

# Statistical simulation report for the respiratory research question within the TipTop platform: sample size and operating characteristics for an adaptive design.

*Lead author and simulation programming*

*Munya Dimairo ([m.dimairo@sheffield.ac.uk](mailto:m.dimairo@sheffield.ac.uk))  
Clinical Trials Research Unit  
Sheffield Centre for Health and Related Research (SCHARR)  
Division of Population Health  
School of Medicine and Population Health*

*Contributors*

*Christopher Partlett  
Nottingham Clinical Trials Unit  
School of Medicine  
University of Nottingham*

## Abbreviations

BPD	Bronchopulmonary dysplasia
ICC	Intra-cluster correlation coefficient
nCPAP	Nasal Continuous Positive Airway Pressure
NIPPV	Nasal intermittent positive pressure ventilation
PMA	Postmenstrual age
RD	Risk difference

## Table of Contents

1	Introduction .....	5
2	Design parameters .....	5
2.1	The primary and adaptation outcome, and control event rate .....	5
2.2	Justification for the treatment effect .....	5
2.3	Unit of randomisation, clustering due to multiple births per mother, and dropout rate .....	5
2.4	Trial adaptations.....	5
2.5	Interim decision-making rules.....	5
3	Simulation objectives .....	6
4	Key statistical considerations .....	6
5	Simulation approach .....	7
6	Simulation scenarios .....	7
7	Simulation results.....	7
7.1	Impact on decision-making errors .....	7
7.2	Impact on futility early stopping probability.....	8
7.3	Impact on the maximum sample sizes.....	11
7.4	Impact of uncertainty around the control event rate on statistical power.....	13
8	Selected designs, sample size and operating characteristics .....	14
9	Conclusions .....	16
10	References.....	16

## List of Tables

Table 1. Simulation scenarios.....	7
Table 2. Sample sizes and operating characteristics of selected designs.....	15

## List of Figures

Figure 1. The overall probability of stopping early for futility across stages. ....	8
Figure 2. The overall probability of stopping for futility across interim analyses. ....	9
Figure 3. Probability of stopping at the first interim analysis. ....	10
Figure 4. Probability of stopping at the second interim analysis. ....	11
Figure 5. Sample size adjusted for trial adaptations and clustering. ....	12
Figure 6. Sample size adjusted for trial adaptations, clustering and dropouts. ....	13
Figure 7. Impact of uncertainty around control event rate on statistical power. ....	14

## 1 Introduction

This statistical report summarises the statistical simulation work and related results which informed the choice of appropriate adaptive designs that can be used to address the respiratory research question within the TipTop platform.

## 2 Design parameters

### 2.1 The primary and adaptation outcome, and control event rate

The primary outcome is survival without bronchopulmonary dysplasia (BPD) at 36 weeks postmenstrual age (PMA). This is the same adaptation outcome that will be used for interim analyses to inform trial adaptations described in Section 2.4. Based on a Cochrane review <sup>1</sup>, the BPD rate in the nasal Continuous Positive Airway Pressure (nCPAP, control) group ranges from 10% to 37% in similar populations. Following discussions with the clinical team, it was felt that 25% was a conservative estimate of mortality or BPD by 36 weeks PMA. As a result, we expect around 75% survival without BPD in the control group in this population.

### 2.2 Justification for the treatment effect

The nasal intermittent positive pressure ventilation (NIPPV) intervention would need to show a 6.75% absolute increase (9% relative increase) in the event rate of survival without BPD compared to the nCPAP (control) group to claim superiority.

### 2.3 Unit of randomisation, clustering due to multiple births per mother, and dropout rate

Based on advice from PPI partners, infants from the same mother will be randomised to the same intervention. It is expected that 13.8% and 1.2% of pregnancies result in twins and triplets, respectively <sup>2</sup>. This will result in approximately 1.16 mean infants per mother. There is also clustering of outcomes within twins which should be accounted for and the intra-cluster correlation coefficient (ICC) for BPD is expected to be around 0.46 (95% confidence interval: 0.26 to 0.72) <sup>3</sup>. Therefore, we expected a design effect of 1.08 assuming an ICC of 0.5 and mean infants of 1.16 per mother. That is, the sample size for a fixed design or adaptive design will need to be inflated by 8%. Finally, we expected a 5% missing data on the primary outcome.

### 2.4 Trial adaptations

The research team wanted the ability to stop early if the intervention indicates futility. That is, if it is likely to be harmful or unlikely to result in meaningful improvement in survival without BPD. The interim futility decision rules to inform this early stopping are discussed in Section 2.5. Early stopping for efficacy was considered not important as the intervention was unlikely to result in overwhelming evidence required at an interim analysis to trigger early stopping. There is also some uncertainty around the control event rate and as such, sample size re-estimation based on the observed control event rate will also be considered. The impact of this uncertainty on the design such as statistical power is also explored through simulation as described in Section 7.4.

### 2.5 Interim decision-making rules

Choosing the appropriate timing and frequency of interim analyses and interim decision rules to trigger trial adaptations is critical to the successful delivery of an adaptive design. As illustrated in the design of the cord clamping research question ([10.15131/shf.data.25393327](https://www.shf.data.ox.ac.uk/shf-data-25393327)), the value of interim analysis diminishes as the number of interim analyses increases above two when low futility thresholds (e.g., around 0 critical value) are used. Of note, most adaptive trials <sup>4-6</sup>, especially those with early stopping options are designed with 1 or 2 interim analyses. For this research question, two interim analyses are

considered when between 40% and 75% information fraction is accrued as informed by lessons from the cord clamping research question ([10.15131/shef.data.25393327](https://doi.org/10.15131/shef.data.25393327)) and statistical considerations discussed in Section 4. Non-binding futility stopping rules are considered for flexibility in the interim decision-making process considering the totality of the evidence.

