Original Research Article

A seamless Phase I/II platform design with a time-to-event efficacy endpoint for potential COVID-19 therapies

Thomas Jaki[1,2] (iD)**, Helen Barnett[3]** (iD)**, Andrew Titman[3] and Pavel Mozgunov[2]** (iD)

Abstract

In the search for effective treatments for COVID-19, the initial emphasis has been on re-purposed treatments. To maximize the chances of finding successful treatments, novel treatments that have been developed for this disease in particular, are needed. In this article, we describe and evaluate the statistical design of the AGILE platform, an adaptive randomized seamless Phase I/II trial platform that seeks to quickly establish a safe range of doses and investigates treatments for potential efficacy. The bespoke Bayesian design (i) utilizes randomization during dose-finding, (ii) shares control arm information across the platform, and (iii) uses a time-to-event endpoint with a formal testing structure and error control for evaluation of potential efficacy. Both single-agent and combination treatments are considered. We find that the design can identify potential treatments that are safe and efficacious reliably with small to moderate sample sizes.

Keywords

1 Introduction

The emergence of COVID-19 and the ensuing pandemic has led to a widespread, frantic, search for treatments. Despite large uncertainty about the underlying pathogen and the natural history of the disease, trials must start rapidly to identify treatments to save lives, but also so that effective treatments can be used in the response to the outbreak. A consequence of this is that trials in COVID-19 in the first few weeks and months of the outbreak have focused on re-purposed treatments.[1–3]

While recently, some success with using re-purposed treatments has been achieved,[4–6] it is crucial that development of treatments specifically developed for COVID-19 is also undertaken in order to maximize the chances of finding therapies to successfully treat patients. The crucial difference of trials investigating novel therapies (in contrast to re-purposed treatments) is that the range of safe and likely effective doses is unknown. Therefore, an efficient dose-finding design identifying safe and active doses to be studied in larger trials is essential. While there exist a number of dose-finding designs for early phase dose-finding trials evaluating toxicity and efficacy simultaneously, for example, Wages and Tait[7] and Mozgunov and Jaki[8] and references therein, many of them consider a binary efficacy endpoint with few recent extension to other endpoints.[9–11] Time-to-event endpoints with censoring at 28 days have previously been used as a clinically meaningful measure in a number of COVID-19 trials[1,2,4] and the argument has been made that they should be considered in all COVID-19 trials.[12]

[1]Faculty for Informatics and Data Science, University Regensburg, Germany
[2]MRC Biostatistics Unit, University of Cambridge, UK
[3]School of Mathematical Sciences, Lancaster University, UK

Corresponding author:
Thomas Jaki, Faculty for Informatics and Data Science, University Regensburg, Regensburg, Germany.
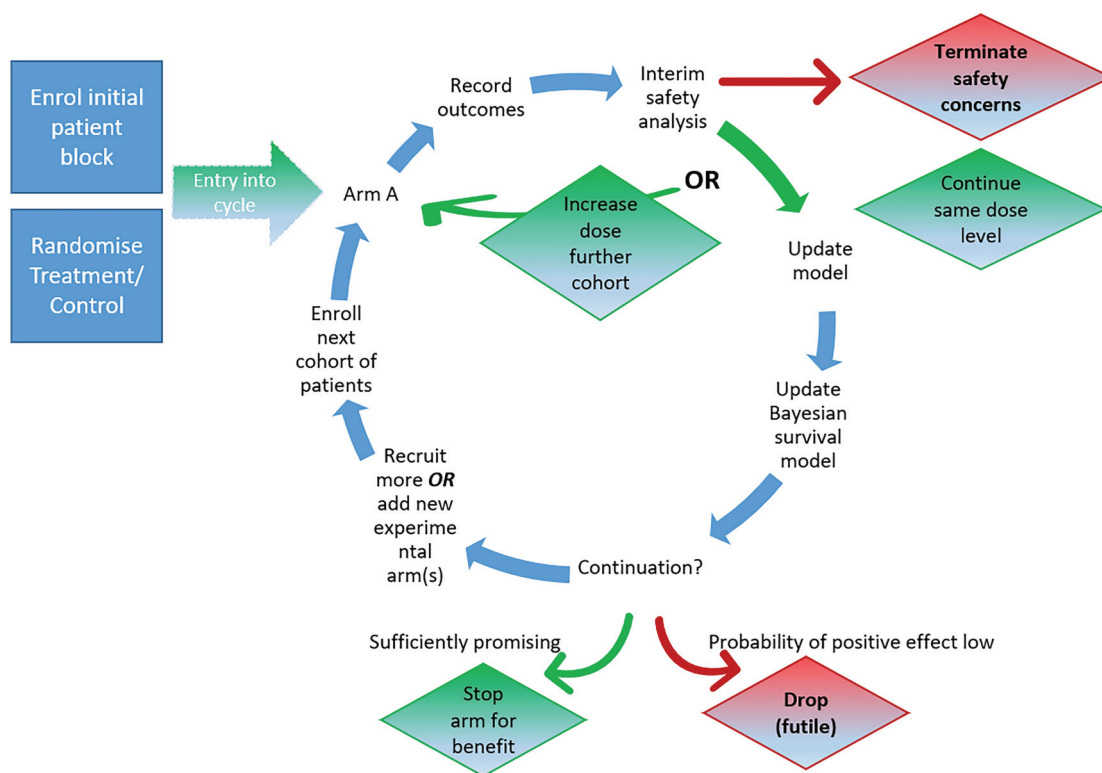Email: thomas.jaki@ur.de

Figure 1. Illustration of the AGILE platform design.

While the majority of Phase I dose-finding trials, particularly in oncology, are non-randomized, it is agreed that in later phases, the gold standard for evaluating novel treatments are well-conducted blinded randomized controlled clinical trials. At the same time, in light of the uncertainty about the symptoms caused by COVID-19 – especially at the beginning of the pandemic – it is essential to conduct randomized dose-finding trials to ensure that the risk of adverse events is correctly attributed to the drug under study rather than to the disease itself. Moreover, it has been argued that adaptive designs[13,14] are particularly suitable during a pandemic, also in the light of the uncertainty about a novel disease.[15] Therefore, a randomized adaptive dose-finding design evaluating both toxicity and time-to-event efficacy would allow to answer the research questions of interest in novel therapies for treating COVID-19.

It is also recognized that there are a number of novel therapies that have the potential to be efficient in fighting COVID-19. Therefore, it is crucial to have a structure in place that allows rapid enrolment of novel therapies to ensure rapid decision-making, and, importantly, would allow for efficient use of information between the studies, that is, utilizing the data from the control treatment across different compounds. This can be achieved via a platform trial.[16]

In this paper, we describe and evaluate the bespoke design developed and implemented for the AGILE platform,[17] an adaptive randomized seamless Phase I/II dose-finding trial platform that seeks to quickly establish a safe range of doses and investigates treatments for potential efficacy using a Bayesian sequential trial design (see a visualisation of the design for one compound in Figure 1). The proposed design is unique as it

(i) utilizes randomization during dose-finding to allow COVID-19 induced symptoms to be distinguished from drug side-effects,

(ii) shares control arm information across the platform in order to maximize efficiency, and

(iii) uses a time-to-event endpoint with a formal testing structure and error control for evaluation of potential efficacy,

making the design particularly suitable for the pandemic setting. We also extend the design for trials studying dual-agent combinations of treatments.

The rest of the article is organized as follows. Section 2 describes the platform for single treatments while its performance is evaluated in simulations in Section 3. The design for dual-agent combinations is proposed in Section 4 and subsequently evaluated in Section 5. We conclude with a discussion (Section 6).

2 Single-agent design

2.1 Setting

Consider a randomized controlled dose-escalation clinical trial in which m increasing doses $d_1 < d_2 < \cdots < d_m$ of a single experimental treatment are studied. Let $d_0 = 0$ be a dose of zero of the treatment, which is subsequently referred to as the *control arm* (or, simply control). The inclusion of the control arm is motivated by the emerging nature of the symptoms associated with COVID-19 when planning the study and the desire to avoid labelling potential treatments as unsafe due to misclassifying non-treatment-related symptoms. Note that information on the control arm can come from either patients randomized to control for this particular candidate treatment within the platform, or from the control arm of other candidates within the platform provided that they are concurrent. A binary outcome of a random variable Y, $y = 0$ is observed if no dose-limiting event (DLE) is observed within $t_{\text{safe}} > 0$ days after randomization, and $y = 1$, otherwise. Let p_j be the probability for a patient to experience a DLE if given dose d_j. It is assumed that the risk of DLE is a non-decreasing function of dose, $p_0 \leq p_1 \leq \cdots \leq p_m$ and prior information for the DLE probability of the control arm, p_0, is available.

