baricitinib.

2.2 PRIMARY OUTCOME

The primary outcome is a 30-category ordinal variable combining days free of organ support up to day 28 for survivors (values 0 to 28) with death (value -1). This will be analysed using proportional odds logistic regression, with the effect of each treatment represented as a proportional odds ratio (POR), with $POR > 1$ relative to usual care indicating a favourable treatment.

2.3 NUMBER OF SITES AND RECRUITMENT RATE

The target number of sites is 70 but the platform will open with 30 sites. Assuming that the number of sites increases linearly to 70 sites at 9 months, and allowing for each site to recruit 1 participant per month, we anticipate recruiting 3180 participants over the initial 4 years of the platform, comprising 2226 hypoinflammatory and 954 hyperinflammatory participants based on an assumed 70:30 ratio in the population.

2.4 TREATMENT ELIGIBILITY AND ALLOCATION

Each intervention carries its own particular contraindications and therefore some platform participants will be ineligible for one or other of the active interventions. This is expected to affect a minority of participants. To ensure unbiased comparisons, the comparator group for each active intervention will comprise only of participants randomised to usual care meeting the eligibility requirements for that active intervention.

Treatment allocation will be by minimisation, stratified by subphenotype and eligibility for active intervention. Within each subphenotype and eligibility class, minimisation is carried out by region. For the purpose of allocation concealment, only 90% of allocations (selected at random) will be to the arm that minimises imbalance; the remaining 10% will go to the second best arm. Ties are resolved by 1:1 randomisation.

2.5 SAMPLE SIZE CAP

Within each active intervention in each subphenotype g , a recruitment limit C_g will be imposed so that the sample size does not exceed the maximum sample size in a frequentist group sequential design that broadly parallels the Bayesian design of the study. This is 504 in hypoinflammatory or 529 in hyperinflammatory. For details of how these frequentist sample sizes were calculated, see *Appendix: Frequentist sample size cap*.

2.6 ADAPTIVE ANALYSES

Regular adaptive analyses will assess the accumulated evidence within each arm and subphenotype; if there is sufficient evidence to declare efficacy or futility of an active intervention within a subphenotype then randomisation to that intervention will be stopped within that subphenotype. For each adaptive analysis i , active intervention t and subphenotype g , the stopping rules will be of the form:

- Stop for efficacy (graduate) if $Prob_{gti}(POR > por_{eff}) \geq p_{eff}$
- Stop for futility (reject) if $Prob_{gti}(POR < por_{fut}) \geq p_{fut}$
- Also stop for futility (reject) if $N_{gti} \geq C_g$

where

- N_{gti} is the number of participants in treatment arm t and subphenotype g at adaptive analysis i
- C_g is the per-arm sample size cap in subphenotype g
- por_{eff}, por_{fut} are threshold values for the POR
- p_{eff}, p_{fut} are threshold probability values
- $Prob_{gti}(A)$ is the posterior probability of event/outcome A in treatment arm t and subphenotype g at adaptive analysis i

2.7 SELECTION OF ADAPTIVE ANALYSIS SCHEDULE

2.7.1 Frequency of adaptive analyses

To maximise the potential for early stopping, adaptive analyses are intended to take place as frequently as possible in the larger (hypoinflammatory) subphenotype. This has been judged to be every three months on the basis that it would not be operationally practicable to produce data extracts, conduct/report analyses and schedule DMC meetings more frequently than this. In the hyperinflammatory subphenotype it was decided to schedule adaptive analyses every 6 months, as this would result in analyses taking place at broadly similar recruitment levels in each subphenotype while also aligning with the 3-month cycle in the hypoinflammatory subphenotype. Preliminary simulations confirmed that following this adaptive analysis schedule results in similar power and type I error rates for the same

treatment effect size in both subphenotypes if the same statistical triggers are used in both. To simplify the design and reduce the number of parameters to be optimised, it was therefore decided to adopt the same triggers por_{eff} , por_{fut} , p_{eff} , p_{fut} in both subphenotypes.

2.7.2 Timing of first adaptive analyses

The timing of the first adaptive analysis in each subphenotype was again motivated by a wish to allow stopping as early as possible and hence maximise efficiency of the platform. However this had to be balanced against inflation of type I error rates, which are highest in the early part of a trial when the sample size is still small. In preliminary simulations (results not shown here), if adaptive analyses were to begin any earlier than 80 participants per arm, no combination of statistical thresholds could be found that would yield type I error below 20% and power above 70% to detect the minimum clinically important treatment effect in both subphenotypes.

Combined with the recruitment projections, this has led to the adoption of the following adaptive analysis timing schedule:

- Hypoinflammatory ($g=1$): First adaptive analysis takes place at the end of the calendar month when 240 hypoinflammatory participants are recruited (80 per arm). Further adaptive analyses take place every 3 months thereafter.
- Hyperinflammatory ($g=2$): First adaptive analysis takes place at the same time as the next scheduled hypoinflammatory adaptive analysis after 240 hyperinflammatory participants are recruited (80 per arm). Further adaptive analyses take place every 6 months thereafter.

Based on the recruitment projections set out above, this results in the anticipated timings and N_{gti} values set out in Table 1 for the first 10 adaptive analyses. The simulations described in this document assume this recruitment pattern applies unless otherwise indicated.

Table 1 - Anticipated schedule of interim analyses based on recruitment projections. Each subphenotype is only included in those analyses where its sample size per arm is shown in bold text.

Adaptive analysis number (i)	Month	Hypoinflammatory sample size per treatment arm (N_{1ti})	Hyperinflammatory sample size per treatment arm (N_{2ti})
1	8	89	38
2	11	138	59
3	14	187	80
4	17	236	101
5	20	285	122
6	23	334	143
7	26	383	164
8	29	432	185
9	32	481	206
10	35	529 (capped)	227

3 AIM OF SIMULATIONS

The simulation exercise aims to address the following objectives:

1. To identify optimal values for the statistical triggers por_{eff} , p_{eff} , por_{fut} and p_{fut} .
2. To determine the platform's operating characteristics (power, sample size and time taken to evaluate treatments), using the identified statistical triggers, under various assumed scenarios regarding recruitment / treatment effects.

4 SIMULATION METHODS

4.1 EVIDENCE SYNTHESIS FOR PRIMARY OUTCOME

The distribution of the primary outcome for patients receiving usual care in both subphenotypes is assumed to reflect the distribution of ventilator free days observed in a population of patients with ARDS who were invasively ventilated from the HARP-2 study [1], which was extracted from the original individual patient level study data. The distribution of ventilator free days was similar in a population which included patients receiving both non-invasive respiratory support and invasive ventilation [2]. In the HARP-2 study, a POR of 1.74 was observed in hyperinflammatory patients treated with simvastatin. To model the distribution of the primary outcome for patients on treatment, given an assumed proportional odds ratio OR_{treat} compared to usual care, the following procedure was used:

- i) For each possible level y of the primary outcome Y , the log odds ratio comparing the chance of observing y between treated and untreated hyperinflammatory ($g=2$) patients in the HARP-2 data was calculated as

$$\lambda_y = \text{logit}(\text{Prob}_{g=2}^{HARP2}(Y = y | \text{treated})) - \text{logit}(\text{Prob}_{g=2}^{HARP2}(Y = y | \text{untreated}))$$

- ii) A crude estimate of the log odds (and hence probabilities) of observing the outcome in the hypothetical treatment arm in either subphenotype under the given scenario was obtained by scaling λ_y in proportion with the desired change in the overall proportional odds ratio, i.e.

$$\begin{aligned} & \text{logit}(\text{Prob}_g(Y = y | \text{treated})) \\ &= \text{logit}(\text{Prob}_g^{HARP2}(Y = y | \text{untreated})) + \lambda_y \frac{\log(POR_{treat})}{\log(1.74)} \end{aligned}$$

- iii) The crude probability estimates were normalised to sum to 1 across all possible outcome values y from -1 to 28, providing the target distribution in each subphenotype.

The resulting probability distributions for control subjects and those treated with a minimally effective treatment in each subphenotype are shown in section 5.1.

For sensitivity analyses with an altered baseline mortality rate, the above procedure was applied twice: first to yield a distribution in each subphenotype with the desired baseline

mortality rate; then again to construct distributions for simulation with PORs relative to the new baseline.

4.2 SIMULATION ASSUMPTIONS

4.2.1 Recruitment and participant characteristics

The main assumptions for recruitment are as set out above in section **Error! Reference source not found.** It is also assumed that 10% of participants are eligible for simvastatin only, 10% are eligible for baricitinib only and 80% are eligible for both.

Sensitivity analyses will examine the impact of:

- halving or doubling the assumed recruitment rate at each site (reflecting what might occur if the incidence of ARDS is substantially lower or higher than anticipated);
- increasing the number of participating sites to 100;
- reducing the assumed proportion of hyperinflammatory patients to 25% or increasing it to 35%
- reducing the assumed mortality rate in each subphenotype.

4.2.2 Withdrawal/loss to follow-up

No allowance is made for participants withdrawing from the study or otherwise being lost to follow-up. Minimal withdrawals are expected owing to the nature of the condition being treated (participants will require constant support in ICU). Any anticipated withdrawals would need to be added to recruitment targets to maintain the operating characteristics of the platform.

We further assume that at the end of a given calendar month, primary outcome data will be observed for all participants recruited in that month. In reality it may take up to 28 days to observe each participant's primary outcome, but such a delay will not have a substantial effect on the timings referred to here, shifting them by at most one month.

4.2.3 Treatment effects

The minimal clinically important treatment effect is an OR of 1.4 for hypoinflammatory and 1.3 for hyperinflammatory, corresponding to an absolute reduction of ~5-6% for mortality (from an assumed untreated baseline of 18% in hypoinflammatory and 45% in hyperinflammatory). This was based on clinical consensus among the study team.

Power and type I error statistics are calculated for a range of assumed treatment effects up to $POR=1.5$ and are reported on a per-treatment basis. Although the platform will evaluate multiple treatments in two subphenotypes, no adjustment is made for multiple testing. Adjustment would not be considered appropriate since the active treatments to be evaluated are a priori unrelated to one another and are expected to have heterogeneous effects across the two subphenotypes.

For calculations which evaluate the sample size / time required for evaluation of simvastatin and baricitinib, it is assumed that both treatments have no effect ($POR=1$) in the hypoinflammatory subphenotype and $POR=1.3$ in hyperinflammatory.

4.2.4 Scope of platform

The study is anticipated to function as a perpetual platform trial (contingent on funding). Simulations assume perpetual 1:1:1 allocation to two interventions and control with no upper limit on the overall sample size (although individual arms will be capped as per section 2.5). However, operating characteristics relating to the initial 4-year funded period will be among the results presented and considered when selecting design parameters.

4.3 SIMULATION METHODS

4.3.1 Data generating mechanism

Within each iteration of the simulations j , for each subphenotype $g=1,2$ and treatment arm $t=0,1,2$ (0 representing control), simulated outcome values Y_{gtmj} for virtual participants $n = 1, \dots, M_g$ are generated as draws from the multinomial distribution for the primary outcome in participants of subphenotype g receiving treatment t . (See section 4.1 for the construction of the required multinomial probabilities).

4.3.2 Platform estimand

The primary estimand for the platform is set out in Table 2. This estimand is the target of analysis within each iteration of the simulations.

Table 2 - Primary estimand

Estimand attribute	Primary estimand
Population	Patients meeting the inclusion criteria and no exclusion criteria
Treatment conditions	(Active treatment + usual care) vs usual care
Outcome variable	An ordinal outcome: composite of organ support-free days up to 28 days and death
Population-level summary measure	Proportional odds ratio comparing each active treatment vs usual care
Intercurrent event: strategies	<p>Death: Composite strategy (included in the outcome)</p> <p>Protocol non-adherence: Treatment policy strategy</p> <p>Use of other effective medications: Treatment policy strategy</p>

4.3.3 Analysis model

The analysis of the simulated data is via Bayesian proportional odds regression, as planned for the actual adaptive analyses.

The regression model (excluding covariate adjustment) can be written as

$$\text{logit}(P(Y \leq y)) = \alpha_y + \beta_S X_S + \beta_B X_B$$

where α_y is the intercept term for level y , β_S is the log POR for simvastatin, β_B is the log POR for baricitinib, and X_S, X_B are indicator variables for treatment assignment to simvastatin and baricitinib respectively.