There is a trade-off between the choice of decision rules and the value of an adaptive design. For example, lowering the bar of the level of evidence required to trigger early stopping would increase the chance of early stopping; however, this will be at the expense of increasing the chances of making incorrect decisions. It is, therefore, important to quantify this trade-off and select appropriate interim decision rules that balance the robustness of the design in addressing research questions and the benefits of an adaptive design under some known underlying treatment effects.

Of note, discussions with the clinical team highlighted that NIPPV intervention is unlikely to be harmful, cheap and easy to implement in practice. As such, the clinical team preferred a design that minimises the probability of stopping early for futility when the treatment effect is small to moderate as these small benefits may still be clinically useful even though the trial may not be powered for such benefits for feasibility reasons. Statistical simulations are therefore used to inform the choice of robust futility early stopping rules by the research team.

### 3 Simulation objectives

As highlighted in Sections 2.4 and 2.5, the key specific objectives for simulations are to help the research team:

- 1) decide on interim decision rules for futility early stopping,
- 2) quantify the risks in the decision-making process under different scenarios of the underlying treatment effect,
- 3) investigate the impact of uncertainty around the control event rate to note regions when sample size re-estimation may be required.

### 4 Key statistical considerations

Unlike the cord clamping and impacted fetal head research questions where the primary outcomes are immediate and their timing (i.e., the start of the trial clock) is consistent across participants, the timing of the primary outcome for this research question depends on PMA at delivery. That is, infants born early (in terms of gestational age) will take longer for their primary outcome data to mature and contribute to the interim analyses. Conversely, primary outcome data of infants born late will mature quickly to contribute to the interim analyses. The impact of this on interim decision-making is unknown. For example, it is unlikely to be an issue to bias interim decisions if the treatment effect is consistent across gestational age. Otherwise, interim analyses will be dominated by infants born late and interim decisions may be incompatible with the population of infants born early. Mitigating measures may include:

- a) delaying the first interim analysis to increase the representation of infants by gestational age at interim analyses,
- b) stratify randomisation by gestational age (where possible) to increase the balance in baseline characteristics between treatment groups concerning gestational age at interim analyses,
- c) independent monitoring of trial population representation by gestational age at interim analyses,
- d) independent monitoring of the treatment effect stratified by gestational age at interim analyses and,

- e) use of non-binding early stopping futility rules.

## 5 Simulation approach

Sample sizes for competing designs using design parameters described in Section 1 and the timing, frequency and decision rules described in Section 6 were estimated in R using version 3.4.0 of “rpact” package <sup>7</sup>. We adopted a similar simulation approach used for the cord clamping research question and these two research questions share design similarities including the nature of the primary outcome and futility early stopping as a key trial adaptation ([10.15131/shef.data.25393327](https://doi.org/10.15131/shef.data.25393327)). Furthermore, the simulations also explored the impact of uncertainty around the control event rate. The simulation scenarios considered are summarised in Section 6. Metrics to evaluate the performance of competing designs are described in the cord clamping simulation report ([10.15131/shef.data.25393327](https://doi.org/10.15131/shef.data.25393327)).

All simulations were performed in R using version 3.4.0 of “rpact” package <sup>7</sup> and the simulation code is accessible via the [GitHub](#) repository. To achieve a reasonably small Monte Carlo error within a feasible computational time,  $10^5$  simulation replicates were used for each simulation scenario. For selected proposed designs in Section 8,  $10^6$  simulation replicates were used.

## 6 Simulation scenarios

Statistical power of 90%, 2.5% one-sided type I error, 75% event rate, and equal allocation were fixed design parameters across all simulations. The simulation scenarios considered are summarised in Table 1. The underlying treatment effect varied from -2.5% to 10.5% capturing the level of evidence supporting harm to the overwhelming benefit of NIPPV. To explore the impact of uncertainty of statistical power, the control event rate was varied from 64% to 80% while keeping the targeted effect size of 6.75% risk difference (RD) constant across simulation scenarios.

Table 1. Simulation scenarios.

Design aspect	Scenarios
Timing of interim analyses (1 <sup>st</sup> , 2 <sup>nd</sup> )	(0.40, 0.60), (0.4, 0.65), (0.45, 0.65), (0.45, 0.70), (0.50, 0.70), (0.5, 0.75)
Interim decision rules on critical value scale (1 <sup>st</sup> , 2 <sup>nd</sup> )	(0, 0), (0, 0.1), (0, 0.2), (0, 0.3), (0, 0.4), (0, 0.5), (0, 0.6), (0, 0.7)
Underlying treatment effect on risk difference scale	-2.5%, -1%, 0%, 2.5%, 5%, 6.75%*, 8.0%, 10.0%

A critical value of 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7 corresponds to a one-sided p-value of 0.5, 0.4602, 0.4207, 0.3821, 0.3446, and 0.3085, 0.2743, and 0.242 respectively. \*A 6.75% risk difference was the targeted treatment effect.

## 7 Simulation results

### 7.1 Impact on decision-making errors

As evident and expected from Figure 1, the competing designs preserve the desired power and type 1 error of 2.5% (one-sided). That is, the chances of making correct and incorrect decisions about efficacy are preserved as desired. This is expected as the maximum sample sizes for these designs are different and calculated to preserve these decision errors.