As it is expected that the control arm is associated with a non-negative (unknown) risk of DLE (or symptoms of the disease that cannot be distinguished from DLEs), the primary goal of the dose-escalation is formulated in terms of the additional risk of a dose-limiting event (ADLE) defined in terms of the expected difference in DLE risk between the doses of the agent and the control. Specifically, we therefore seek to identify the dose that corresponds to an additional risk of $\gamma = 0.20$ which equates to finding the dose $d_{j^\star}$ such that

$$j^\star = \arg \min_{j=0,\dots,m} | (p_j - p_0) - \gamma |$$

2.2 Bayesian dose-escalation model

The following randomized Bayesian dose-escalation design that builds on the proposal by Mozgunov et al.[18] is used. Assume that the DLE probability has the functional form

$$\varphi(\tilde{d}_j, \theta_1, \theta_2) = \frac{\exp(\theta_1 + \theta_2 \tilde{d}_j)}{1 + \exp(\theta_1 + \theta_2 \tilde{d}_j)} \tag{1}$$

where θ_1 and θ_2 are unknown parameters, and $\tilde{d}_j$ is a standardized dose level (also referred to as a skeleton) corresponding to dose j, which is constructed given the information about the prior DLE toxicities (details are given below). We require that $\theta_2 > 0$ and enforce this constraint through the construction of the prior distribution (see Section 3.3.1). This model choice was found to result in good statistical properties in terms of the target dose identification in a randomized dose-finding trial.[18] We require that the standardized dose level corresponding to control is equal to $\tilde{d}_0 = 0$. This will guarantee that a sequential update of the slope parameter θ_2 will not contribute to the DLE probability estimation on the control arm yet all data are used for its estimation.[18]

Denote the prior distribution of the vector $\theta = (\theta_1, \theta_2)$ by $f_0(.)$. To construct the standardized levels, $\tilde{d}_j$, we represent them in terms of prior estimates of the DLE probabilities $\hat{p}_j^{(0)}$ associated with doses d_j $j = 0, \dots, m$

$$\tilde{d}_j = \frac{\text{logit}(\hat{p}_j^{(0)}) - \hat{\theta}_1^{(0)}}{\hat{\theta}_2^{(0)}} \tag{2}$$

where $\hat{\theta}_1^{(0)}$ and $\hat{\theta}_2^{(0)}$ are prior point estimates of the model parameters, and $\text{logit}(x) = \log(x/(1-x))$ is the logit transformation of x. To satisfy $\tilde{d}_0 = 0$, the prior needs to be chosen such that $\text{logit}(\hat{p}_0^{(0)}) = \hat{\theta}_1^{(0)}$.

Assume that n patients, potentially including concurrent control patients from other evaluations within the platform, have already been assigned to doses $\tilde{d}(1), \dots, \tilde{d}(n)$ and binary responses $\mathbb{Y}_n = [y_1, \dots, y_n]^{\text{T}}$ were observed, respectively. The model updates the posterior distribution of θ using Bayes' theorem

$$\begin{aligned} f_n(\theta) &= \frac{f_{n-1}(\theta)\phi(\tilde{d}(n), y_n, \theta)}{\int_{\mathbb{R}^h} f_{n-1}(u)\phi(\tilde{d}(n), y_n, u)\mathrm{d}u} \\ &= \frac{f_0(\theta) \prod_{i=1}^{n} \phi(\tilde{d}(i), y_i, \theta)}{\int_{\mathbb{R}^h} f_0(u) \prod_{i=1}^{n} \phi(\tilde{d}(i), y_i, u)\mathrm{d}u} \end{aligned} \tag{3}$$

where $\phi(d(n), y_n, \theta) = \varphi(d(n), \theta)^{y_n}(1 - \varphi(d(n), \theta))^{1-y_n}$. This posterior distribution is then used to make the escalation/de-escalation decision.[19] Specifically, the first set of safe doses is defined as the doses d_j for which

$$\mathbb{P}\left(p_j - p_0 \geq \gamma + 2\delta\right) < c_{\text{overdose}} \tag{4}$$

where γ is the target ADLE risk, δ is the width of the interval of DLE risk which we consider acceptable, c_{overdose} is the threshold controlling overdosing, and the probability is found with respect to the updated posterior distribution. Amongst the safe doses, the dose which maximizes

$$\mathbb{P}\left(p_j - p_0 \in [\gamma - \delta, \gamma + \delta]\right) \tag{5}$$

is selected as the target dose.

2.3   Efficacy design

2.3.1   Bayesian efficacy model

In this study, we assess the potential efficacy of the treatment for a particular dose instead of modelling efficacy across all doses. Although other approaches are possible, our approach allows us to make conclusions about a given dose alone without sharing information from other arms and enables control of the type I error for the assessment of a given dose. A Cox proportional hazards model is assumed where the hazard of recovery at time t is given by $h(t; z) = h_0(t; z)\psi^z$, where $z = 1$ corresponds to a treatment and $z = 0$ to control. We use a two-point prior for ψ (detailed below) for computational efficiency and a parametric model is assumed in order to maximize power in the light of small sample sizes and in the absence of a clear understanding of potential reasons for deviation from the proportionality assumptions. We study the impact of violating this assumption in a sensitivity analysis presented in the supplemental materials. Initially, the cohort of patients who have graduated from the dose-escalation stage are followed up for a total of t_{eff} days. Based on their outcomes, a decision is made to either stop for futility, stop for efficacy or recruit a further cohort of patients. To improve power, controls recruited from other candidate treatments or other doses of the same treatment are also used within the evaluation, but this is restricted to using only the most recent, concurrent, n_c such controls to mitigate the risk of bias due to population drift.

A Bayesian criterion is adopted for the stopping rule at each stage. In line with Bayesian thinking, we set the stopping rules to be the same for each stage k. Specifically denoting all data up to stage k on dose j by $\mathcal{D}_k^{(j)}$ and for a given desirable treatment effect, ψ^*,

- evaluation is stopped for efficacy if $\pi_{E|k}^{(j)} := \mathbb{P}(\psi = \psi^* \mid \mathcal{D}_k^{(j)}) > u$,
- evaluation is stopped for futility if $\pi_{E|k}^{(j)} < l$, or
- an additional cohort of patients is recruited, otherwise.

In order to ensure a decision is made at the final stage $k = K$, efficacy for dose d_j is established if $\pi_{E|K}^j > u$ and is considered futile otherwise. This is in line with traditional group-sequential designs where the lower and upper boundaries are made equal at the final analysis.

A point prior of the form $\mathbb{P}(\psi = 1) = 1 - \pi_E^{(j)}$, $\mathbb{P}(\psi = \psi^*) = \pi_E^{(j)}$ is assumed for ψ. Here $\pi_E^{(j)}$ represents the degree of optimism or scepticism towards the likely efficacy of dose d_j.

An advantage of the point prior is that obtaining the posterior probability $\pi_{E|k}^{(j)}$ is computationally straightforward allowing comprehensive evaluations of the design via simulations. The posterior probability under this model is

$$\pi_{E|k}^{(j)} = \frac{\pi_E^{(j)} L(\psi^* \mid \mathcal{D}_k^{(j)})}{\pi_E^{(j)} L(\psi^* \mid \mathcal{D}_k^{(j)}) + (1 - \pi_E^{(j)})L(1 \mid \mathcal{D}_k^{(j)})} \tag{6}$$

where $L(u \mid \mathcal{D}_k^{(j)})$ is the Cox partial likelihood with respect to the data for dose j up to the kth period evaluated at a hazard ratio of u. While the Cox partial likelihood is often not considered compatible with a Bayesian analysis since it does not use the full information of the data, Bayesian justifications for its use are available.[20,21]

2.3.2 Setting the boundaries

To set the boundaries, (l, u), a large number of trajectories of $L(1 \mid \mathcal{D}_k^{(j)})$ and $L(\psi^* \mid \mathcal{D}_k^{(j)})$ can be simulated under both the null and alternative hypothesis, where in all cases the simulation continues until the maximum period K. An assumption regarding the proportion of patients recovered by time t_{eff} under the null is needed, in addition to the hazard ratio between treatments. In the absence of censoring due to drop-out, the results will be otherwise invariant to the precise survival distribution assumed since the Cox partial likelihood only uses the order of events.