β_S and β_B are assigned vague Normal priors with mean 0 and precision 0.1. The intercepts α_y are assigned improper Normal priors with mean 0 and precision 0.

4.3.4 Simulation outputs

Within each iteration the following outputs are recorded:

- Posterior por_{igt} within each subphenotype and active intervention
- Result (efficacy or futility determined, sample size cap reached, or active intervention continues) within each subphenotype and active intervention at each adaptive analysis
- Number of adaptive analyses needed for result within each subphenotype and active intervention
- Actual sample size needed for result within each subphenotype and active intervention
-

4.3.5 Performance measures

Summarising the simulation outputs over all iterations provides the following performance measures for a given design in relation to the chosen scenario:

- Cumulative power in each subphenotype at each adaptive analysis (equal to 100% minus type II error percentage)
- Cumulative type I error percentage in each subphenotype at each adaptive analysis (equal to 100% minus percentage chance of rejecting null treatment)
- Mean and quantiles of cumulative sample size distribution at each adaptive analysis
- Mean and quantiles of number of adaptive analyses (and hence stopping time) needed for result within each subphenotype and active treatment arm.

4.3.6 Identification of optimal triggers

To determine the optimal values of the statistical triggers por_{eff} , p_{eff} , por_{fut} and p_{fut} , simulations were carried out across a wide range of potential values.

The three main key criteria used to assess and compare designs are:

- **Type I error rate** should be minimised and should not exceed 20% for each treatment in either subphenotype (considered to be a reasonable limit for a phase II platform, as successful treatments will be evaluated further at phase III). As the

futility triggers are non-binding, a conservative estimate (or upper bound) of the type I error rate must assume that active interventions are never stopped for futility.

- **Power** should be maximised and should be at least 70% for each treatment in either subphenotype within the initial 4-year funded period.
- **Expected sample size** should be minimised. Particular attention will be paid to the expected sample size required to evaluate simvastatin and baricitinib under the assumption that baricitinib achieves the minimum clinically important odds ratio of 1.4 in the hypoinflammatory subphenotype and simvastatin achieves the minimum clinically important odds ratio of 1.3 in the hyperinflammatory subphenotype, with neither treatment having any effect in the other subphenotype.

The ranges of values evaluated for each parameter were initially set as follows:

- Efficacy odds ratio threshold por_{eff} : 1.00 to 1.125
- Efficacy probability threshold p_{eff} : 0.8 to 0.95
- Futility odds ratio threshold por_{fut} : 1.025 to 1.25
- Futility probability threshold p_{fut} : 0.6 to 0.8

Later, extensions to these ranges were also explored.

4.3.7 Estimation and software

The Integrated Nested Laplace Approximation (INLA) technique has been used to estimate Bayesian models owing to the substantially faster runtime compared to Markov Chain Monte Carlo (MCMC) methods [3]. Simulations have been run in R v4.2.0 using the INLA package for Bayesian analyses [4]. A limitation of this package is that ordinal outcomes for proportional odds logistic regression can have a maximum of 10 categories. The operating characteristic simulations are therefore based on a modified version of the primary outcome, which combines the original values into categories as shown in

Table 3.

Table 3 – Collapsing the primary outcome down to 10 categories

Combined category	Organ support free days
-1	-1 (death)
0	0
1	1 to 9
2	10 to 13
3	14 to 17
4	18 to 19
5	20 to 21
6	22 to 23
7	24 to 26
8	27

A limited number of simulations (not shown here) using the full 30 primary outcome values and MCMC estimation, implemented using the Rjags package [5], have confirmed that the 10-category INLA approximation has no substantial impact on the results.

4.3.8 Number of iterations

Initially simulations have been run with the number of iterations set to 5000. Promising designs were then re-evaluated with 8000 iterations

5 RESULTS

5.1 PROBABILITY DISTRIBUTIONS FOR PRIMARY OUTCOME

Table 4 and Table 5 show the synthesised probability distributions for control subjects and those treated with a minimally effective treatment in the hypoinflammatory and hyperinflammatory subphenotypes respectively. For each phenotype the control group distribution is illustrated graphically, in Figure 2 and Figure 3 respectively.

Table 4 - Anticipated probabilities for each of the categories in the hypoinflammatory subphenotype

Ordered categories	Corresponding organ support free days	Control group probabilities from HARP2 data	Intervention group probabilities based on OR of 1.4
1 (least favourable)	-1(=death)	0.16	0.12
2	0	0.25	0.212
3	1	0	0
4	2	0.003	0.003
5	3	0.006	0.005
6	4	0.007	0.007
7	5	0.008	0.008
8	6	0.011	0.011
9	7	0.011	0.011
10	8	0.011	0.011
11	9	0.011	0.011
12	10	0.017	0.016
13	11	0.02	0.019
14	12	0.02	0.02
15	13	0.017	0.017
16	14	0.014	0.014
17	15	0.014	0.014
18	16	0.014	0.015
19	17	0.014	0.015
20	18	0.023	0.024
21	19	0.031	0.033
22	20	0.042	0.047
23	21	0.045	0.051
24	22	0.045	0.053
25	23	0.045	0.055
26	24	0.051	0.064
27	25	0.051	0.067
28	26	0.039	0.054
29	27	0.017	0.023
30 most favourable	28	0.002	0.003

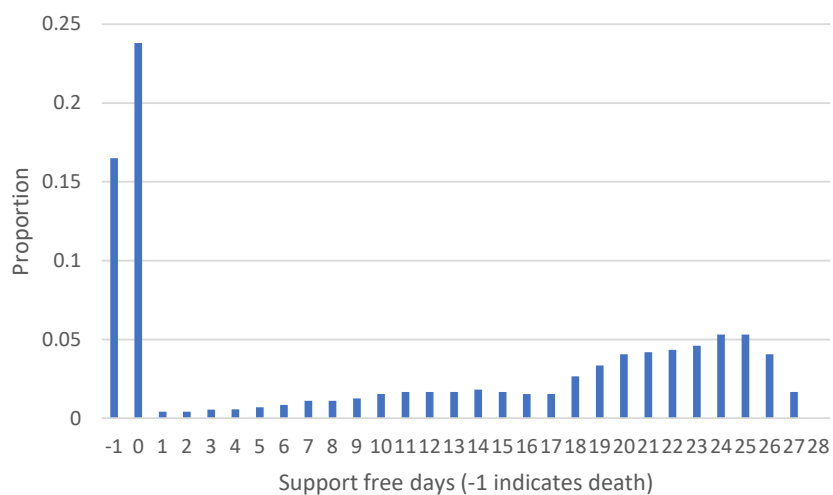


Figure 2 - Anticipated primary outcome distribution in control group (hypoinflammatory subphenotype)

Table 5 - Anticipated probabilities for each of the categories in the hyperinflammatory subphenotype

Ordered categories	Corresponding organ support free days	Control group probabilities from HARP2 data	Intervention group probabilities based on OR of 1.3
1 least favourable	-1 (=death)	0.45	0.386
2	0	0.308	0.32
3	1	0.003	0.003
4	2	0.004	0.004
5	3	0.005	0.005
6	4	0.005	0.005
7	5	0.004	0.004
8	6	0.004	0.004
9	7	0.005	0.005
10	8	0.006	0.006
11	9	0.007	0.009
12	10	0.009	0.011
13	11	0.009	0.011
14	12	0.007	0.009
15	13	0.006	0.007
16	14	0.007	0.009
17	15	0.007	0.009
18	16	0.007	0.009
19	17	0.007	0.009
20	18	0.013	0.016
21	19	0.017	0.02
22	20	0.018	0.023
23	21	0.017	0.021
24	22	0.015	0.018
25	23	0.013	0.016
26	24	0.015	0.019
27	25	0.015	0.019
28	26	0.013	0.017
29 most favourable*	27	0.006	0.007

*Probability of 28 organ support free days = 0

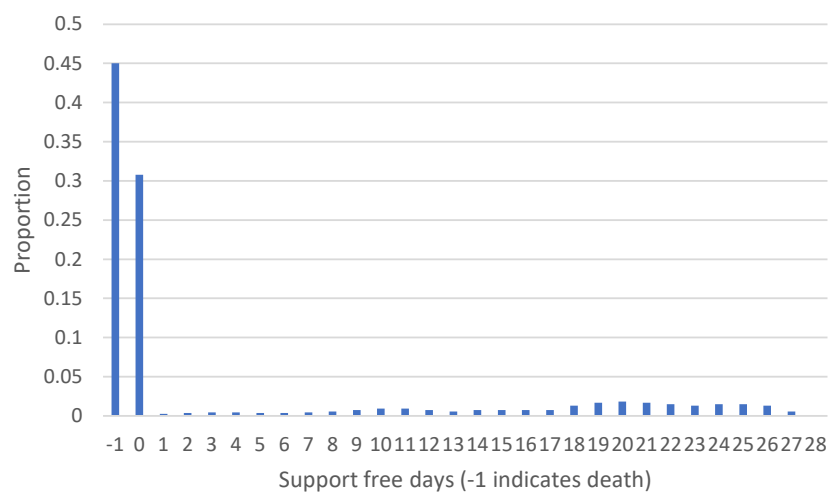


Figure 3 - Anticipated primary outcome distribution in control group (hyperinflammatory subphenotype)

5.2 OPTIMAL STATISTICAL TRIGGERS

As a reminder, for each adaptive analysis i , active intervention t and subphenotype g , the stopping rules will be of the form:

- Stop for efficacy (graduate) if $Prob_{gti}(POR > por_{eff}) \geq p_{eff}$
- Stop for futility (reject) if $Prob_{gti}(POR < por_{fut}) \geq p_{fut}$
- Also stop for futility (reject) if $N_{gti} \geq C_g$

where

- N_{gti} is the number of participants in treatment arm t and subphenotype g at adaptive analysis i
- C_g is the per-arm sample size cap in subphenotype g
- por_{eff}, por_{fut} are threshold values for the POR
- p_{eff}, p_{fut} are threshold probability values
- $Prob_{gti}(A)$ is the posterior probability of event/outcome A in treatment arm t and subphenotype g at adaptive analysis i
- This section is concerned with identification of the optimal thresholds $por_{eff}, por_{fut}, p_{eff}, p_{fut}$.

5.2.1 Efficacy triggers

Figure 4 shows an array of equally spaced points in the 2-dimensional efficacy trigger space, with the odds ratio threshold por_{eff} on the horizontal axis and the probability threshold p_{eff} on the vertical axis. The green region shows parameter combinations for which the estimated upper bound of the type I error rate (assuming no futility stopping) for a single treatment does not exceed 20% in either subphenotype. This rate is calculated at the 20th adaptive analysis, expected at 5.5 years, as an indicator of operating characteristics beyond the funded period. This is a conservative approach since type I error over the funded period will be the same or lower. All efficacy parameter combinations outside the green region can be discounted as exhibiting excessive type I error.

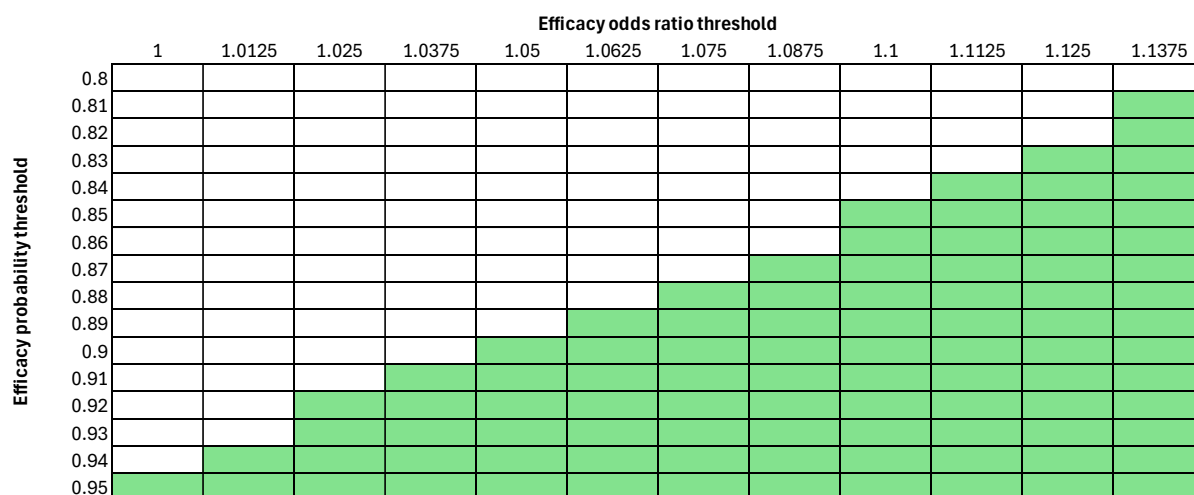


Figure 4 - Efficacy parameter combinations with acceptable Type I error control over 5.5 years

In Figure 5, the green region indicates parameter combinations where power for a single treatment in both subphenotypes exceeds 70% at the end of the 4-year funded period under the main assumptions, based on 5000 simulations for each combination. This is a conservative estimate of power since power over longer time horizons will be the same or greater. Again, no futility stopping has been allowed for, and as such the threshold of 70% represents an upper bound on the power under a non-binding futility stopping rule. Stopping for futility will reduce power, so any efficacy parameter combinations outside the green region in Figure 5 will have power below 70% regardless of the futility stopping rule, and can therefore be discounted as underpowered.