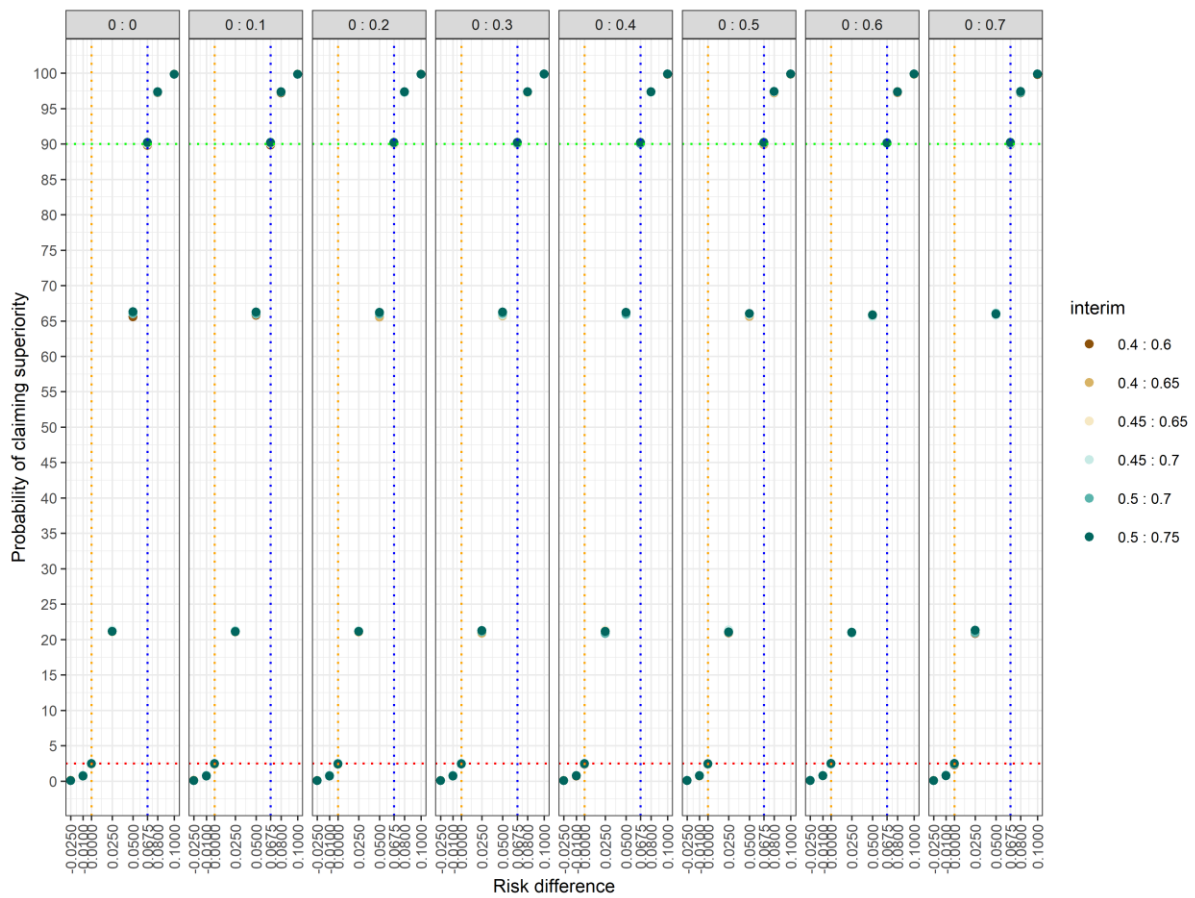


Figure 1. The overall probability of stopping early for futility across stages.

## 7.2 Impact on futility early stopping probability

The overall probabilities of futility early stopping across stages (interim analyses) are displayed in Figure 2. The probabilities for stopping at the first and second interim analyses are shown in Figure 3 and Figure 4. In summary, both futility interim analyses are valuable and the probabilities of stopping increase with increasing futility threshold at the second interim analysis. However, this increase is also at the expense of increasing the probability of stopping in regions where the underlying treatment effect is small to moderate. Finally, the performance of competing designs is comparable except that designs with delayed first interim analysis minimise the probability of stopping when the underlying treatment effect is small to moderate, which is preferred by the clinical team.