The effect of varying (l, u) can then be explored by converting the pairs of likelihoods into a posterior probability and imposing the boundary-stopping rules. For any given set of boundaries, the type I error, power, expected number of patients under the null, expected number of patients under the alternative and probabilities of stopping for futility or efficacy at each stage, can be approximated. The boundaries can then be set to optimize some criterion, for instance sum of expected sample sizes under the null and alternative, subject to some constraints, for instance controlling type I error, keeping power above some level or limiting the chance of early stopping under the alternative.

The inclusion of historic controls will increase both the power and type I error of any procedure, for example, Schmidli et al.[22] To ensure type I error is controlled for the evaluation of the given dose, the boundaries are set assuming the maximum, n_c, previous controls are available, with the consequence that the type I error will be lower for the first treatment evaluated in the platform. This will also mean the power will be lower for the first few evaluated treatments. However, given the dose-finding design, it is anticipated that the first evaluations will be of less importance, as for safety reasons evaluations tend to start at sub-optimal low doses.

2.4 Overall design

The overall design of the platform allows for multiple different compounds to be evaluated and, by sharing concurrent control group data, efficiency is gained. For any compound in the platform, patients are allocated in cohorts of size $c = c_1 + c_2$, where c_1 is the number of patients in the cohort assigned to an active dose and c_2 is the number of patients in cohort assigned to the control arm, $\tilde{d}_0$, throughout its evaluation. Below is an outline of the overall procedure for one compound made up of both safety and efficacy evaluation:

Safety evaluation

1. The first cohort of $c_1 + c_2$ patients is assigned to the first dose and to the control arm, respectively.
2. After t_{safe} days, short-term DLE outcomes are collected and the posterior distribution of the parameters is updated using equation (4).
3. The set of safe doses is found using equation (4).
   - If no doses are safe, the trial is stopped for safety;
   - if only the current dose is safe, the next cohort of $c_1 + c_2$ patients is assigned to the current dose and to the control arm, respectively;
   - otherwise, the next cohort of $c_1 + c_2$ patients is assigned to the adjacent, safe dose level for which the probability (5) is maximized and control arm, respectively, that is, no skipping of dose levels is allowed.
4. Once efficacy information is available for two cohorts on a safe dose, that dose is graduated to the efficacy evaluation.

Efficacy evaluation

1. If a dose d_j is deemed safe, the efficacy outcome is observed up to day, t_{eff}.
2. The posterior probability, $\pi_{E|k}^{(j)}$, following equation (6) is then computed where k corresponds to the number of times this dose has been evaluated for efficacy.
   - If $\pi_{E|k}^{(j)} < l$ evaluation of dose d_j is stopped for futility;
   - if $\pi_{E|k}^{(j)} > u$ evaluation is stopped for efficacy and the corresponding candidate and dose d_j recommended for further testing;
   - otherwise if $k < K$, an additional cohort of $c_1 + c_2$ patients is recruited on the current dose and control arm, respectively.

The evaluation of a dose continues until the maximum number of patients N on a dose has been reached unless it is stopped for efficacy, futility or safety before. Once all doses are stopped, the evaluation of this candidate stops. Note that this structure implies that different doses of one compound that are deemed safe can independently be evaluated against control in this design. The overall design for one compound is depicted in Figure 1.

3 Evaluations of proposed design

3.1 Setting

We will now evaluate, for one compound, safety and efficacy across the study together in a simulation study and evaluate the impact of shared control data that are gradually accumulated over the course of the trial, thereby assessing the added benefit of the platform structure. To our knowledge there are no alternative approaches that (i) utilize randomization during dose-finding, (ii) assess efficacy using a time-to-event endpoint within a formal testing framework and (iii) employ a platform structure and hence no comparator is presented here. Our sensitivity analysis presented in the supplemental materials, however, provides a comparison with a design that uses a similar design with a binary outcome for efficacy.

We consider the setting where there are three active doses ($m = 3$) of a single agent and the control arm. As before, the DLE endpoint is binary and corresponds to either experiencing or not experiencing a DLE by time $t_{safe} = 7$ and the efficacy endpoint is time-to-improvement defined as a 1-category improvement on the 10-point World Health Organization scale[23] over 28 days, $t_{eff} = 28$. To generate the efficacy outcome a Weibull distribution is used. For the control group, the rate and shape parameters were 0.085 and 0.797, respectively, resulting in a median recovery time of 14 days and a recovery rate of 70% within 28 days. The rate parameter for the candidate dose has been adapted to match the scenarios described below. The binary safety and time-to-event efficacy responses are assumed to be highly correlated and are generated via a procedure described by Mozgunov et al.[24] using a correlation coefficient of $\rho = 0.8$.

The maximum total intake per dose level is 72 patients assigned to each dose level and control, which equates to a maximum total sample size of 216. In line with the real study, cohort sizes of $c_1 = 4$ (assigned to active doses) and $c_2 = 2$ (assigned to the control) are used. Note that imbalanced randomization between active doses and control is used here, as additional control information from different doses of that compound and from concurrently available control data from other compounds is also used in the efficacy assessment. We also report sensitivity analysis to assess the impact of various cohort sizes.

The objective of the trial is to find all safe efficacious doses to be graduated into a larger Phase II or Phase III clinical trial. The target ADLE risk is $\gamma = 0.20$, and the dose is considered safe if the ADLE risk is ≤ 0.30 ($\delta = 0.05$) and is efficacious if it corresponds to a hazard ratio of at least 1.75. The stopping boundaries for efficacy have been found such that the type I error in each pairwise comparison (i.e. for each dose compared to control) is controlled at 10% one-sided, while the power for each individual dose is 80%. The resulting boundary values are $l = 0.224$ and $u = 0.839$ (see Section 3.3.2 for more details).

3.2 Scenarios

As the trial aims to study novel compounds which have yet to be explored with respect to their mechanism of action in COVID-19 patients, it is crucial that the design has good operating characteristics under a variety of dose-DLE and dose-efficacy scenarios. Therefore, we consider five dose-efficacy scenarios ranging from no doses corresponding to a change in time-to-improvement within 28 days to all doses resulting in a clinically significant reduction; and five dose-DLE scenarios ranging from all doses being safe to all doses being very unsafe. We then consider all combinations of these scenarios, resulting in 25 scenarios explored in total. The five dose-DLE and dose-efficacy scenarios for each (d_0, d_1, d_2, d_3) are presented in Table 1.

We will refer to the scenario with dose–DLE relationship x and dose–efficacy relationship y as 'Scenario x to y'. Each dose under the combination of DLE and efficacy scenario is classified as incorrect, undesirable, acceptable or desirable. If a treatment is unsafe or has a hazard ratio of 1 then it is classed as incorrect. If it is safe, then a hazard ratio of 1.25 is undesirable, 1.5 is acceptable and at least 1.75 is desirable.

For all 25 scenarios, a sensitivity analysis is conducted on varying values of c_1 and c_2 in order to assess the effect of both altering the allocation ratio between control and active doses and the total cohort size, c. We also study the implications of

Table 1. Safety and efficacy scenarios for (d_0, d_1, d_2, and d_3).