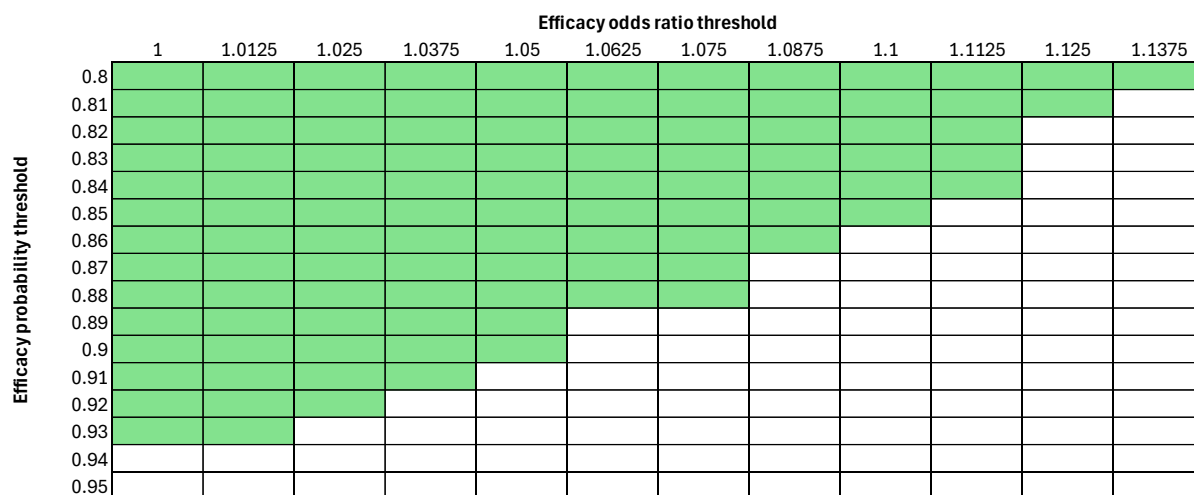


Figure 5 - Efficacy parameter combinations with potentially acceptable power over 4 years

In Figure 6, the green region is the intersection of the green regions from Figure 4 and Figure 5, within which any combination of efficacy triggers provides acceptable Type I error rate over 5.5 years regardless of futility stopping, and sufficient power over 4 years in the absence of futility stopping. These combinations are taken forward to the next stage where power and expected sample size are evaluated under a range of possible futility stopping triggers.

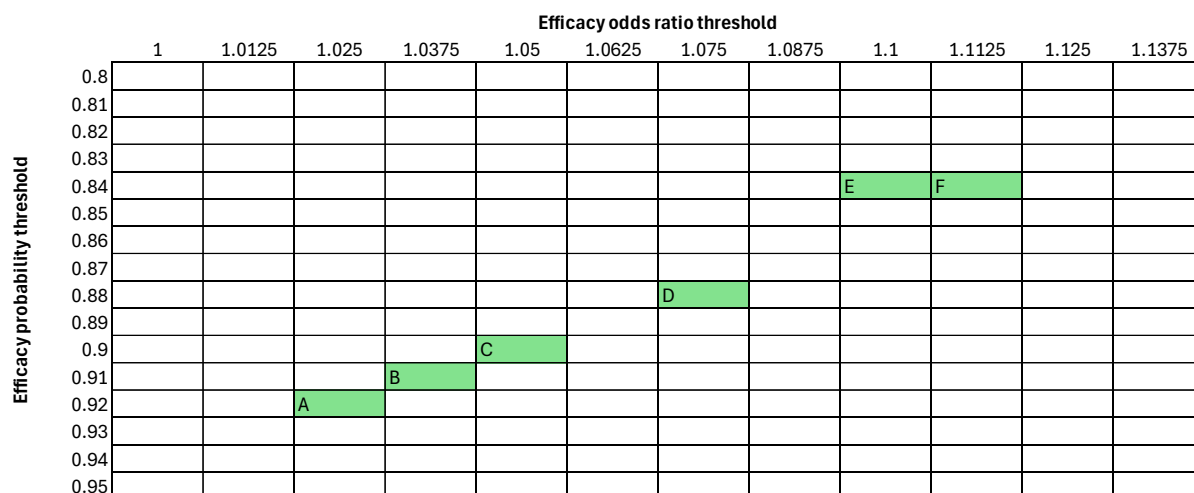


Figure 6 - Efficacy parameter combinations with acceptable Type I error control over 5.5 years and potentially acceptable power over 4 years

5.2.2 Futility triggers

For each of the viable efficacy parameter combinations identified in Figure 6, operating characteristics were evaluated under a range of futility parameter combinations. Efficacy combinations A, D and F yielded no designs with sufficient power over the 4-year funded period.

A further constraint was imposed whereby the expected overall sample size to evaluate simvastatin and baricitinib over the initial 4-year funded period was required to be less than 1000 participants. This eliminated all designs with efficacy parameter combinations B and C.

Figure 7 shows in green, for efficacy parameter combination E, the region of the futility parameter space that yield designs meeting the following criteria:

- power at least 70% over the initial 4 year funded period
- expected sample size (in the active intervention only, assuming treatment effect in line with the minimum clinically important difference) per active treatment less than 160
- expected sample size to evaluate simvastatin and baricitinib over the initial 4-year funded period less than 1000
- upper bounds of 18% (hypoinflammatory) and 20% (hyperinflammatory) on type I error over 5.5 years (already guaranteed due to the efficacy parameters).

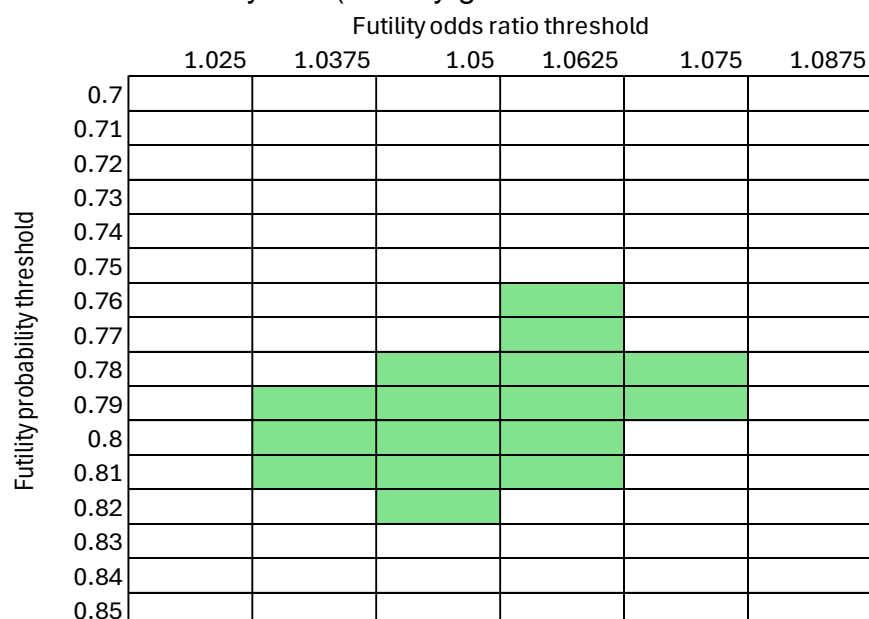


Figure 7 - Acceptable futility threshold combinations for efficacy odds ratio threshold = 1.1 and efficacy probability threshold = 0.84

The key operating characteristics of the designs evaluated within the green region are tabulated in Table 6.

Table 6 – Summary of operating characteristics of acceptable designs

Futility odds ratio threshold	Futility probability threshold	Hypoinflammatory				Hyperinflammatory				Expected platform sample size over 4 years
		Power (lower bound, 4 years)	Power (upper bound, 5.5 years)	Type I error rate (lower bound, 4 years)	Type I error rate (upper bound, 5.5 years)	Power (lower bound, 4 years)	Power (upper bound, 5.5 years)	Type I error rate (lower bound, 4 years)	Type I error rate (upper bound, 5.5 years)	
1.0375	0.79	93.3	96.2	17.8	18.8	70.5	81.2	17.9	19.5	1078
1.0375	0.8	94.1	96.2	19.0	18.8	70.7	81.2	18.1	19.5	1092
1.0375	0.81	93.7	96.2	18.0	18.8	70.8	81.2	17.9	19.5	1116
1.05	0.78	93.2	96.2	17.3	18.8	70.0	81.2	17.8	19.5	1041
1.05	0.79	93.1	96.2	18.1	18.8	70.6	81.2	18.4	19.5	1053
1.05	0.8	93.5	96.2	17.6	18.8	70.4	81.2	17.5	19.5	1066
1.05	0.81	93.6	96.2	17.4	18.8	70.5	81.2	17.6	19.5	1086
1.05	0.82	93.8	96.2	18.3	18.8	71.3	81.2	17.1	19.5	1100
1.0625	0.76	92.3	96.2	17.4	18.8	70.1	81.2	17.8	19.5	988
1.0625	0.77	92.2	96.2	17.8	18.8	70.2	81.2	17.7	19.5	995
1.0625	0.78	92.2	96.2	17.4	18.8	69.8	81.2	18.0	19.5	1008
1.0625	0.79	93.6	96.2	17.0	18.8	70.6	81.2	17.3	19.5	1026
1.0625	0.8	93.1	96.2	17.2	18.8	69.5	81.2	17.5	19.5	1053
1.0625	0.81	93.2	96.2	17.8	18.8	69.6	81.2	17.6	19.5	1057
1.075	0.78	92.4	96.2	17.4	18.8	70.2	81.2	17.9	19.5	980
1.075	0.79	92.4	96.2	17.8	18.8	70.3	81.2	17.8	19.5	996

Since all the designs have more than adequate power in the hypoinflammatory subphenotype and sufficient type I error control, the main criteria on which these designs were compared are power in the hyperinflammatory subphenotype and expected sample size.

The differences are marginal, but since it has the lowest expected sample size and power above 70% in the hyperinflammatory subphenotype, the optimal design has been selected as that with the trigger values:

- Efficacy odds ratio threshold por_{eff} : 1.1
- Efficacy probability threshold p_{eff} : 0.84
- Futility odds ratio threshold por_{fut} : 1.075
- Futility probability threshold p_{fut} : 0.78

This design is indicated in bold in Table 6.

5.3 OPERATING CHARACTERISTICS UNDER CHOSEN TRIGGERS

5.3.1 Main assumptions

5.3.1.1 Probabilities of graduation and rejection

Figure 8 and Figure 9 show the cumulative probability of graduation (power) and rejection (which includes exceeding the sample size cap) at or before the final analysis for a single active intervention in the platform, as the proportional odds ratio for the primary outcome relative to usual care is varied. Within a subphenotype, if the probabilities of graduation and rejection at a given odds ratio do not sum to 100%, the difference indicates the probability of not reaching a conclusion within the given time horizon. In Figure 8 the futility stopping rule is assumed to be binding, yielding lower bounds for power and type I error (and upper bounds for rejection probabilities), and the time horizon is set to the initial 4-year funded period. In Figure 9 it is assumed that no futility stopping occurs, yielding upper bounds for power and type I error over a longer time horizon of 5.5 years (and lower bounds for rejection probabilities, with all rejections taking place in the hypoinflammatory subphenotype due to exceeding the sample size cap).