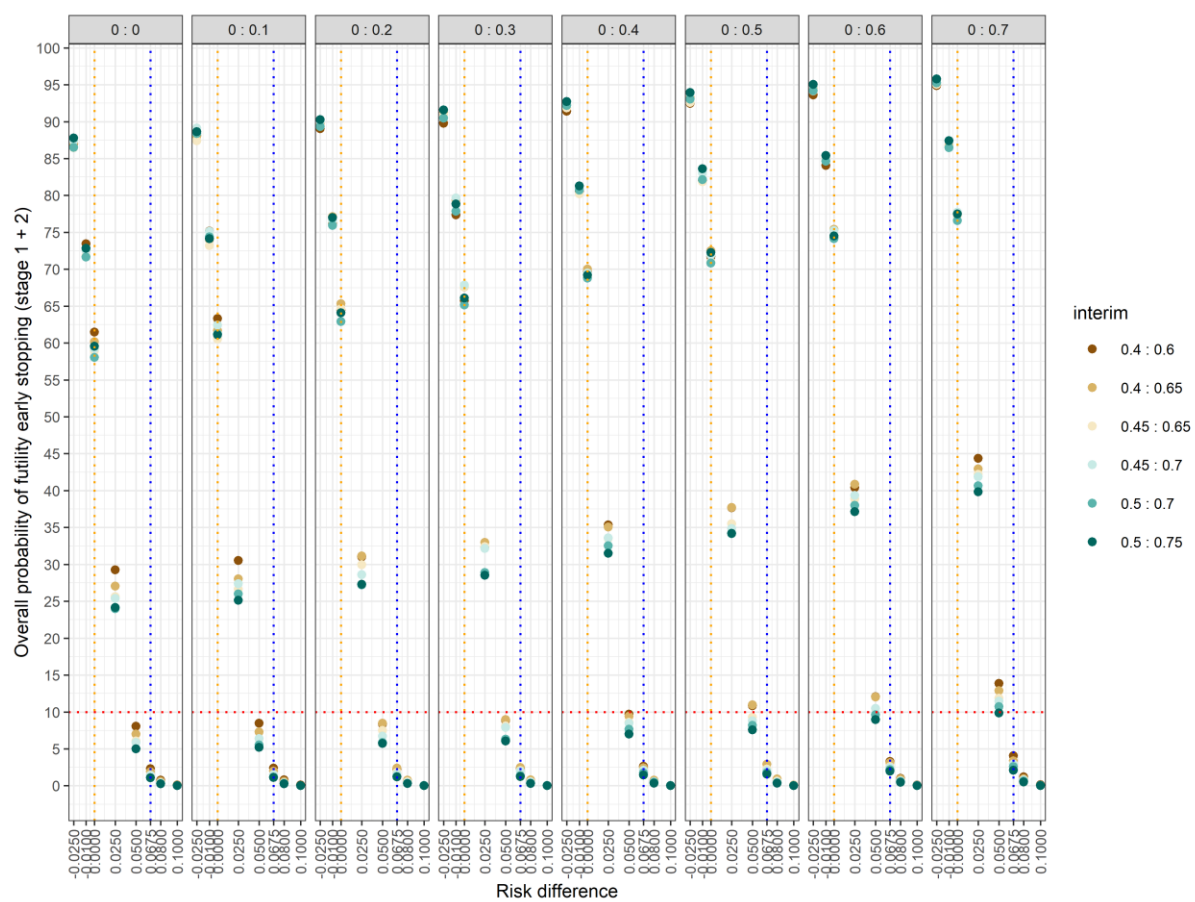


Figure 2. The overall probability of stopping for futility across interim analyses.

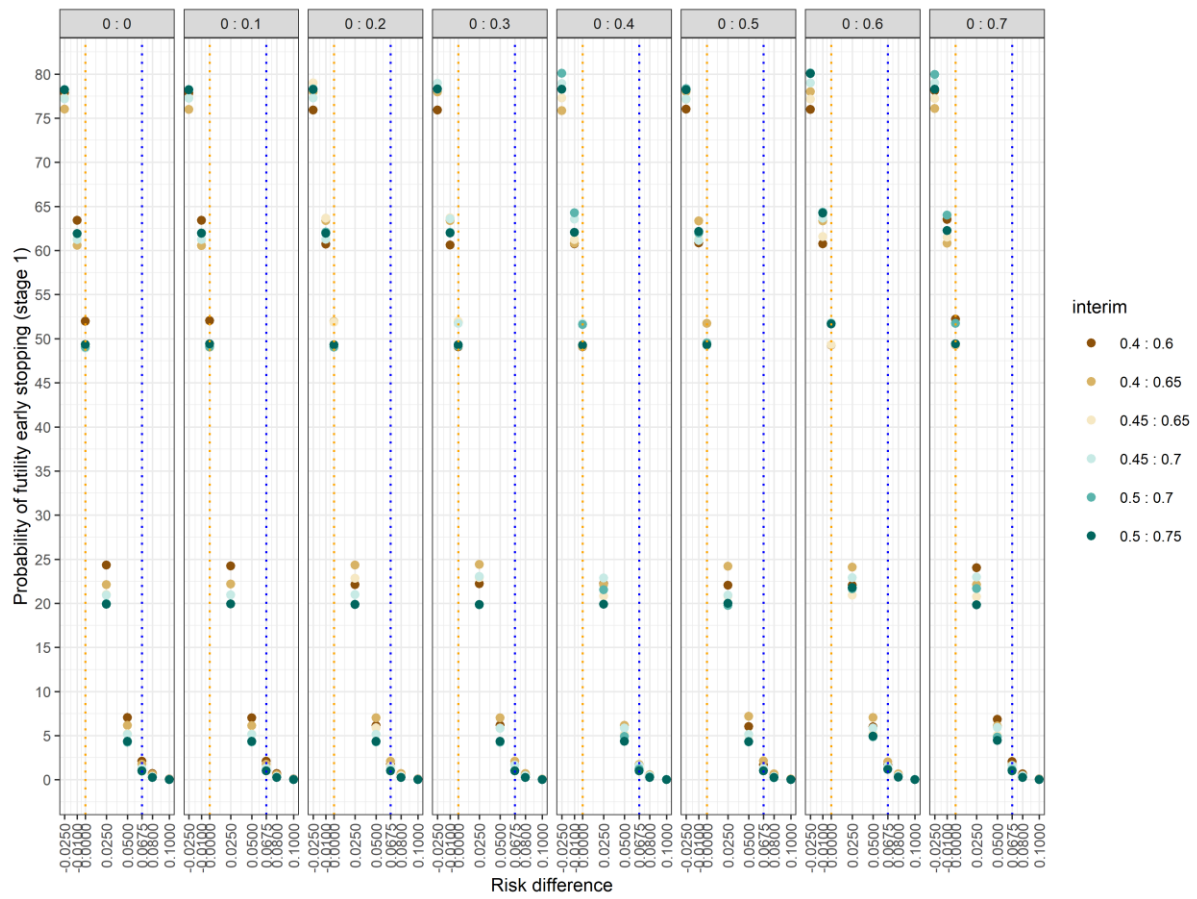


Figure 3. Probability of stopping at the first interim analysis.