	Safety (probability of DLE, p_j)	Efficacy (hazard ratio)
Scenario 0	(0.10, 0.12, 0.13, 0.15)	(1.00, 1.00, 1.00, 1.00)
Scenario 1	(0.10, 0.12, 0.15, 0.30)	(1.00, 1.00, 1.75, 1.75)
Scenario 2	(0.10, 0.15, 0.30, 0.45)	(1.00, 1.50, 1.75, 1.75)
Scenario 3	(0.10, 0.30, 0.45, 0.60)	(1.00, 1.50, 1.75, 2.00)
Scenario 4	(0.10, 0.45, 0.60, 0.60)	(1.00, 1.75, 2.00, 2.00)

DLE: dose-limiting event.

not sharing controls across doses. A total of six settings are considered, ($c_1 = 2$ & $c_2 = 1$ and $c_1 = 2$ & $c_2 = 2$ for only the settings where controls are shared, and $c_1 = 4$ & $c_2 = 2$ and $c_1 = 3$ & $c_2 = 3$ for both settings where controls are shared and where they are not). The maximum number of cohorts per dose varies with cohort size in order to maintain the constant maximum total intake per dose level of 72 across that dose level and control.

Software in the form of R code used to produce the presented results is available on GitHub (https://github.com/dose-finding/covid19-agile).

3.3 Choice of design parameters

3.3.1 Safety model

The proposed design requires the prior and design parameters for both safety and efficacy parts to be pre-specified in advance of the conduct of the trial. The procedure of how these parameters were chosen is given below.

The prior parameters for the safety model were obtained via a calibration procedure[25] over a number of safety scenarios (not taking into account efficacy). We use safety Scenarios 1 to 3 in Table 1 that correspond to the target dose being d_3, d_2, and d_1, respectively, thus covering various locations of the target dose on the dosage grid.

The following prior distribution for the vector of safety model parameters θ was assumed:

$$(\theta_1, \log(\theta_2)) \sim \mathcal{N}(\mu, \Sigma)$$

where $\mu = (\mu_1, \mu_2)^{\mathrm{T}}$ is the vector of means and

$$\Sigma = \begin{bmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{12} & \sigma_2 \end{bmatrix}$$

Given the link between the prior toxicity on the control and the intercept parameter θ_1 as implied by equation (2), $\mu_1 = \mathrm{logit}(\hat{p}_0^{(0)})$ where the prior DLE probability at the control, $\hat{p}_0^{(0)}$. Following discussions with the clinical team, the DLE risk on control was set to $\hat{p}_0^{(0)} = 0.10$. To reduce the computational complexity of the calibration, the covariance between the model parameters was assumed to be $\sigma_{12} = 0$. The rest of the parameters were chosen by conducting simulations using various combinations of values of the parameters on the grid, $\mu_2 = \{-0.5, -0.25, 0, 0.25, 0.5\}, \sigma_1 = \{1.2, 1.3, 1.4, 1.5, 1.6\}, \sigma_2 = \{0.15, 0.25, 0.35, 0.45, 0.55\}$.

Furthermore, to define the standardized doses $\tilde{d}_j$ in equation (2), the prior toxicity probability at each dose should be assumed. As there is no reliable information on the DLE rates in the COVID-19 population, the skeleton was also calibrated. The grid of values for the prior toxicity risks is chosen in terms of the difference in the probability of DLE between the neighbouring doses. Specifically, $\hat{p}_j^{(0)} = \hat{p}_0^{(0)} + v \times j, j = 1, 2, 3$ where v is the difference in the toxicity probabilities between doses, which are then used to find the skeleton using equation (2). The grid of values of v was included as one of the parameters for the calibration $v = \{0.075, 0.100, 0.125, 0.150\}$. Below, we fix $c_{\mathrm{overdose}} = 0.25$ that was previously found to result in good safety properties of the design for the two-parameter logistic model.[26]

The calibration was performed as follows. For each combination of parameters of $v, \mu_2, \sigma_1, \sigma_2$ on the specified grid, 500 simulations were run under each of the three considered scenarios monitoring the proportion of target dose selections. Then, the selected values of the parameters are those that maximized the geometric mean (taken across scenarios for the same combination of values of the parameters) of the proportion of the target dose selection. This resulted in using $v = 0.125, \mu_2 = -0.25, \sigma_1 = 1.40, \sigma_2 = 0.35$ for the further design evaluation.

3.3.2 Efficacy model

The efficacy stopping boundaries for a particular setting were taken as the pair (l, u) that maximizes

$$\text{Criterion} = \text{Power} - \lambda \left\{ \frac{\mathbb{E}(N_0) + \mathbb{E}(N_1)}{2 * \max N} \right\} \tag{7}$$

subject to the constraint that the type I error is $\leq 10\%$. The maximum total sample size, $\max N$, is 216 here, and $\mathbb{E}(N_0)$ and $\mathbb{E}(N_1)$ are the expected sample sizes (across both active dose and control arms) under the null and alternative hypothesis, respectively. The tuning parameter, λ, controls the contribution to the criterion such that $\lambda > 1$ implies more emphasis on expected sample sizes and $\lambda < 1$ more weight on power.

A value of $\lambda = 1.35$ was chosen to allow a power of 80% to be achieved when $\psi = 1.75$ for the main settings where $n_c = 30$ and $c_1 : c_2 = 4 : 2$. For comparability, the same λ was used in the sensitivity analysis where the size of cohorts

Table 2. Boundaries for settings in the sensitivity analysis of cohort sizes. The settings have the same maximum sample sizes and common criteria to trade-off power and average sample size.

Cohort

c_1	c_2	n_c	l	u	Power	Criterion
4	2	30	0.224	0.839	0.800	0.630
3	3	30	0.268	0.841	0.747	0.573
2	1	30	0.192	0.858	0.794	0.634
2	2	30	0.227	0.858	0.744	0.566
4	2	0	0.317	0.815	0.634	0.438
3	3	0	0.271	0.821	0.691	0.484

Table 3. Percentage of 10,000 simulations where each dose is recommended for (d_1, d_2, d_3) for $c_1 = 4$ and $c_2 = 2$, with controls shared across doses.

		Efficacy scenario				
		0	1	2	3	4
Safety	**0**	(0.4, 2.3, 8.9)	(0.4, **24.0**, **67.2**)	(2.9, **24.2**, **67.5**)	(2.9, **24.6**, **80.6**)	(**5.1**, **29.7**, **80.9**)
	1	(1.1, 5.9, 8.0)	(1.2, **47.4**, **52.9**)	(8.6, **47.6**, **53.6**)	(9.1, **47.3**, **62.1**)	(**14.0**, **56.3**, **61.6**)
Scenario	**2**	(5.5, 7.9, 2.4)	(5.7, **48.4**, 9.4)	(31.7, **49.9**, 9.8)	(31.4, **49.3**, 10.9)	(**45.0**, **57.2**, 10.9)
	3	(8.4, 2.1, 0.1)	(8.2, 7.3, 0.2)	(40.4, 7.8, 0.2)	(40.4, 7.9, 0.2)	(**55.0**, 8.3, 0.2)
	4	(5.0, 0.1, 0.0)	(4.9, 0.4, 0.0)	(17.6, 0.3, 0.0)	(17.6, 0.3, 0.0)	(23.1, 0.4, 0.0)

Desirable doses are highlighted in **bold** and acceptable doses are highlighted in *italics*. Note that these may sum to more than 100% for each scenario as more than one dose can be recommended simultaneously.

and/or the use of past controls was varied. In these other settings, power is lower than 80%, ranging from 63% to 79%. The boundaries for the scenarios are shown in Table 2. We can see from these results that a 2 : 1 allocation within a compound is more powerful than an equal allocation. This is due to the inclusion of additional concurrent control patients (up to 30) from within the platform.

Throughout, the point prior is taken to be $\pi_E^{(j)} = 1/2$ for all doses j. Note that if the boundaries are chosen based on (7), the operating characteristics of the efficacy design are not affected by the choice of point prior. Specifically, since (6) can be re-expressed in terms of the posterior log-odds as

$$\log\left(\frac{\pi_{E|k}^{(j)}}{1-\pi_{E|k}^{(j)}}\right) = \log\left(\frac{\pi_E^{(j)}}{1-\pi_E^{(j)}}\right) + \log\left(\frac{L(\psi^*|D_k^{(j)})}{L(1|D_k^{(j)})}\right)$$

changing the value of $\pi_E^{(j)}$ merely has the effect of translating the posterior log-odds by a constant. Hence the boundaries $(\tilde{l}, \tilde{u})$ under the alternative prior will satisfy the relationship $(\text{logit}(\tilde{l}), \text{logit}(\tilde{u})) = (\text{logit}(l) + \xi, \text{logit}(u) + \xi)$, where ξ is the log-odds ratio between the new and old prior odds of efficacy.