In the hypoinflammatory subphenotype:

- power at the minimum clinically important POR of 1.4 is at least 92% over 4 years and at most 96% over 5.5 years.
- type I error rate (POR=1) is at least 17% over 4 years and at most 19% over 5.5 years.

In the hyperinflammatory subphenotype:

- power at the minimum clinically important POR of 1.3 is at least 70% over 4 years and at most 81% over 5.5 years,
- type I error rate (POR=1) is at least 18% over 4 years and at most 20% over 5.5 years.

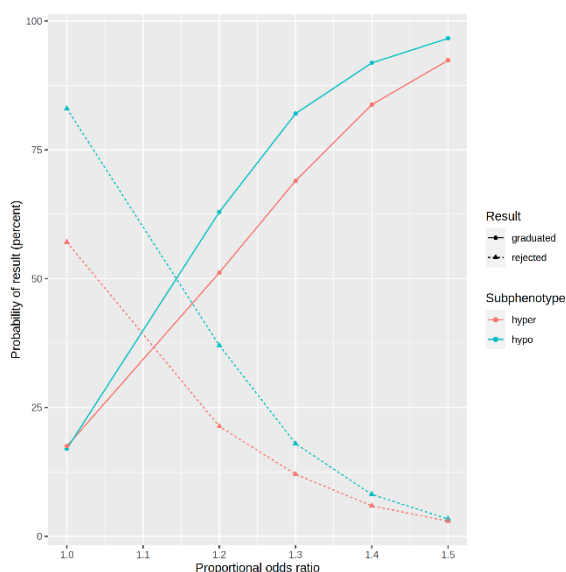


Figure 8 - Graduation and rejection curves under main assumptions – assuming binding futility stopping, 4 year time horizon

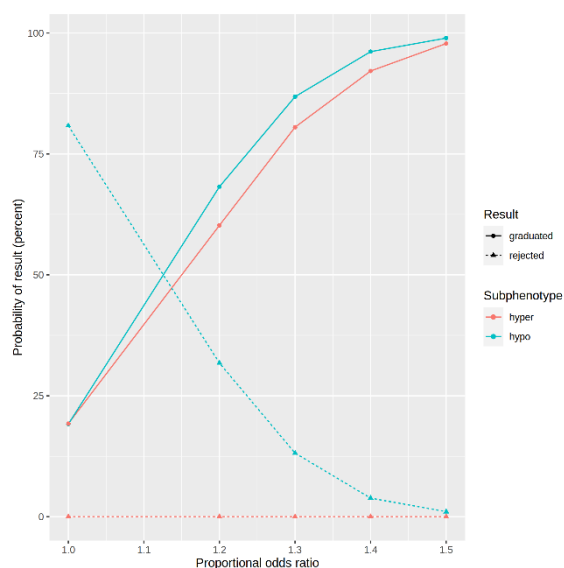


Figure 9 - Graduation and rejection curves under main assumptions – assuming no futility stopping, 5.5 year time horizon

5.3.1.2 Sample size and time required

Platform sample size

Error! Reference source not found. shows the distribution, obtained from simulations, of the total sample size required to evaluate simvastatin and baricitinib up to a maximum of 4 years under the main assumptions, i.e a recruitment rate of 1 participant per site per month, a maximum of 70 sites, 30% hyperinflammatory participants, and mortality of 18% (hypoinflammatory) / 45% (hyperinflammatory).

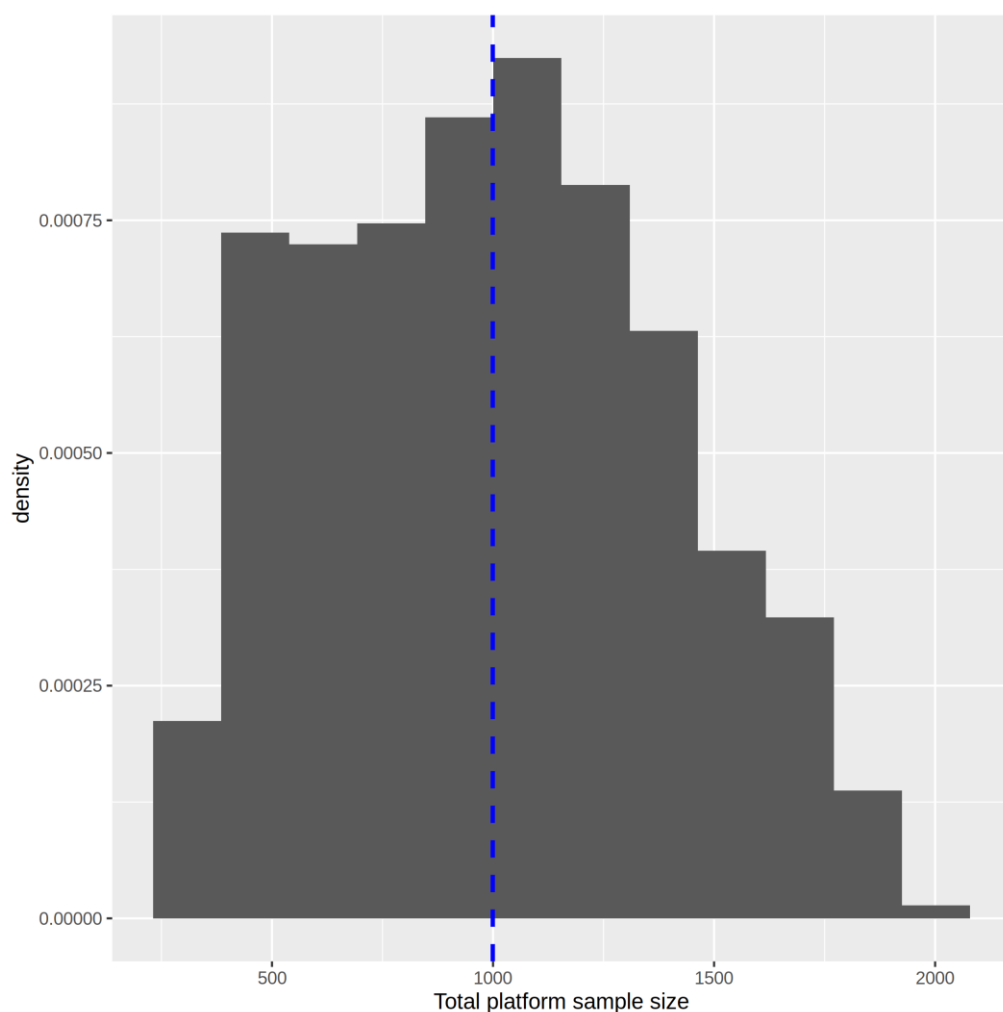


Figure 10 - Histogram showing distribution of total sample size required to evaluate simvastatin and baricitinib. The expected (mean) sample size is indicated by the dashed blue line.

Error! Reference source not found. shows the expected (mean) and 80th percentile sample size in each treatment arm in each subphenotype over the 4-year funded period. The 80th percentile is the value that we are 80% confident will not be exceeded.

When working with averages and percentiles, the sample size tends to be lower in each active arm than in the control arm; this is because each active arm has a chance of stopping before the control arm, while the control arm must always wait for both active arms to stop.

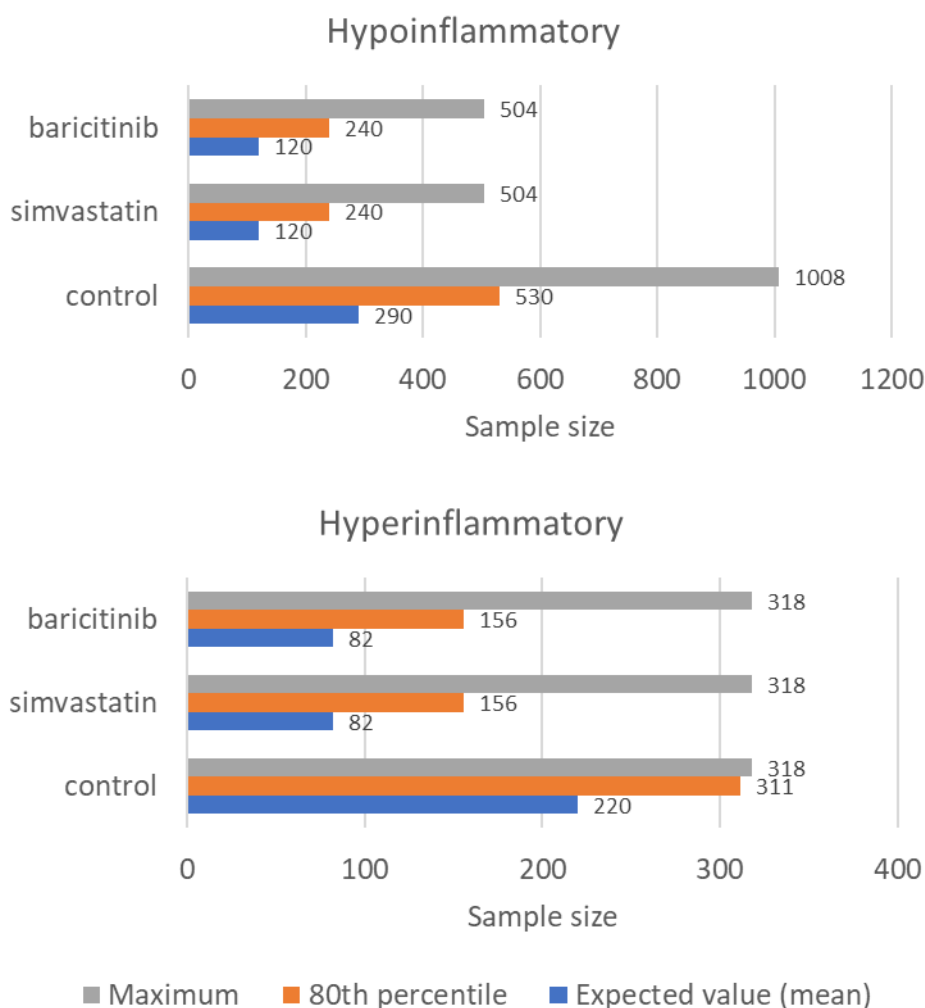


Figure 11 – Expected (mean) and 80th quantile sample size by subphenotype and treatment arm, over 4 years.

Time to fully evaluate simvastatin and baricitinib in both subphenotypes

Error! Reference source not found. shows the distribution, obtained from simulations, of the overall time required to fully evaluate simvastatin and baricitinib under the main assumptions.

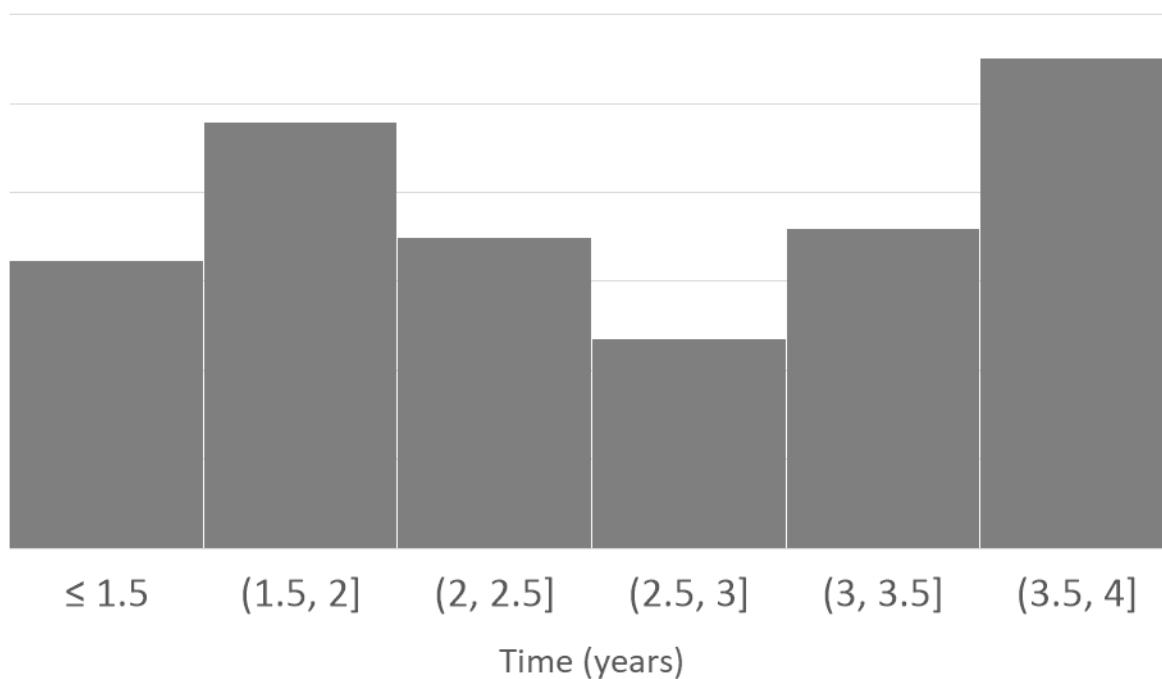


Figure 12 - Histogram showing distribution of time required to evaluate simvastatin and baricitinib in both subphenotypes. The expected (mean) time is indicated by the blue dashed line.