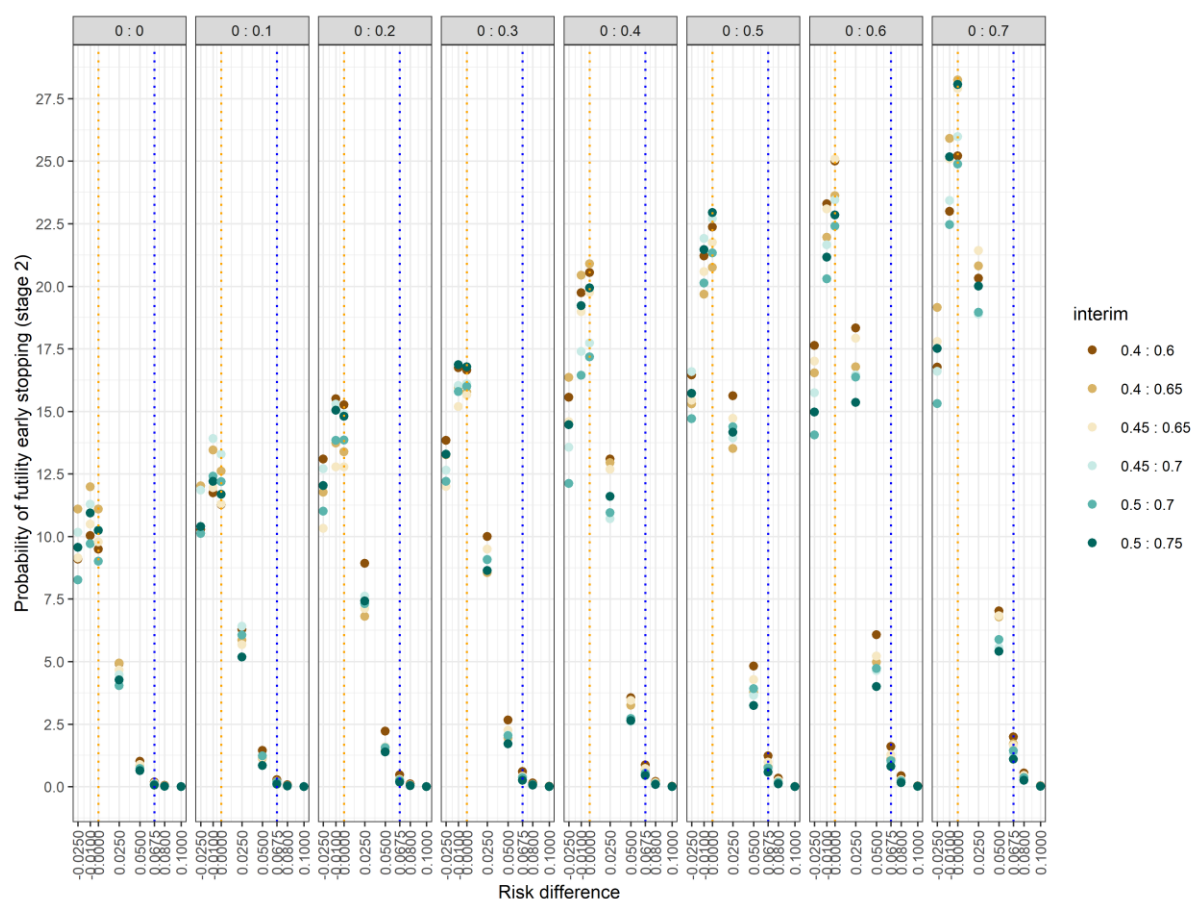


Figure 4. Probability of stopping at the second interim analysis.

### 7.3 Impact on the maximum sample sizes

The overall sample sizes required after accounting for trial adaptations and clustering (design effect of 1.8) are shown in Figure 5. Sample sizes accounting for trial adaptations, clustering (design effect of 1.08), and 5% dropout rate are presented in Figure 6. In general, the total sample size increases with increasing futility threshold for the second interim analysis while the futility threshold for the first interim analysis is held constant. The sample sizes are similar when the second futility threshold is close to that of the first interim analysis. Finally, designs where the first interim analysis is conducted earlier require larger sample sizes.