3.4 Results

Detailed results for the setting with a cohort size of $4 + 2$ and inclusion of up to 30 control patients are presented in Table 3. A comparison of results for the varying cohort sizes is illustrated in Figures 2 and 3. Across settings, the overall type I error rate, that is the percentage of simulations in Scenario 0–0 where any dose is recommended, ranges from 11% to 14%. Note that, for a given dose level, the type I error is controlled at 10% as desired and we observe only a small increase when we take all possible doses into account.

Figure 2 shows the percentage of simulations where all desirable doses are recommended (left) and where any desirable doses are recommended (right). For the baseline setting of $c_1 = 4$ and $c_2 = 2$, the mean percentage of simulations recommending all desirable doses is 28.2%, whilst 71.0% of simulations recommended any desirable dose. As expected, safety/efficacy scenarios where only one dose is desirable have higher percentages of simulations recommending all
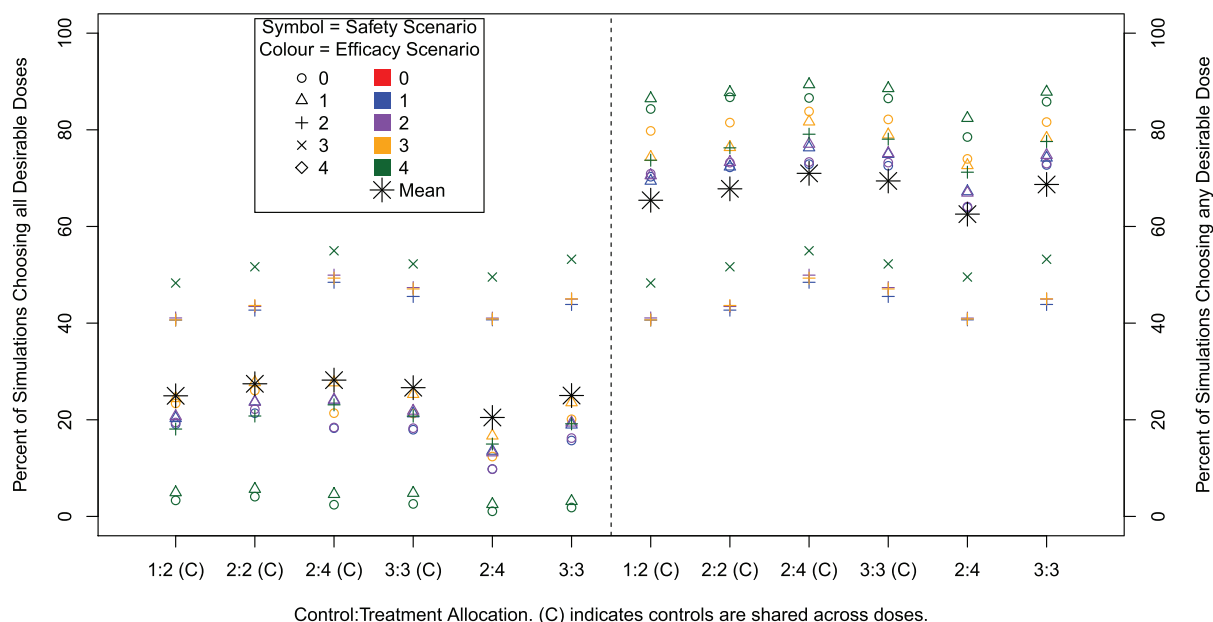
Figure 2. Percentage of simulations that recommend all desirable doses (left) and the percentage of simulations that recommend any desirable dose (right) for different cohort sizes and compositions and with and without sharing control group data. Note that only 13 out of 25 efficacy/safety scenarios contain a desirable dose.
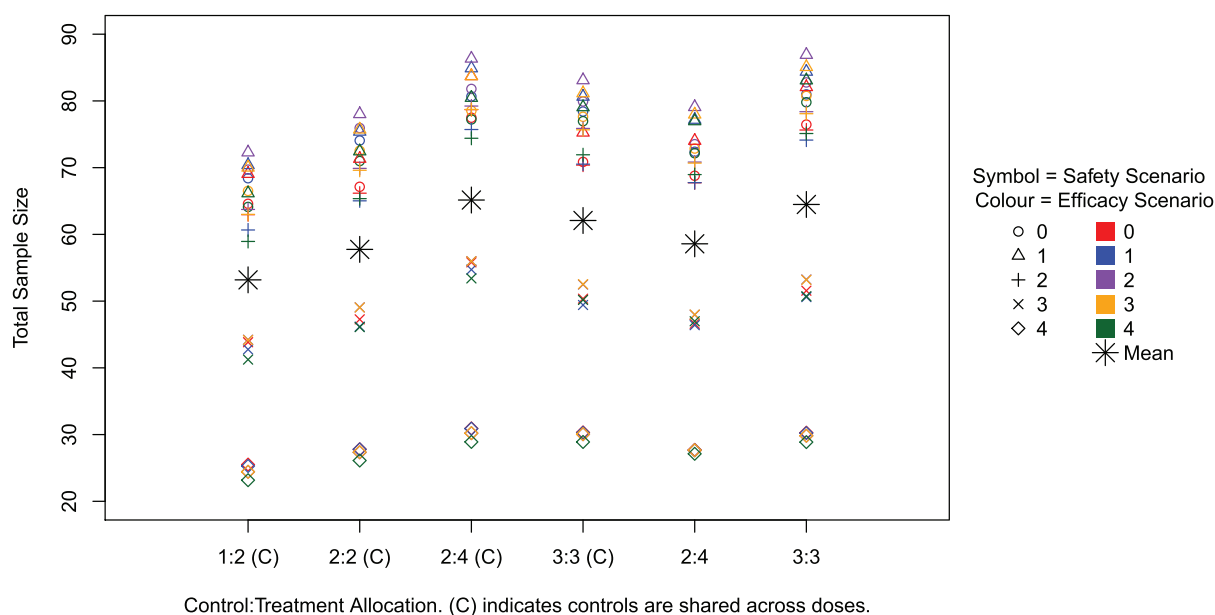

Figure 3. Average total sample size across simulations for all scenarios.

desirable doses, whilst lower percentages of simulations recommending any desirable dose. The highest power for recommending any desirable dose is observed in the safety/efficacy scenarios where all active doses are safe and efficacious, reaching 89% for Scenarios 1 to 4.

The sensitivity analysis across the varying cohort settings shows there is a very small difference in performance. The ordering of performance across safety/efficacy scenarios is identical with only a small numerical difference. However, we can see that not sharing controls decreases the performance. Power for recommending any desirable dose increases for increasing cohort size c, on average increasing 6% for allocation 1 : 2 ($c = 3$ and $c = 6$) and 2% for allocation ratio 1 : 1 ($c = 4$ and $c = 6$). Again unsurprisingly, a lower power is observed when controls are not shared, and within this a higher

power when the allocation ratio between control and active dose within cohort is equal. The highest power is achieved for cohort sizes $c_1 = 4$ and $c_2 = 2$ with controls being shared.

Figure 3 illustrates the average total sample size across the scenarios and settings. On average about 65 patients are required in the setting used in the trial with the total sample size exceeding 150 in only 1% of simulations across scenarios. The scenarios with the smallest sample sizes are those where all doses are unsafe and the trial is therefore stopped early for safety. In such cases, it takes 30 patients across 6 weeks on average to reach the conclusion of stopping early for safety. The scenarios with larger sample sizes are those where all doses are safe and most are acceptable or only just desirable (i.e. not the case where the hazard ratio is 2.00), as in these cases more doses are taken to the efficacy part and more patients are required to detect the smaller difference in hazard ratios.

It can be seen across settings that the larger the total cohort size, the larger the total sample size. This also corresponds to the higher power settings. When controls are shared, altering the control:treatment allocation from 1 : 1 to 1 : 2 decreases the total average sample size, whereas when controls are not shared this increases. This also links to the relationship with the power in the corresponding settings; a higher power is achieved for equal allocations when controls are not shared. The sharing of controls means more patients can be allocated to the novel treatment and hence more can be learnt about it. In all cases, the average total sample size is below 90.