Time until first treatment arm stopped in hypoinflammatory subphenotype

Error! Reference source not found. shows the distribution, obtained from simulations, of the time until the first treatment arm is stopped in the hypoinflammatory subphenotype under the main assumptions, i.e. the earliest time when a new intervention could be started in this subphenotype.

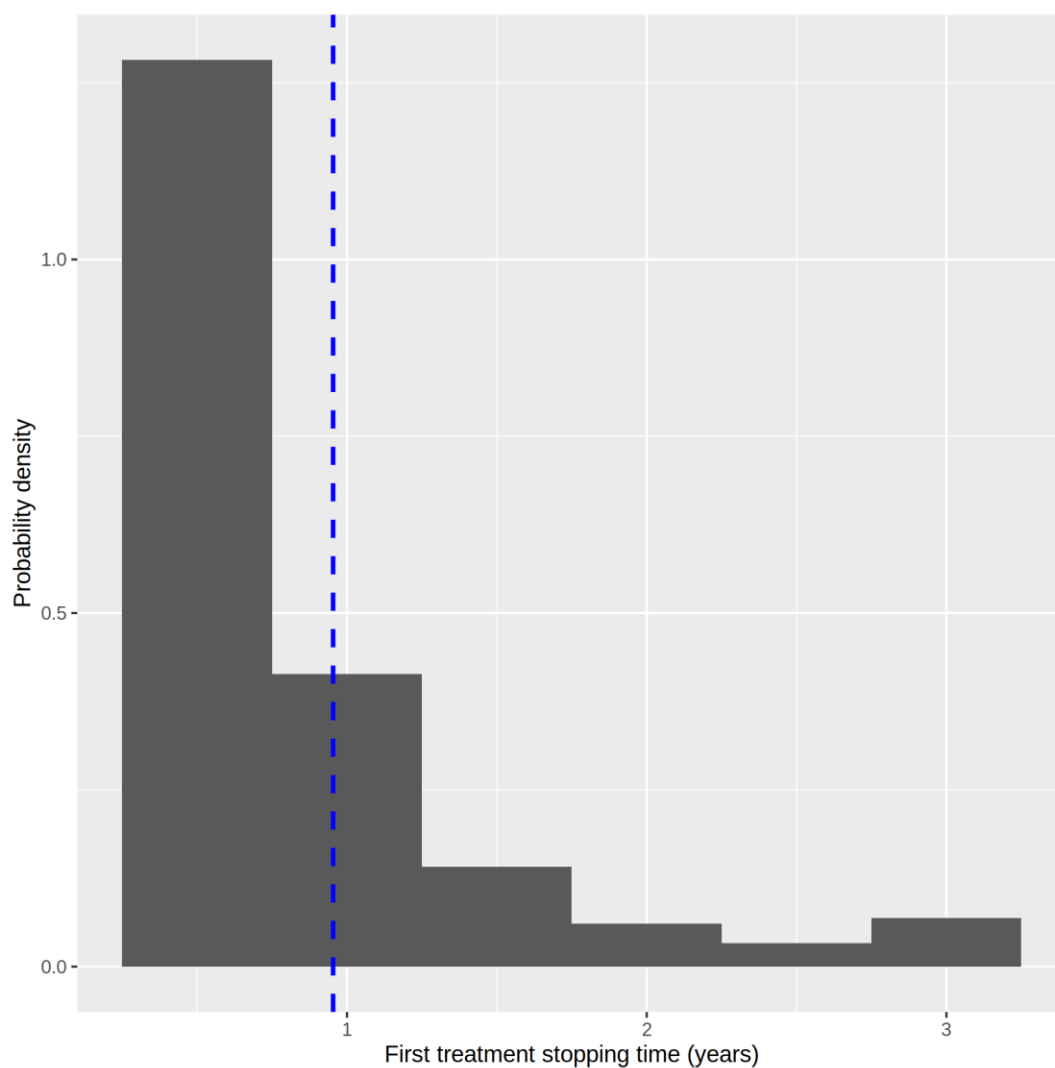


Figure 13 - Histogram showing distribution of time until first treatment is stopped in hypoinflammatory subphenotype The expected (mean) time is indicated by the blue dashed line.

Time until first treatment arm stopped in hyperinflammatory subphenotype
Error! Reference source not found. shows the distribution of the time until the first treatment arm is stopped in the hyperinflammatory subphenotype under the main assumptions, i.e. the earliest time when a new intervention could be started in this subphenotype.

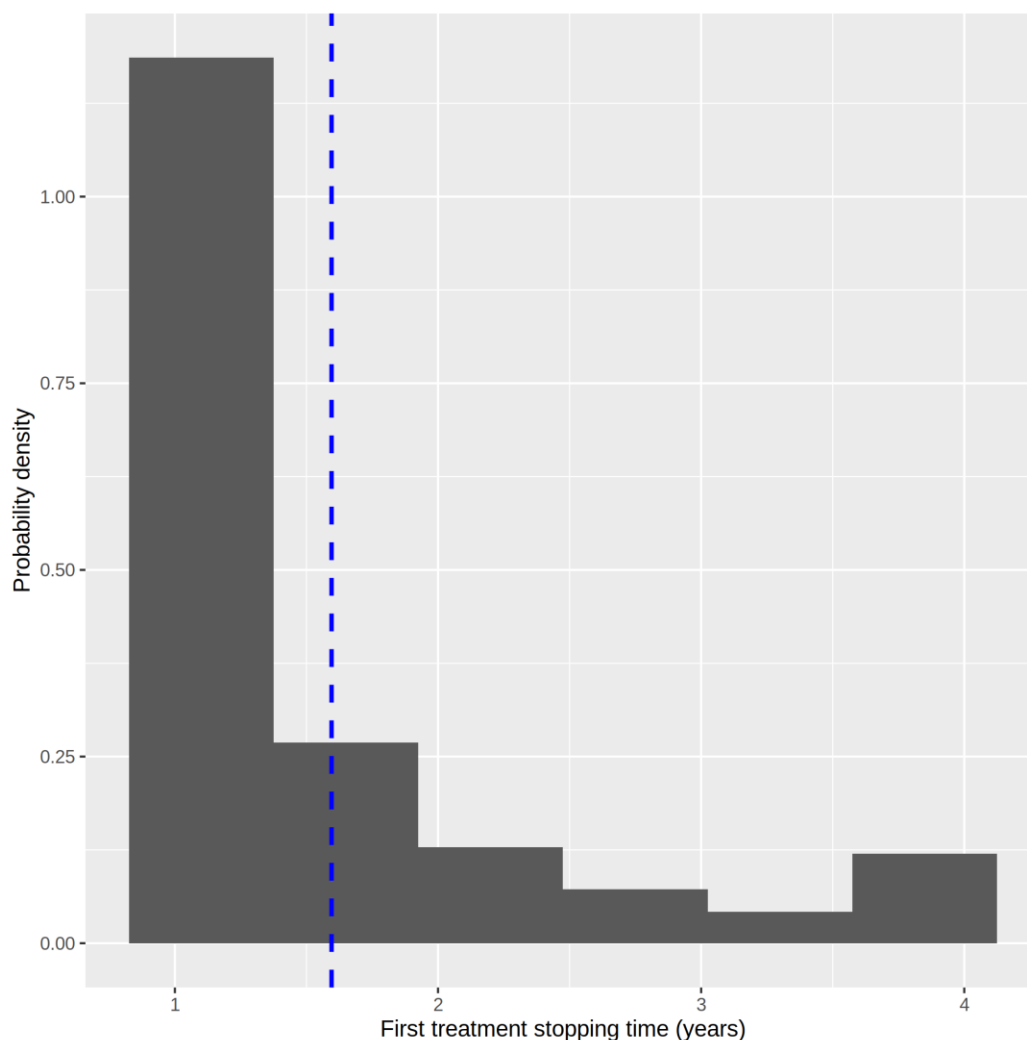


Figure 14 - Histogram showing distribution of time until first treatment is stopped in hyperinflammatory subphenotype. The expected (mean) time is indicated by the blue dashed line.

5.3.2 Sensitivity analyses

5.3.2.1 Probabilities of gradation and rejection

Further graduation and rejection curves are shown below for each modelled recruitment and control mortality scenario. Boundaries for power and type I error rates are also stated in each case.

Double recruitment rate per site to 2 participants per month

Under this scenario the first adaptive analysis (which occurs at the end of the month in which 240 hypoinflammatory participants are recruited) occurs later than previously assumed, but at approximately the same sample size. Subsequent adaptive analyses occur at fixed intervals and will therefore have greater sample sizes than under the main assumptions.

In the hypoinflammatory subphenotype:

- power at the minimum clinically important POR of 1.4 is at least 92% over 4 years and at most 96% over 5.5 years.
- type I error rate (POR=1) is at least 15% over 4 years and at most 15% over 5.5 years.

In the hyperinflammatory subphenotype:

- power at the minimum clinically important POR of 1.3 is at least 80% over 4 years and at most 83% over 5.5 years,
- type I error rate (POR=1) is at least 15% over 4 years and at most 16% over 5.5 years.

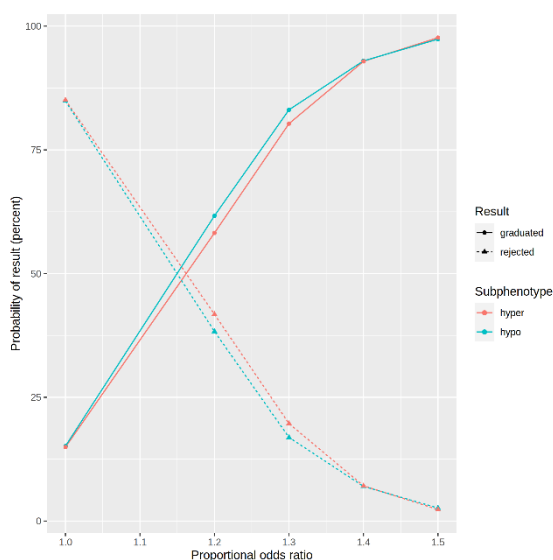


Figure 15 - Graduation and rejection curves under doubled recruitment rate assumption – assuming binding futility stopping, 4 year time horizon

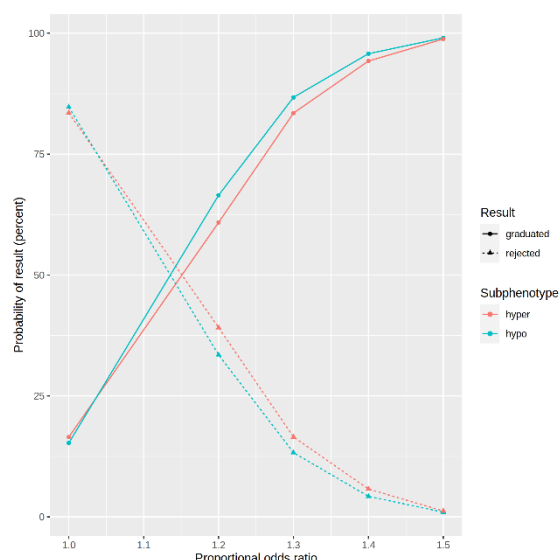


Figure 16 - Graduation and rejection curves under doubled recruitment rate assumption – assuming no futility stopping, 5.5 year time horizon

Halve recruitment rate per site to 0.5 participants per month

Under this scenario, the first adaptive analysis (which occurs at the end of the month in which 240 hypoinflammatory participants are recruited) takes place later than previously assumed, but at approximately the same sample size. Subsequent adaptive analyses

occur at fixed intervals and will therefore have smaller sample sizes than under the main assumptions.