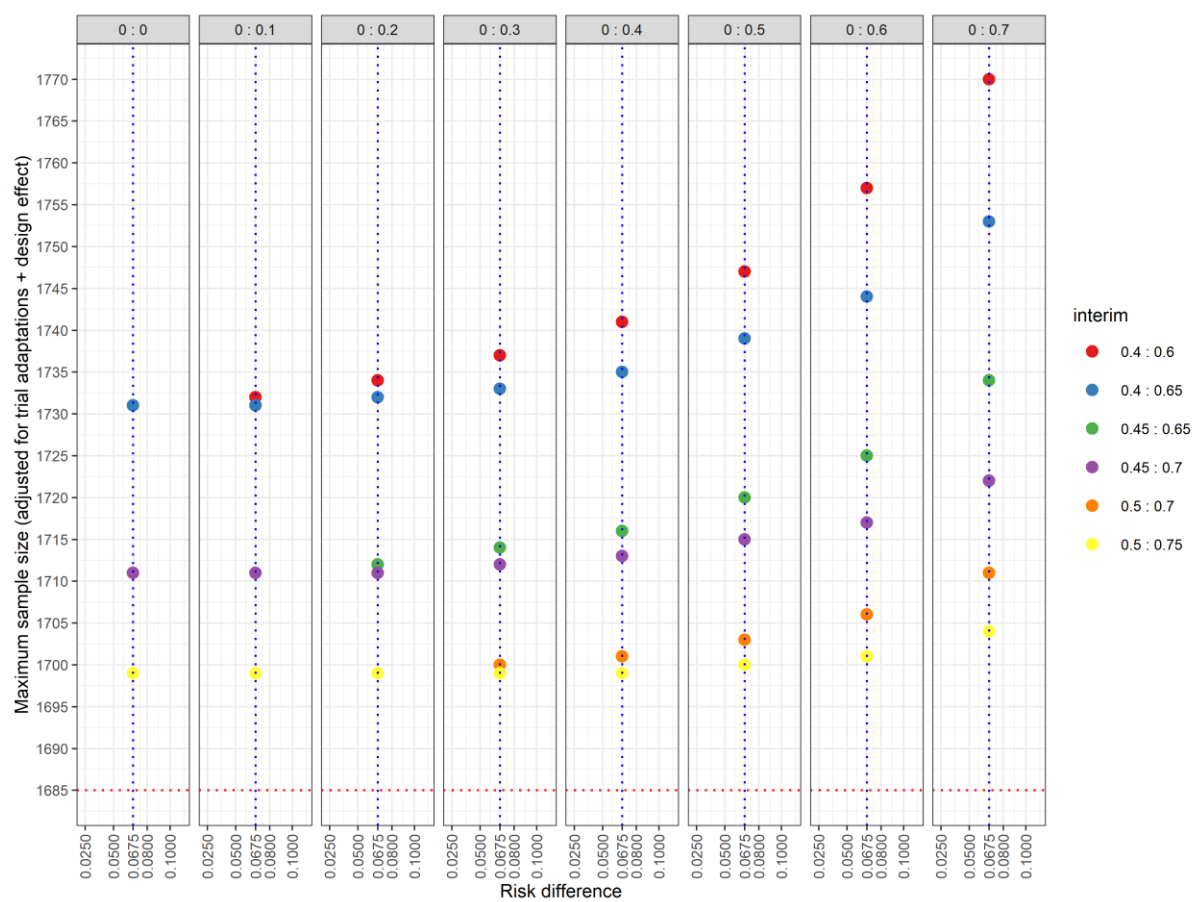


Figure 5. Sample size adjusted for trial adaptations and clustering.

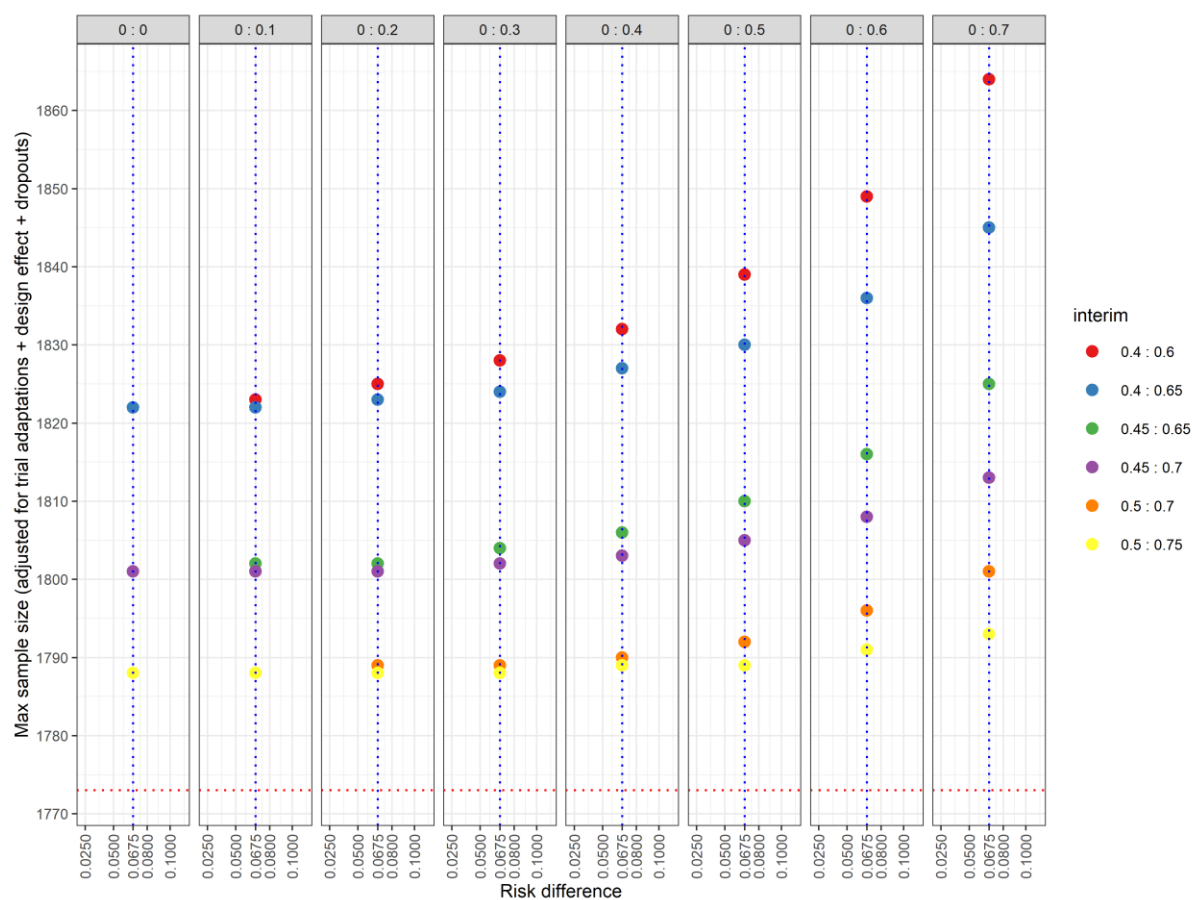


Figure 6. Sample size adjusted for trial adaptations, clustering and dropouts.