Table 3 gives more detail into which doses are recommended across simulations. For example, in efficacy Scenario 2, where the lowest dose is acceptable and the higher two are desirable in terms of efficacy. In safety Scenario 0, where all doses are safe, the highest dose is chosen most often. In safety Scenario 2 where only the highest dose is unsafe, the middle dose is chosen most often, although less often than the highest desirable dose in Scenario 0. In safety Scenario 4 where all doses are unsafe, the lowest dose is chosen only 17.6% of the time. It is clear that desirable doses are recommended most often, with incorrect and undesirable doses rarely recommended. This gives insight that the procedure is successful in identifying desirable doses of a single agent.

Our additional sensitivity analyses investigating the violation of proportional hazards and providing a comparison against a binary efficacy outcome, presented in the supplemental materials, show that the design is fairly robust to violations of the proportional hazards assumption. Major violations of this assumption yield increased power at the expense of higher type I error. At the same time, these analyses show that the proposed time-to-event approach is superior to using a binary efficacy endpoint as expected.

4 Extension to combination treatments

4.1 Setting

Consider now a randomized controlled dose-escalation dual-agent clinical trial studying the combinations of J doses $d_1 < d_2, \ldots, d_J$ of the first compound (referred to as agent A) and of L doses $s_1 < s_2 < \cdots < s_L$ of the second compound (referred to as agent B). As before, let $d_0 = s_0$ be a zero dose of each compound, respectively, correspondingly to the control treatment, and denote the combination of dose d_j of A and dose s_l of B by (d_j, s_l). Within each agent (assuming the second agent is fixed), it is known prior to the trial that the risk of a DLE monotonically increases with the dose. The objective of the trial is then to study the safety of the combinations and to establish the maximum tolerated combination, the combination corresponding to the target ADLE over the control arm of 20%. Denoting the probability of DLE at combination with doses d_j and s_l by $p(d_j, s_l)$, the probability of each agent given individually by $p(d_j)$ and $p(s_l)$, and the probability of DLE at the control by $p(d_0, s_0) = p_0$, the aim is to find the combination $(d_{j\star}, s_{l\star})$ minimising

$$| \left(p(d_j, s_l) - p_0 \right) - \gamma |$$

The fundamental difference to the single-agent setting introduced in Section 2.1 is that one cannot order all of the combinations of the compounds with respect to the monotonically increasing risk of DLE despite the monotonicity assumption being satisfied within each compound. For example, comparing combination (d_1, s_2) and (d_2, s_1), the dose of one agent is increased and another is decreased, and it is unknown prior to the trial which of these effects prevails in the overall DLE risk associated with the combinations. Consequently, the model-based design for monotherapies in Section 2.2 cannot be used as an alternative dose-finding. Below, the extension of the single-agent model-based design is suggested.

4.2 Dual-agent Bayesian dose-escalation model

For the considered randomized dual-agent combination setting, under the assumption of independence of the compounds, the probability of a DLE associated with combination (d_j, s_l) can be written as

$$p_0(d_j, s_l) = 1 - (1 - p(d_j))(1 - p(s_l)) \tag{8}$$

To allow for the interaction of the compounds in terms of the probability of a DLE, we use the proposed model by Neuenschwander et al.[27]

$$\text{odds}(p(d_j, s_l)) = \text{odds}(p_0(d_j, s_l)) \times \exp\left(\eta \times \tilde{d}_j \times \tilde{s}_l\right)$$

where $\text{odds}(p) = p/(1-p)$ is the odds transformation of the probability p, and η is the interaction coefficient, positive values of which correspond to synergistic DLE risk, zero corresponds to additive effect without interaction, and negative values correspond to the antagonistic risk of DLE, and $\tilde{d}_j, \tilde{s}_l$ are standardized dose levels corresponding to the same DLE probability as the doses d_j, s_l.

Note that $p(\cdot)$ in equation (8) is the probability of a DLE associated with one compound given as monotherapy as in Section 2.2. Therefore, we adopt the two-parameter logistic model given in equation (1) for each agent separately. Specifically, let

$$p(d_j) = \varphi(\tilde{d}_j, \theta_1, \theta_{21}) \tag{9}$$

and

$$p(s_l) = \varphi(\tilde{s}_l, \theta_1, \theta_{22}) \tag{10}$$

where $\theta = (\theta_1, \log(\theta_{21}), \log(\theta_{22}), \eta)$ are the unknown parameters with a normal prior distribution, $\theta \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = (\mu_1, \mu_{21}, \mu_{22}, \mu_\eta)^{\mathrm{T}}$ is the vector of means and

$$\Sigma = \begin{bmatrix} \sigma_1 & \sigma_{1,21} & \sigma_{1,22} & 0 \\ \sigma_{1,21} & \sigma_{21} & 0 & 0 \\ \sigma_{1,22} & 0 & \sigma_{22} & 0 \\ 0 & 0 & 0 & \sigma_\eta \end{bmatrix}$$

As before, we require the standardized dose level corresponding to the control treatment to be equal to $\tilde{d}_0 = \tilde{s}_0 = 0$, so that the intercept parameter of the two-parameter model (8) relates to the probability of DLE on the control only. Therefore, both single-agent models employ the same intercept parameter θ_1 as it corresponds to the probability of a DLE at the same control treatment. Consequently, for small to moderate values of probability of DLE, the intercept parameter in the single-agent model, θ_1 approximately equals the logit inverse-logit transformation of the half of the probability of DLE on the control treatment subject $\text{logit}(p(d_0, s_0)/2) = \theta_1$. This is used to construct the standardized dose levels $\tilde{d}_j, \tilde{s}_l$ using the prior means of the parameters θ and the prior probabilities of a DLE at each combination similarly to the construction in equation (2).

Parameters of the vector θ are the unknown quantities that define the combination–DLE relationship. As in the single-agent design in Section 2.2, the posterior distribution of these is sequentially updated using the data collected during the trial using Bayes' theorem. Specifically, denote the joint prior distributions of vector θ by $f_0(.)$. Assume that n patients have received the combinations $(d1), s(1)) \ldots, (d(n), s(n))$ and binary responses $\mathbb{Y}_n = [y_1, \ldots, y_n]^{\mathrm{T}}$ were observed, respectively. The models update the posterior distribution of θ as

$$\begin{aligned} f_n(\theta) &= \frac{f_{n-1}(\theta)\phi((d(n), s(n)), y_n, \theta)}{\int_{\mathbb{R}^4} f_{n-1}(u)\phi((d(n), s(n), y_n, u)\mathrm{d}u} \\ &= \frac{f_0(\theta) \prod_{i=1}^{n} \phi((d(i), s(i), y_i, \theta)}{\int_{\mathbb{R}^4} f_0(u) \prod_{i=1}^{n} \phi((d(i), s(i), y_i, u)\mathrm{d}u} \end{aligned} \tag{11}$$

where $\phi((d(n), s(n), y_n, \theta) = p(d(n), s(n), \theta)^{y_n}(1 - p(d(n), s(n), \theta))^{1-y_n}$.

This posterior distribution is then used to make the escalation and de-escalation decision during the trials as proposed below.

4.3 Dual-agent dose-escalation design

The above combination-DLE model is then used in the design in Section 2.4 in place of the single-agent model. As in the single-agent setting, escalation can only occur to adjacent doses. As a consequence, no dose skipping is allowed and only

Table 4. Safety and efficacy scenarios for dual-agent combinations *A* and *B*. It is assumed the control arm remains with a probability of dose-limiting event (DLE) 0.10 and a hazard ratio 1.00.