In the hypoinflammatory subphenotype:

- power at the minimum clinically important POR of 1.4 is at least 87% over 4 years and at most 96% over 5.5 years.
- type I error rate (POR=1) is at least 23% over 4 years and at most 26% over 5.5 years.

In the hyperinflammatory subphenotype:

- power at the minimum clinically important POR of 1.3 is at least 63% over 4 years and at most 66% over 5.5 years,
- type I error rate (POR=1) is at least 19% over 4 years and at most 20% over 5.5 years.

,

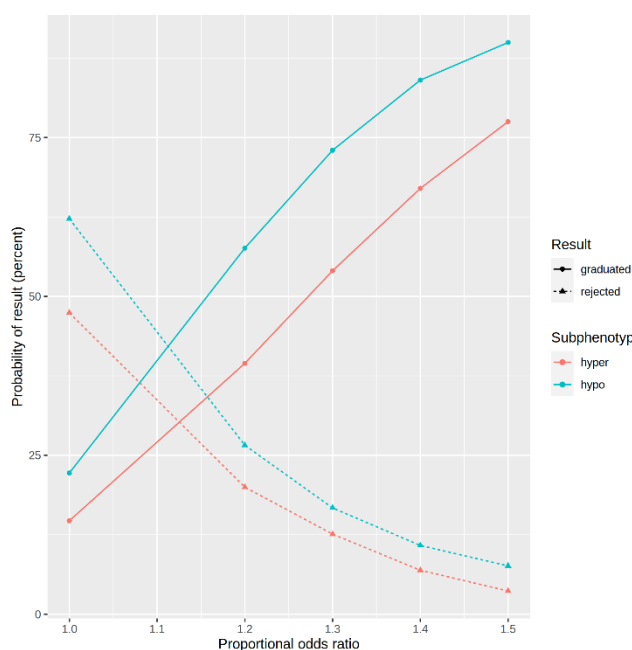


Figure 17 - Graduation and rejection curves under halved recruitment rate assumption – assuming binding futility stopping, 4 year time horizon

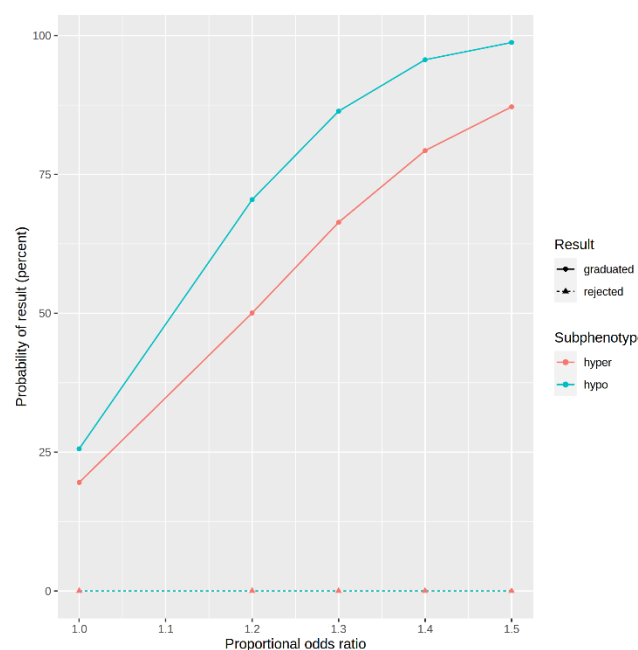


Figure 18 - Graduation and rejection curves under halved recruitment rate assumption – assuming no futility stopping, 5.5 year time horizon

Increase number of sites to 100

Under this scenario, the first adaptive analysis (which occurs at the end of the month in which 240 hypoinflammatory participants are recruited) occurs earlier but at approximately the same sample size. Subsequent adaptive analyses occur at fixed intervals and will therefore have greater sample sizes than under the main assumptions.

In the hypoinflammatory subphenotype:

- power at the minimum clinically important POR of 1.4 is at least 92% over 4 years and at most 96% over 5.5 years.
- type I error rate (POR=1) is at least 17% over 4 years and at most 18% over 5.5 years.

In the hyperinflammatory subphenotype:

- power at the minimum clinically important POR of 1.3 is at least 74% over 4 years and at most 85% over 5.5 years,
- type I error rate (POR=1) is at least 17% over 4 years and at most 18% over 5.5 years.

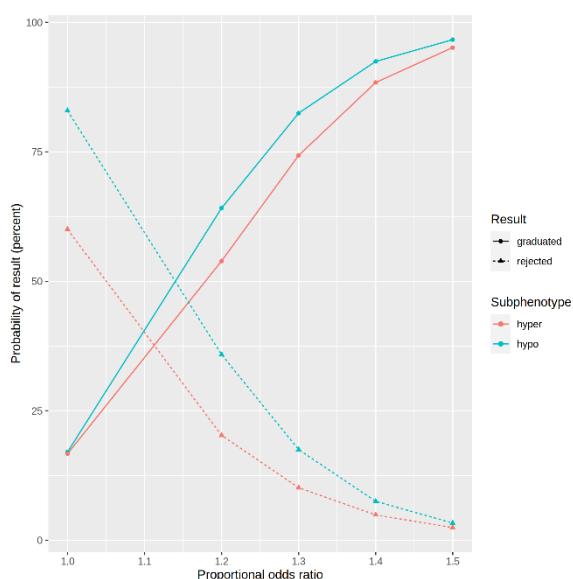


Figure 19 - Graduation and rejection curves under increased sites assumption – assuming binding futility stopping, 4 year time horizon

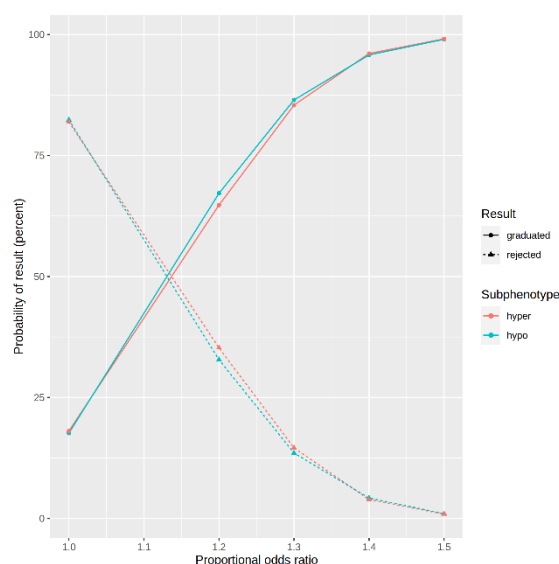


Figure 20 - Graduation and rejection curves under increased sites assumption – assuming no futility stopping, 5.5 year time horizon

Increase proportion in hyperinflammatory subphenotype to 35%

Under this scenario, the first adaptive analysis (which occurs at the end of the month in which 240 hypoinflammatory participants are recruited) occurs later but at approximately the same sample size. Subsequent adaptive analyses will have smaller sample sizes than under the main assumptions in the hypoinflammatory subphenotype, but larger sample sizes in the hyperinflammatory subphenotype.

In the hypoinflammatory subphenotype:

- power at the minimum clinically important POR of 1.4 is at least 92% over 4 years and at most 96% over 5.5 years.
- type I error rate (POR=1) is at least 18% over 4 years and at most 19% over 5.5 years.

In the hyperinflammatory subphenotype:

- power at the minimum clinically important POR of 1.3 is at least 74% over 4 years and at most 83% over 5.5 years,
- type I error rate (POR=1) is at least 18% over 4 years and at most 18% over 5.5 years.

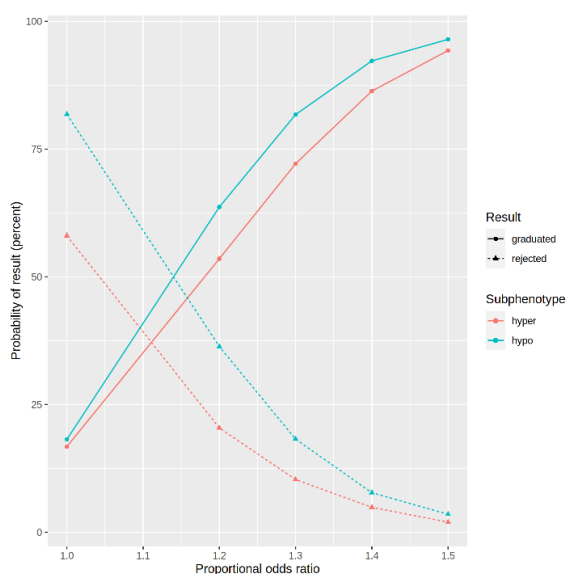


Figure 21 - Graduation and rejection curves under increased hyperinflammatory prevalence assumption – assuming binding futility stopping, 4 year time horizon

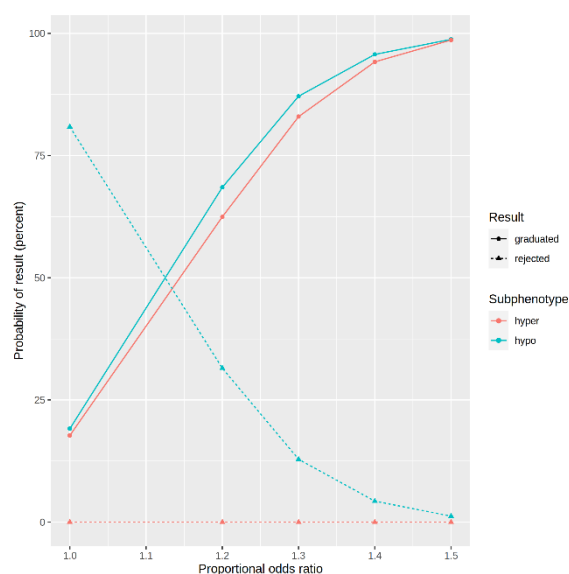


Figure 22 - Graduation and rejection curves under increased hyperinflammatory prevalence assumption – assuming no futility stopping, 5.5 year time horizon

Reduce proportion in hyperinflammatory subphenotype to 25%

Under this scenario, the first adaptive analysis (which occurs at the end of the month in which 240 hypoinflammatory participants are recruited) occurs earlier but at approximately the same sample size. Subsequent adaptive analyses will have smaller sample sizes than under the main assumptions in the hyperinflammatory subphenotype, but larger sample sizes in the hypoinflammatory subphenotype.

In the hypoinflammatory subphenotype:

- power at the minimum clinically important POR of 1.4 is at least 92% over 4 years and at most 95% over 5.5 years.
- type I error rate (POR=1) is at least 17% over 4 years and at most 18% over 5.5 years.

In the hyperinflammatory subphenotype:

- power at the minimum clinically important POR of 1.3 is at least 68% over 4 years and at most 77% over 5.5 years,
- type I error rate (POR=1) is at least 17% over 4 years and at most 19% over 5.5 years.