#### 7.4 Impact of uncertainty around the control event rate on statistical power

As evident in Figure 7, the impact of uncertainty around the event rate of survival without BPD in the nCPAP (control) arm on statistical power is similar across competing designs (i.e., timing of interim analyses and futility decision rules) considered. The power drops to lower than 85% if survival without BPD in the control arm falls below around 70%; as such, if the interest is in preserving at least 85% power, sample size re-estimation will be required based on observed control event rate if it falls within this region. On the other hand, if the interest is to preserve the 90% power or as close as possible, sample size re-estimation at the first interim analysis can be triggered if the observed control event rate falls below 74% or is at least 77.5%; that is, sample size will need to be increased or reduced, accordingly. An increase in sample size will be conditional on feasibility and funding extension. Finally, changes to the original sample size after the first interim analysis may require updating the timing of the second interim analysis.

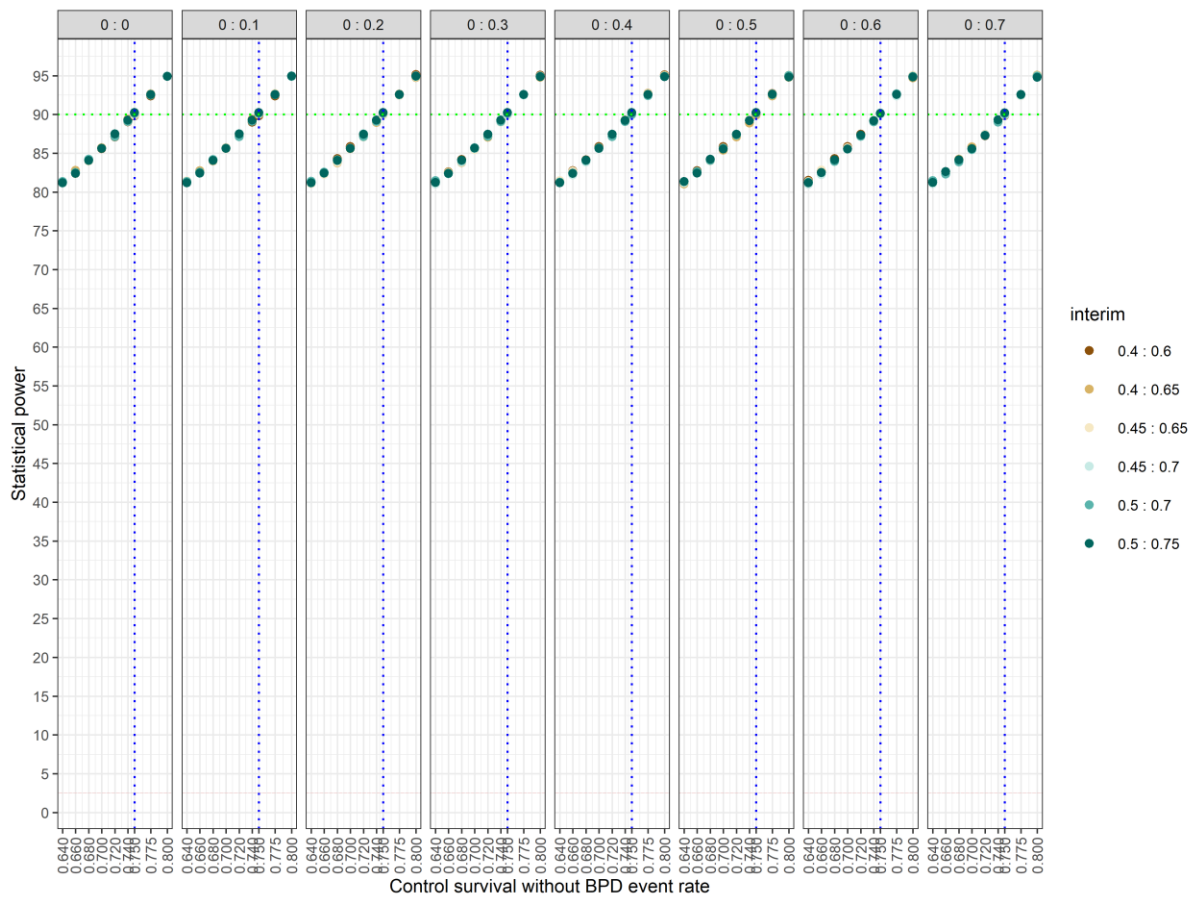


Figure 7. Impact of uncertainty around control event rate on statistical power.

## 8 Selected designs, sample size and operating characteristics

It was clear during discussions with the clinical team that the goal of preferred designs would be to minimise the probability of stopping early when treatment is small to moderate while achieving a high probability of stopping early when the treatment is harmful. Delaying the first interim analysis and using low futility thresholds at both interim analyses achieve this while mitigating a statistical concern discussed in Section 4. As such, designs with futility thresholds of 0 at both interims (i.e., only stop if NIPPV is the same as or worse compared to nCPAP) and delayed first interim analysis (i.e., at 50%), which are presented in Table 2 are better options.

The sample sizes required for design with interim analyses are conducted at (50%, 70%) and (50%, 75%) are the same (Table 2 and Figure 6). With equal allocation, a total sample size (rounded upwards to the nearest even number) of 1700 (850 per arm) adjusted for interim analyses and clustering will be required to preserve a 90% power and 2.5% one-sided type 1 error. This assumes stopping early for futility at the first or second interim analysis if the effect of NIPPV is the same or worse than CPAP. This is equivalent to observing a one-sided p-value of at least 0.5, a critical value of less than or equal to 0, or RD of less than or equal to 0 at each interim analysis. If we adjust for an expected 5% dropout rate, a total sample size of 1788 (894 per group) will be needed. If interim analyses are performed at 50% and 75% and the effect of NIPPV is the same as nCPAP (red row), there is a 49.2% and 10.2% probability of stopping early for futility at the first and second interim analyses, respectively (i.e., 59.4% overall probability). As expected, there is a 90.1% chance of declaring superiority of NIPPV if a targeted 6.75% RD is observed (green row) with only a 1% chance of stopping early for futility.