		Safety			Efficacy	
		(Probability of DLE, p_j)			(Hazard ratio)	
		d_1	d_2		d_1	d_2
Scenario 0	s_1	0.10	0.12		1.00	1.00
	s_2	0.13	0.15		1.00	1.00
	s_3	0.15	0.18		1.00	1.00
Scenario 1	s_1	0.10	0.12		1.00	1.25
	s_2	0.25	0.30		1.25	1.50
	s_3	0.50	0.55		1.50	1.75
Scenario 2	s_1	0.15	0.30		1.00	1.50
	s_2	0.25	0.35		1.25	1.75
	s_3	0.30	0.45		1.50	2.00
Scenario 3	s_1	0.40	0.45		1.00	1.50
	s_2	0.45	0.50		1.50	1.75
	s_3	0.50	0.55		1.75	1.75

escalation of one agent in the combination is permitted. In the case of equal probability for two eligible combinations, randomization is used. As the efficacy part of the dose-escalation design proposed for monotherapies considered each dose individually, the efficacy part of the combination study remains the same. Once the combination of the compounds is established to be safe, it is graduated into the efficacy part following the single-agent proposal and the same decision rules for dropping for futility and safety.

5 Evaluation of combination treatment design

5.1 Scenarios

In order to evaluate the dual-agent design, we conduct a simulation study comprising scenarios with two dose levels of agent *A* (d_1 & d_2) and three dose levels of agent *B* (s_1, s_2 & s_3). We consider four dose-DLE scenarios ranging from a situation where all combinations are safe to a case where all are unsafe. Four dose-efficacy scenarios ranging from no efficacious combination to a steep monotonic within agent relationship are considered to yield 16 safety–efficacy scenarios. The dual-agent scenarios are presented in Table 4 with definitions of incorrect, undesirable, acceptable and desirable dose combinations remaining as they were previously defined for single-agent doses in Section 3.2. Here we fix cohort sizes to the previous baseline setting of $c_1 = 4$ (assigned to active dose combinations) and $c_2 = 2$ (assigned to the control) and allow controls to be shared across dose combinations. The maximum number of patients per dose combination is 72 as before.

5.2 Parameter calibration

To define the parameters of the combination model, a calibration procedure similar to the procedure described in Section 3.3.1 was applied. Safety Scenarios 0 to 3 in Table 4 that correspond to different steepness of the combination–toxicity relationship and different locations of the target combination were used. We then choose the hyperparameters for the prior distribution of the parameters of the model, $\theta = (\theta_1, \log(\theta_{21}), \log(\theta_{22}), \eta)$, given in equations (9) and (10).

Given the link between the prior toxicity on the control arm and the intercept parameter θ_1 that is common for both single-agent parameter models, we set $\mu_1 = \text{logit}(\hat{p}_0^{(0)}/2)$, where $\hat{p}_0^{(0)} = 0.10$ is the DLE risk on the control, as before. To reduce the computational complexity, the covariance between all model parameters was assumed to be $\sigma_{1,21} = \sigma_{1,22} = 0$, and the variance of the slope parameters in each single-agent model is the same $\sigma_{21} = \sigma_{22}$. Finally, we set the mean of the distribution of the interaction parameter $\mu_\eta = 0$ to reflect that either synergistic or antagonistic effects are possible. The rest of the parameters were chosen by simulations using combinations of values on the grid, $\mu_{21} = \{-0.4, 0.2, 0.0, 0.2\}$, $\mu_{22} = \{-0.4, 0.2, 0.0, 0.2\}$, $\sigma_1 = \{0.4, 0.6, 0.8, 1.0, 1.2\}$, $\sigma_{21} = \sigma_{22} = \{0.05, 0.15, 0.25, 0.35, 0.45\}$ and $\sigma_\eta = \{0.02, 0.04, 0.10, 0.20, 1\}$.

As in the single-agent setting, the standardized doses $\tilde{d}_j, \tilde{s}_l$ were calibrated in terms of the difference in the probability of DLE between the neighbouring doses. Specifically, $\hat{p}_j^{(0)}(d_j) = \hat{p}_0^{(0)} + v_d \times j$, and $\hat{p}_j^{(0)}(s_l) = \hat{p}_0^{(0)} + v_s \times j$, $l, j = 1, 2, 3$ where v_d, v_s are the differences in the toxicity probabilities between doses of the first and second agents, respectively. The
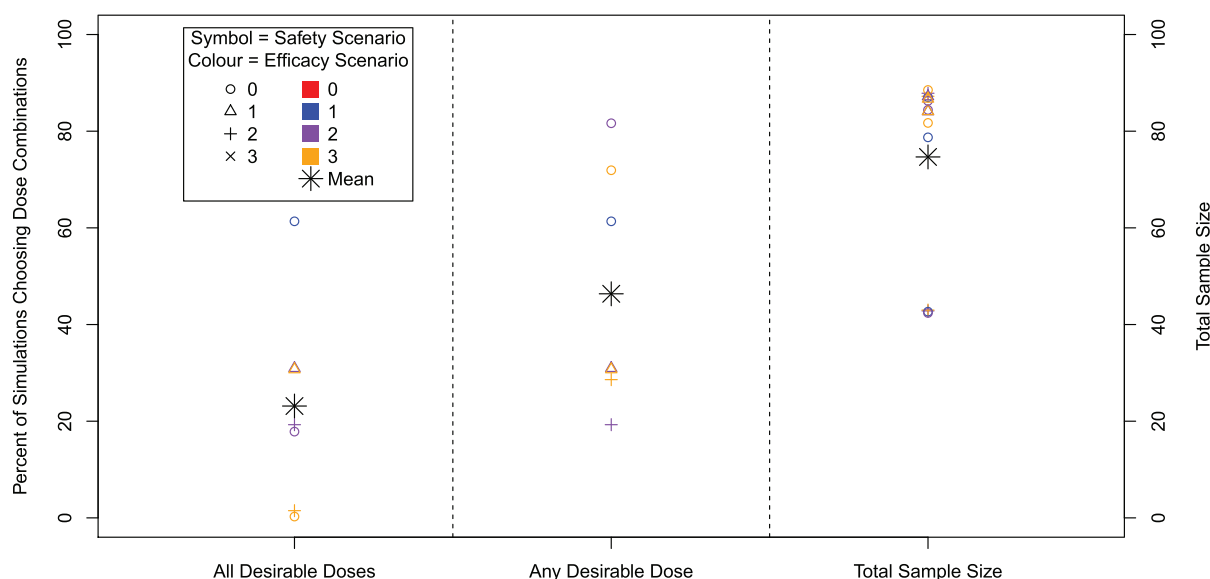
Figure 4. Percentage of 10,000 simulations that recommend all desirable dose combinations (left), the percentage of simulations that recommend any desirable dose combination (centre) and average total sample size. Note that only 7 out of 16 efficacy/safety scenarios contain a desirable dose combination.

Table 5. Percentage of 10,000 simulations where each dose combination is recommended.

		Efficacy scenario							
		0		1		2		3	
Safety scenario	**0**	0.2	0.8	0.2	2.5	0.1	*5.1*	0.2	*5.0*
		0.8	2.9	2.7	*17.3*	2.5	**25.2**	*5.0*	**25.1**
		0.3	7.9	*1.6*	**61.4**	*1.4*	**74.4**	**2.2**	**61.3**
	1	1.1	3.0	1.2	9.5	1.2	*17.3*	1.3	*16.6*
		5.5	5.3	15.2	*23.3*	15.1	**31.0**	28.1	**30.7**
		0.5	0.5	1.0	1.7	1.2	1.9	1.5	1.7
	2	2.8	3.2	3.2	9.3	3.1	*15.8*	3.1	*15.8*
		5.7	4.0	14.7	*14.8*	15.0	**19.3**	26.2	**19.0**
		2.1	0.7	8.6	3.6	8.2	4.1	**11.1**	3.6
	3	6.7	1.1	6.5	2.2	6.6	2.8	6.6	2.9
		1.0	0.1	2.1	0.2	2.1	0.2	2.9	0.2
		0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0

Desirable dose combinations are highlighted in **bold** and acceptable dose combinations are highlighted in *italics*. Note that these may sum to more than 100% for each scenario as more than one dose combination can be recommended simultaneously.

following values of differences were tried $v_d = \{0.025, 0.05, 0.075, 0.10, 0.125\}$, $v_s = \{0.025, 0.05, 0.075, 0.10, 0.125\}$. Finally, we fix $c_{\text{overdose}} = 0.25$ for the overdose control constraint.