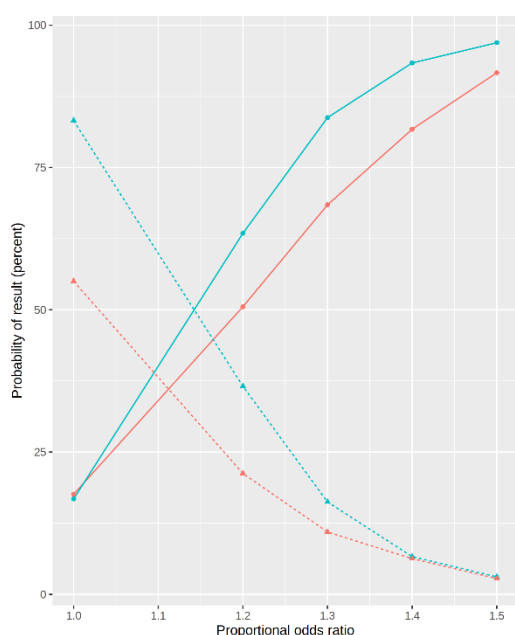


Figure 23 - Graduation and rejection curves under reduced hyperinflammatory prevalence assumption – assuming binding futility stopping, 4 year time horizon

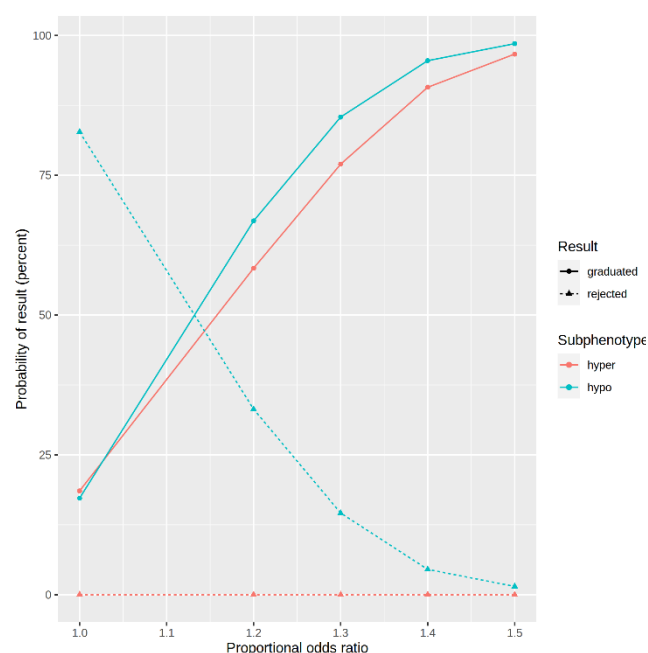


Figure 24 - Graduation and rejection curves under reduced hyperinflammatory prevalence assumption – assuming no futility stopping, 5.5 year time horizon

Reduce control mortality rate to 13% (hypoinflammatory) / 30% (hyperinflammatory)

Under this scenario, recruitment is unchanged, as is the proportional odds ratio associated with treatment in each subphenotype, but the baseline mortality rates in the control arm are assumed to be lower. This influences the precision of the estimated odds ratios in the proportional odds model, which affects the chances of the stopping rules being triggered.

In the hypoinflammatory subphenotype:

- power at the minimum clinically important POR of 1.4 is at least 80% over 4 years and at most 93% over 5.5 years.
- type I error rate (POR=1) is at least 24% over 4 years and at most 27% over 5.5 years.

In the hyperinflammatory subphenotype:

- power at the minimum clinically important POR of 1.3 is at least 78% over 4 years and at most 88% over 5.5 years,
- type I error rate (POR=1) is at least 17% over 4 years and at most 19% over 5.5 years.

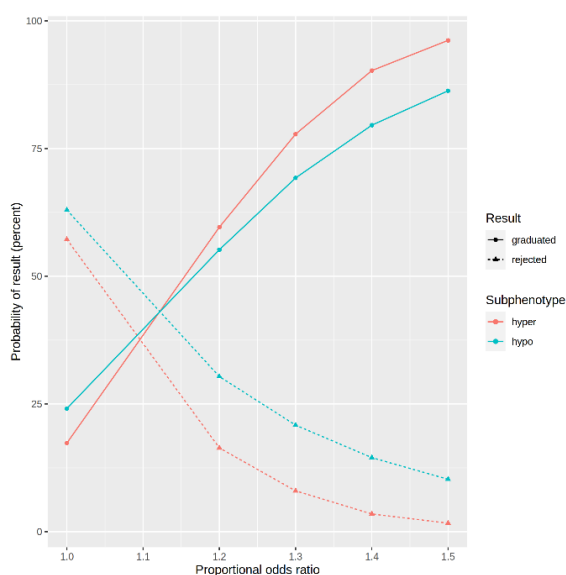


Figure 25 - Graduation and rejection curves under reduced mortality assumption – assuming binding futility stopping, 4 year time horizon

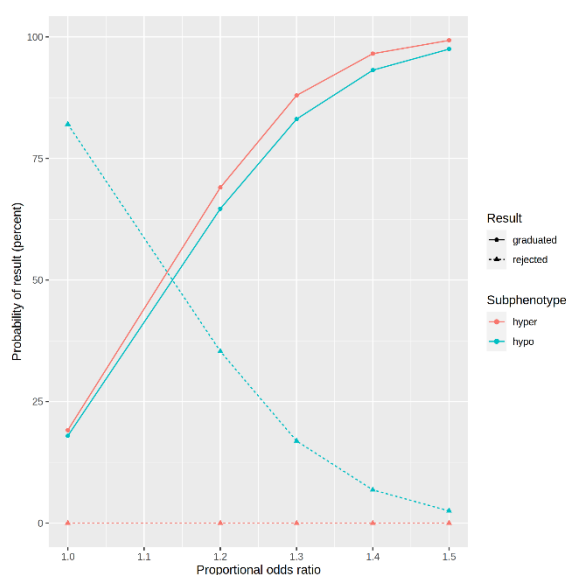


Figure 26 - Graduation and rejection curves under reduced mortality assumption – assuming no futility stopping, 5.5 year time horizon

Side-by-side comparison

Error! Reference source not found. shows the lower bound (assuming binding futility stopping) on power over 4 years to detect the minimum clinically important treatment effect in each subphenotype for each set of recruitment and control mortality assumptions.

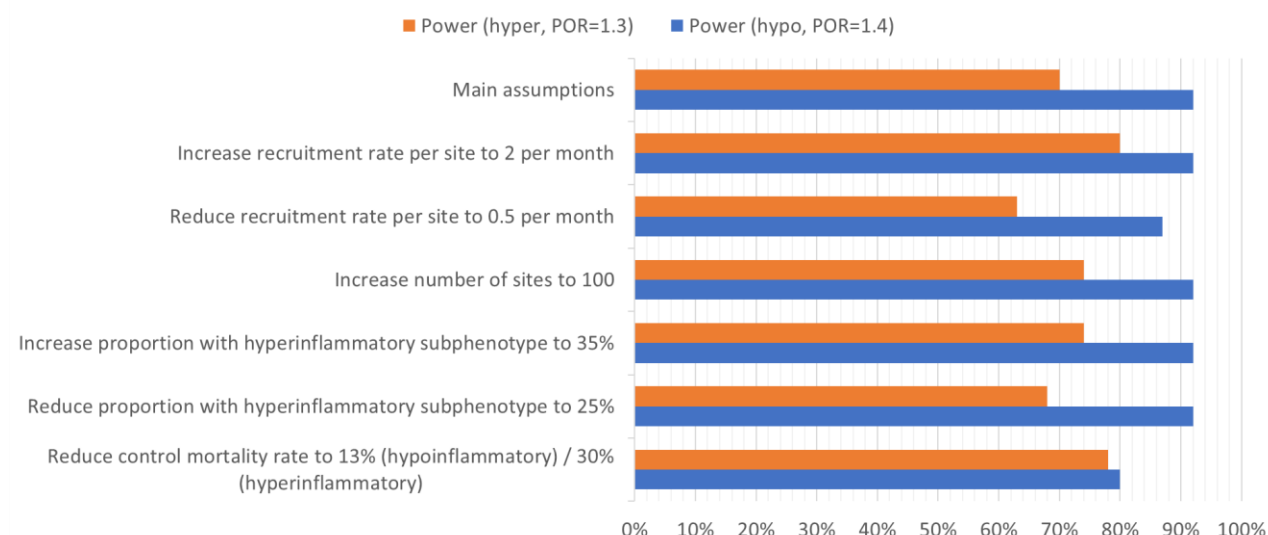


Figure 27 - Sensitivity to assumptions (power, lower bound over 4 years). The main assumptions are a recruitment rate of 1 participant per site per month, a maximum of 70 sites, 30% hyperinflammatory participants, mortality of 18% (hypoinflammatory) / 45% (hyperinflammatory), and binding futility stopping.

Error! Reference source not found. shows the type I error rate in each subphenotype for each set of recruitment and control mortality assumptions.

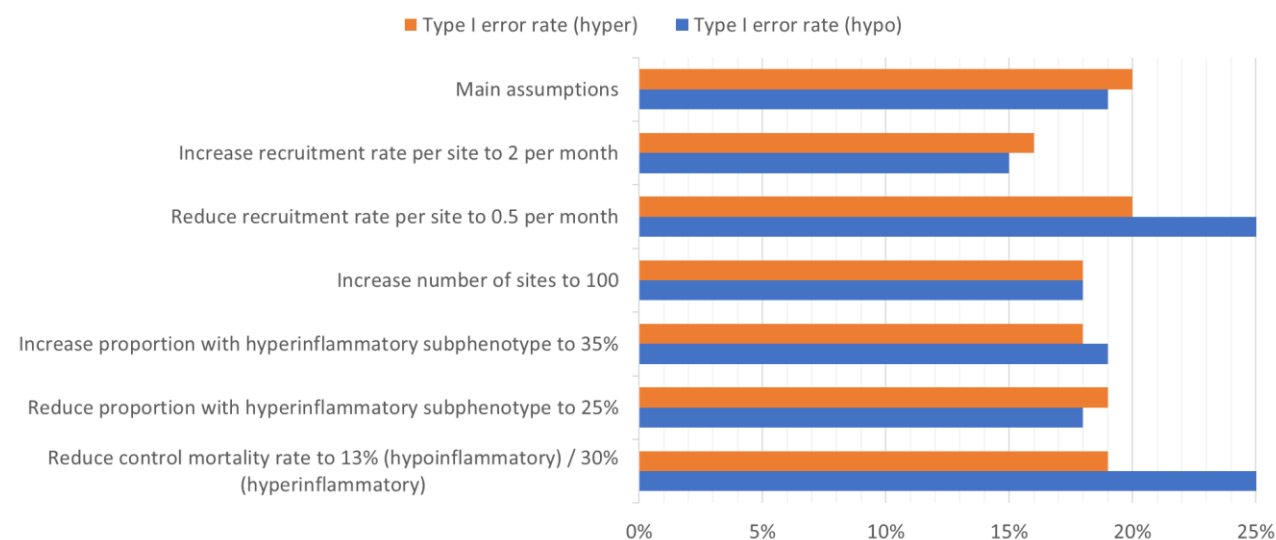


Figure 28 - Sensitivity to assumptions (type I error, upper bound over 5.5 years). The main assumptions are a recruitment rate of 1 participant per site per month, a maximum of 70 sites, 30% hyperinflammatory participants, mortality of 18% (hypoinflammatory) / 45% (hyperinflammatory), and no futility stopping.

5.3.2.2 Sample size and time required

Figure 29 shows for each set of alternative recruitment and control mortality assumptions the expected value (mean), 80th percentile and maximum of the distribution of the sample size required to evaluate simvastatin and baricitinib over the 4-year funded period.

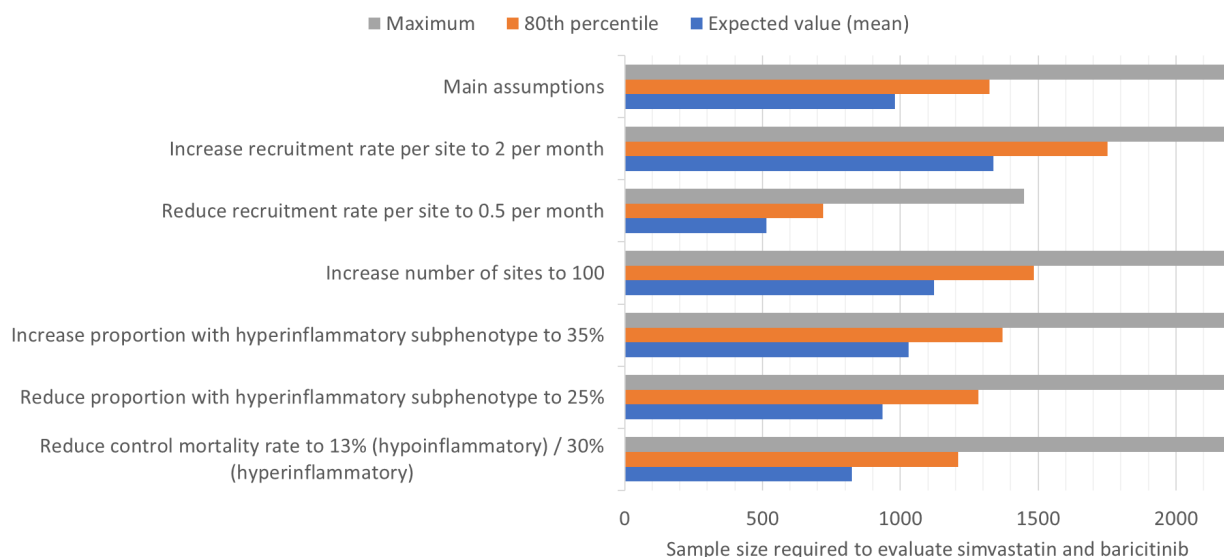


Figure 29 - Sensitivity to recruitment and control mortality assumptions (platform sample size over 4 years)

Figure 30 shows for each set of recruitment and control mortality assumptions the expected value (mean), 80th percentile and maximum of the distribution of the time required to evaluate simvastatin and baricitinib up to a maximum of 4 years.