Table 2. Sample sizes and operating characteristics of selected designs.

nCPAP event rate	NIPPV event rate	RD	Information fraction at interim analyses (1 <sup>st</sup> : 2 <sup>nd</sup> )	Futility decision thresholds at interim analyses (1 <sup>st</sup> : 2 <sup>nd</sup> )			Probability of stopping at interim analyses (1 <sup>st</sup> : 2 <sup>nd</sup> )	Overall probability of futility stopping	Power	Fixed design sample size without & with clustering adjustment, respectively	Sample sizes without & with clustering adjustment at interim & final analyses, respectively		
				Critical value	P-value	RD					1 <sup>st</sup>	2 <sup>nd</sup>	Final
0.75	0.7250	-0.0250	0.5: 0.7	0: 0	0.5: 0.5	0: 0	78.2%: 8.4%	86.6%	0.1%	1560: 1685	787: 850	1101: 1189	1573: 1699
0.75	0.7400	-0.0100					61.9%: 9.7%	71.5%	0.8%				
0.75	0.7500	0.0000					49.2%: 9.0%	58.2%	2.5%				
0.75	0.7750	0.0250					19.9%: 4.1%	24.0%	21.1%				
0.75	0.8000	0.0500					4.4%: 0.7%	5.1%	65.9%				
0.75	0.8175	0.0675					1.0%: 0.1%	1.1%	90.1%				
0.75	0.8300	0.0800					0.3%: 0.0%	0.3%	97.4%				
0.75	0.8500	0.1000					0.0%: 0.0%	0.0%	99.9%				
0.75	0.7250	-0.0250	0.5: 0.75	0: 0	0.5: 0.5	0: 0	78.2%: 9.6%	87.8%	0.1%	1560: 1685	787: 850	1180: 1274	1573: 1699
0.75	0.7400	-0.0100					61.9%: 11.0%	72.9%	0.8%				
0.75	0.7500	0.0000					49.2%: 10.2%	59.4%	2.5%				
0.75	0.7750	0.0250					19.9%: 4.4%	24.2%	21.1%				
0.75	0.8000	0.0500					4.4%: 0.7%	5.1%	65.9%				
0.75	0.8175	0.0675					1.0%: 0.1%	1.1%	90.1%				
0.75	0.8300	0.0800					0.3%: 0.0%	0.3%	97.4%				
0.75	0.8500	0.1000					0.0%: 0.0%	0.0%	99.9%				

RD, risk difference, total sample size adjusted for adaptation, design effect and dropouts is 1788 (894 per group).

## 9 Conclusions

The statistical performances of competing designs have been explored through simulations. These simulations together with discussions with the clinical team helped to narrow down options of robust adaptive designs to achieve specific goals. Simulations were important to the research team to understand the implications of the competing designs, interim decision rules and timing of interim analyses on decisions. Whilst the statistical performances of the timing of interim analyses between 40% and 75% were comparable, designs with delayed first interim analysis and with low futility thresholds are both interim analyses are preferred in this context. In addition, these designs yield smaller maximum sample sizes although the differences in sample sizes between designs are relatively small.

A noteworthy limitation of this work is that the statistical impact of potential overlap in trial populations between the cord clamping and respiratory research questions is unknown as there are no prior data to draw any inference and inform any related simulation work if needed. However, the clinical team believes that this overlap in trial populations will be very small and negligible. This is something that should be monitored during the trial and statistical modelling alongside the trial can be used to explore implications.

## 10 References

1. Lemyre B, Deguise MO, Benson P, Kirpalani H, Ekhuagere OA, Davis PG. Early nasal intermittent positive pressure ventilation (NIPPV) versus early nasal continuous positive airway pressure (NCPAP) for preterm infants. *Cochrane Database Syst Rev*. 2023;2023(7). doi:10.1002/14651858.CD005384.PUB3/MEDIA/CDSR/CD005384/IMAGE\_N/NCD005384-CMP-002.04.SVG
2. Dorling J, Hewer O, Hurd M, et al. Two speeds of increasing milk feeds for very preterm or very low-birthweight infants: the SIFT RCT. *Health Technol Assess (Rockv)*. 2020;24(18):1-94. doi:10.3310/HTA24180
3. Yelland LN, Sullivan TR, Collins CT, Price DJ, McPhee AJ, Lee KJ. Accounting for twin births in sample size calculations for randomised trials. *Paediatr Perinat Epidemiol*. 2018;32(4):380-387. doi:10.1111/PPE.12471
4. Stevely A, Dimairo M, Todd S, et al. An Investigation of the Shortcomings of the CONSORT 2010 Statement for the Reporting of Group Sequential Randomised Controlled Trials: A Methodological Systematic Review. Shamji M, ed. *PLoS One*. 2015;10(11):e0141104. doi:10.1371/journal.pone.0141104
5. Wang Y, Yao M, Liu J, et al. A systematic survey of adaptive trials shows substantial improvement in methods is needed. *J Clin Epidemiol*. 2024;167. doi:10.1016/j.jclinepi.2024.111257
6. Zhang J, Saju C. A systematic review of randomised controlled trials with adaptive and traditional group sequential designs – applications in cardiovascular clinical trials. *BMC Med Res Methodol* 2023 231. 2023;23(1):1-9. doi:10.1186/S12874-023-02024-1
7. Wassmer G, Pahlke F. RPACT Package Overview | RPACT. Accessed April 15, 2019. <https://www.rpact.org/>