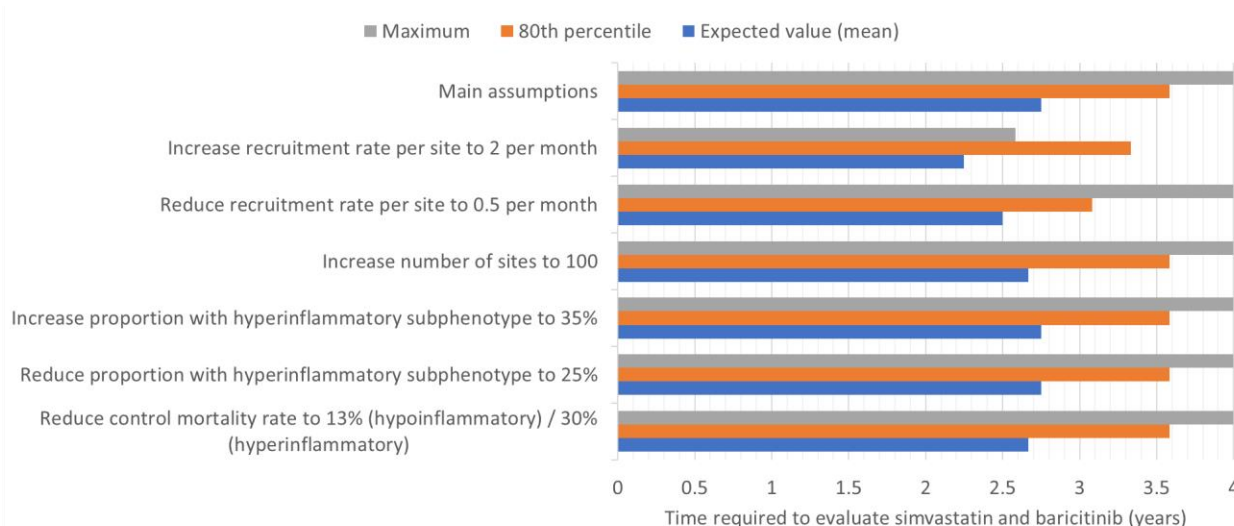


Figure 30 - Sensitivity to recruitment and control mortality assumptions (platform time required up to 4 years)

Table 7 summarises the sensitivity of the platform to the recruitment and control mortality assumptions. Again, the power estimates are lower bounds over 4 years (i.e. binding futility is assumed) while the type I error estimates are upper bounds over 5.5 years (assuming no futility stopping). Sample sizes also assume binding futility stopping.

Table 7 - Overall summary of sensitivity analysis results

Change in assumptions	Expected sample size to evaluate simvastatin and baricitinib assuming 4-year time horizon	80th percentile sample size to evaluate simvastatin and baricitinib assuming 4-year time horizon	Max sample size to evaluate simvastatin and baricitinib assuming 4-year time horizon	Power to detect OR=1.4/1.3 in hypo/hyperinflammatory subphenotype (4 years, binding futility stopping)*	One-sided type I error probability in hypo/hyperinflammatory subphenotype (5.5 years, no futility stopping)*	Average time to evaluate simvastatin and baricitinib in both phenotypes (capped at 4 years)
Main assumptions	980	1324	2445	92% / 70%	19% / 20%	2y 9m
Increase recruitment rate per site to 2 per month	1338	1752	3378	92% / 80%	15% / 16%	2y 3m
Reduce recruitment rate per site to 0.5 per month	516	720	1449	87% / 63%	26% / 20%	2y 8m
Increase number of sites to 100	1122	1486	2766	92% / 74%	18% / 18%	2y 8m
Increase proportion with hyperinflammatory subphenotype to 35%	1031	1371	2601	92% / 74%	19% / 18%	2y 9m
Reduce proportion with hyperinflammatory subphenotype to 25%	936	1284	2289	92% / 68%	18% / 19%	2y 9m
Reduce control mortality rate to 13% (hypoinflammatory) / 30% (hyperinflammatory)	824	1211	2445	80% / 78%	27% / 19%	2y 8m

* The futility rule is non-binding in practice but assuming it is binding yields a conservative lower bound for power. Assuming it never applies yields a conservative lower bound for type I error.

The operating characteristics in the hypoinflammatory subphenotype are fairly insensitive to recruitment assumptions since the maximum sample size cap is always expected to be reached. In the hyperinflammatory subphenotype, greater recruitment leads to higher power and lower type I error while lower recruitment goes in the opposite direction. A lower mortality rate would worsen the operating characteristics in the hypoinflammatory subphenotype but improve them in the hyperinflammatory subphenotype.

Figure 31 shows how the expected sample size for evaluation of simvastatin and baricitinib varies if the assumed effect size in each subphenotype for **one** active treatment is changed (the other active treatment is still assumed to have no effect in hypoinflammatory and a proportional odds ratio of 1.3 in hyperinflammatory).

For example, if the following treatment effect assumptions are made:

- Simvastatin: no effect (POR=1.0) in hypoinflammatory, POR=1.3 in hyperinflammatory (as per main assumptions)
 - Baricitinib: POR=1.2 in hypoinflammatory, POR=1.4 in hyperinflammatory
- then the expected sample size is 934.

		Proportional Odds Ratio in hyperinflammatory subphenotype							
		0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
Proportional Odds Ratio in hypoinflammatory subphenotype	0.8	782	806	834	844	834	824	796	780
	0.9	839	868	886	900	893	880	855	832
	1.0	909	939	965	970	969	944	924	899
	1.1	946	969	997	1007	992	987	959	934
	1.2	925	959	971	981	977	962	934	913
	1.3	872	896	927	933	934	918	887	869
	1.4	838	862	887	896	889	869	848	831
	1.5	802	837	853	860	857	840	818	795

Figure 31 - Sensitivity of expected sample size to assumed treatment effects (changing the effect of one treatment only). The numbers in the coloured cells show the expected sample size for evaluation of simvastatin and baricitinib, with red shades corresponding to large values and green shades to small values.

6 APPENDIX: FREQUENTIST SAMPLE SIZE CALCULATIONS

Using the method of White et al [6] and the accompanying Stata package `artcat`, we have calculated the frequentist sample size required for a parallel group fixed design (with no early stopping) for an ordered categorical outcome. We have then applied an inflation factor calculated in the R package `gsDesign` to estimate the maximum sample size for a group sequential design with adaptive analysis boundaries and timing broadly reflecting those in the PANTHER platform. This yields an approximate equivalent frequentist maximum sample size as a means to cap the required sample size under our Bayesian design.

For the **hypoinflammatory** subphenotype the full details of the assumed probabilities for each category of the ordinal outcome in the control group are shown in Table 4. For example, the control group was estimated to have a 0.16 probability of the least favourable outcome (death), a 0.25 probability of the next least favourable outcome (0 organ support free days) and so on up to a 0.02 probability of the most favourable outcome (28 days free of organ support).

We are satisfied with 80% power and, owing to the large sample size likely to be available in this subphenotype, can restrict the risk of a type I error to 10%. As the event rate is lower in the hypoinflammatory subphenotype we are also willing to accept a slightly higher proportional odds ratio to correspond to an approximate 4% difference in mortality. Therefore, we calculated the sample size required to have 80% power and 10% type I error if the treatment arm achieves a POR of 1.4 for a favourable outcome in comparison to the control arm. This is expressed in the `artcat` command by setting the left most category as the least favourable outcome and inverting the odds ratio such that:

- `artcat, pc(0.16, 0.25, 0.0000001, 0.002820144, 0.005640288, 0.00705036, 0.008460432, .011280576, 0.011280576, 0.011280576, 0.011280576, 0.016920863, 0.019741007, 0.019741007, 0.016920863, 0.014100719, 0.014100719, 0.014100719, 0.014100719, 0.022561151, 0.031021583, 0.042302158, 0.045122302, 0.045122302, 0.045122302, 0.05076259, 0.05076259, 0.039482014, 0.016920863) or(1/1.4) unfavourable alpha(0.1)`

The fixed design sample size is 334 per arm.

The group sequential design inflation factor is 1.51, based on adaptive analyses in line with the design schedule and O'Brien-Fleming asymmetrical boundaries for efficacy and futility. The number of stages was refined iteratively to match the number of adaptive analyses required to reach the maximum sample size, resulting in a 10-stage group sequential design.

This means **the required maximum sample size is 504 per arm.**

For the **hyperinflammatory** subphenotype the full details of the assumed probabilities for each category of the ordinal outcome in the control group are shown in Table 5. For example, the control group was estimated to have a 0.45 probability of the least favourable outcome (death), a 0.308 probability of the next least favourable outcome (0 organ support free days) and so on up to a 0.006 probability of the most favourable outcome (27 days free of organ support), there was a 0 probability of 28 days free of organ support in the HARP-2 study.

In the hyperinflammatory subphenotype we are willing to accept a higher type I error owing to the smaller sample size likely to be available, since treatments that graduate will be further evaluated in phase III trials. As the event rate is higher in the hyperinflammatory subphenotype we also want to detect a lower proportional odds ratio to maintain an important absolute change in mortality. We calculated the sample size required to have 80% power and 25% type I error if the treatment arm achieves a POR of 1.3 for a favourable outcome in comparison to the control arm. This is expressed in the arcat command as follows:

- artcat, pc(0.45, 0.308, 0.002770992, 0.003694656, 0.004618321, 0.004618321, 0.003694656, 0.003694656, 0.004618321, 0.005541985, 0.007389313, 0.009236641, 0.009236641, 0.007389313, 0.005541985, 0.007389313, 0.007389313, 0.007389313, 0.012931298, 0.016625954, 0.018473282, 0.016625954, 0.014778626, 0.012931298, 0.014778626, 0.014778626, 0.012931298, 0.005541985) or(1/1.3) unfavourable power(0.8) alpha(0.25)

The fixed design sample size is 387 per arm (total 1161).

The group sequential design inflation factor is 1.366, based on adaptive analyses in line with the design schedule and O'Brien-Fleming asymmetrical boundaries for efficacy and futility. The number of stages was refined iteratively to match the number of adaptive analyses required to reach the maximum sample size, resulting in a 12-stage group sequential design.

This means **the required maximum sample size is 529 per arm.**

REFERENCES

1. McAuley, D.F., et al., *Simvastatin in the Acute Respiratory Distress Syndrome*. New England Journal of Medicine, 2014. **371**(18): p. 1695-1703.
2. Force*, T.A.D.T., *Acute Respiratory Distress Syndrome: The Berlin Definition*. JAMA, 2012. **307**(23): p. 2526-2533.
3. Rue, H., S. Martino, and N. Chopin, *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2009. **71**(2): p. 319-392.
4. *R-INLA Project*. Available from: www.r-inla.org.
5. *Package 'rjags'*. 2023; Available from: <https://cran.r-project.org/web/packages/rjags/rjags.pdf>.
6. White, I.R., et al., *artcat: Sample-size calculation for an ordered categorical outcome*. The Stata Journal, 2023. **23**(1): p. 3-23.

7 REVISION HISTORY

Version	Date	Summary of changes
1.0	19 FEB 2025	First version