Using 500 simulations under each scenario and each combination of hyperparameters, the values $\mu_{21} = 0.0, \mu_{22} = 0.0, \sigma_1 = 0.6, \sigma_{21} = \sigma_{22} = 0.25, \sigma_\eta = 0.10, v_d = v_s = 0.075$ were found to maximize the geometric mean of the proportion of correct combination selections across the scenarios.

5.3 Results

For the 16 scenarios considered in the simulation study, the percentage of simulations recommending any desirable dose combination, the percentage of simulations recommending all correct dose combinations and the mean total sample size are presented in Figure 4, with further detail on individual dose combination recommendations given in Table 5. The overall type I error rate that is the percentage of simulations recommending *any* dose combination in Scenario 0–0, is 12.1%. By construction, the type I error for a given dose is controlled at 10%. In the extension from monotherapies to dual-agent therapies, some similar patterns are maintained in the results although there are some notable differences.

It can be seen in Figure 4 that the spread of powers across scenarios for dose combinations is larger than for monotherapies, for both the selection of all and any desirable dose combinations. The maximum power of 81.7% to select any desirable dose combination is achieved in Scenarios 0 to 2, where all doses are safe and there is a steep monotonic relationship within agents for efficacy. Even though there is an extra desirable dose combination in Scenarios 0 to 3, we observe a slightly reduced power since the most efficacious dose combination has a lower hazard ratio. The power to recommend all desirable doses ranges up to 61% in Scenario 0 to 1 where only one dose is desirable, with the lowest in Scenarios 0 to 3 and 2 and 3 where the desirable dose combinations are across separate agents' dose-escalation (i.e. to select all desirable doses requires a de-escalation in one agent and then escalation in the other agent).

Across scenarios, the mean total sample size is 75, ranging from 42 to 89, a narrower range than for the single agent. However similar to the single agent, the smaller sample sizes correspond to scenarios where all dose combinations are unsafe and therefore the trial stopped early for safety. When this is not the case, there is little variation across scenarios in terms of mean total sample size.

Table 5 shows further details of dose recommendations in the simulations. Especially of note is the emphasis on recommendations of acceptable doses. For example, in Scenarios 2 and 3, where the power to detect desirable dose combinations is low, a large proportion of simulations also recommend an acceptable dose combination. It is also clear that inefficacious and/or unsafe doses are rarely recommended.

6 Discussion

We introduce and evaluate the bespoke statistical design of the AGILE platform which seeks to quickly establish safe doses and potential for efficacy. The novel design utilizes a platform structure that allows the sharing of control data, includes a randomized dose-finding component and yields well-powered decisions about the activity of the treatments while controlling the type I error. We find that the design can identify potential treatments with good accuracy and show that the approach is easily extended to combinations of treatments.

The design uses a recently proposed randomized dose-finding design to ensure that differences between symptoms of COVID can be distinguished from side effects of the investigated treatment while a very simple Bayesian model is used to capture the potential efficacy of the treatments. The latter is in line with the objective of the trial: make reliable decisions about potential quickly, rather than using more complex methods that allow more precise estimation. At the same time, this approach guarantees that the whole platform structure can be simulated quickly to enable the study design to be fixed quickly.

The design has been constructed in a flexible manner using a time-to-event outcome and we based our simulations on time-to-improvement – an endpoint that has been shown recently to be a highly powered and relatively easy to collect.[12] The platform has, however, been constructed to also be able to investigate mild disease in which case a primary endpoint used would be time-to-negative viral titres in nose and/or throat swab. Provided that the event rate in this setting is the same, we expect that the performance reported here will be similar.

In line with Yeung et al.,[28] we have opted for separate models for safety and efficacy to allow the timing of information assessment on safety (7 days) and efficacy (28 days) to be different in order to increase the speed of the dose-escalation. At the same time, Cunanan and Koopmeiners[29] found that in their evaluations using a joint model did not yield improved performance.

In setting up the AGILE platform and more generally when considering Phase I/II trials, several important choices, such as error rates, power and sample size need to be made. Given the exploratory nature of such studies, we believe that, in light of the small sample sizes, it is preferable to allow a somewhat larger type I error in order to achieve adequate power which will prevent missing potentially useful treatments at this early stage of development, something previously highlighted by Lindborg et al.[30] Future development will seek to extend the design to more general prior distributions for the efficacy model and consider extensions that allow the duration of treatment to be explored in addition to dose.

Data availability

Software in the form of R code used to produce the presented results is available on GitHub (https://github.com/dose-finding/covid19-agile).

Declaration of conflicting interests

Funding

ORCID iDs

Thomas Jaki (ID) https://orcid.org/0000-0002-1096-188X
Helen Barnett (ID) https://orcid.org/0000-0001-7466-7033
Pavel Mozgunov (ID) https://orcid.org/0000-0001-6810-0284

Supplemental material

Supplemental material for this article is available online.
A sensitivity analysis for the assumption of proportional hazards is available as part of the online article.

References

1. Cao B, Wang Y, Wen D et al. A trial of Lopinavir–Ritonavir in adults hospitalized with severe COVID-19. *N Engl J Med* 2020; **382**: 1787–1799.
2. Wang Y, Zhang D, Du G et al. Remdesivir in adults with severe COVID-19: A randomised, double-blind, placebo-controlled, multicentre trial. *Lancet* 2020; **395**: 1569–1578.
3. Horby P, Mafham M, Linsell L et al. Effect of hydroxychloroquine in hospitalized patients with COVID-19. *N Engl J Med* 2020; **383**: 2030–2040.
4. Beigel JH, Tomashek KM, Dodd LE et al. Remdesivir for the treatment of COVID-19. *N Engl J Med* 2020; **383**: 1813–1836.
5. Horby P, Campbell M, Staplin N et al. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): A randomised, controlled, open-label, platform trial. *Lancet* 2021; **397**: 1637–1645.
6. Horby P, Lim WS, Emberson JR et al. Dexamethasone in hospitalized patients with COVID-19. *N Engl J Med* 2021; **384**: 693–704.
7. Wages NA and Tait C. Seamless phase I/II adaptive design for oncology trials of molecularly targeted agents. *J Biopharm Stat* 2015; **25**: 903–920.
8. Mozgunov P and Jaki T. An information theoretic phase I–II design for molecularly targeted agents that does not require an assumption of monotonicity. *J R Stat Soc: Ser C (Appl Stat)* 2019; **68**: 347–367.
9. Hirakawa A. An adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase I oncology trials. *Stat Med* 2012; **31**: 516–532.
10. Mozgunov P and Jaki T. A flexible design for advanced phase I/II clinical trials with continuous efficacy endpoints. *Biom J* 2019; **61**: 1477–1492.
11. Yuan Y and Yin G. Bayesian dose finding by jointly modelling toxicity and efficacy as time-to-event outcomes. *J R Stat Soc: Ser C (Appl Stat)* 2009; **58**: 719–736.
12. Dodd LE, Follmann D, Wang J et al. Endpoints for randomized controlled clinical trials for COVID-19 treatments. *Clin Trials* 2020; **17**: 472–482.
13. Pallmann P, Bedding AW, Choodari-Oskooei B et al. Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Med* 2018; **16**: 1–15.
14. Burnett T, Mozgunov P, Pallmann P et al. Adding flexibility to clinical trial designs: An example-based guide to the practical use of adaptive designs. *BMC Med* 2020; **18**: 1–21.
15. Stallard N, Hampson L, Benda N et al. Efficient adaptive designs for clinical trials of interventions for COVID-19. *Stat Biopharm Res* 2020; **12**: 483–497.
16. Meyer EL, Mesenbrink P, Dunger-Baldauf C et al. The evolution of master protocol clinical trial designs: A systematic literature review. *Clin Ther* 2020; **42**(7): 1330–1360.
17. Griffiths G, Fitzgerald R, Jaki T et al. Agile-accord: A randomized, multicentre, seamless, adaptive phase I/II platform study to determine the optimal dose, safety and efficacy of multiple candidate agents for the treatment of COVID-19: A structured summary of a study protocol for a randomised platform trial. *Trials* 2020; **21**: 1–3.
18. Mozgunov P, Jaki T and Paoletti X. Randomized dose-escalation designs for drug combination cancer trials with immunotherapy. *J Biopharm Stat* 2019; **29**: 359–377.
19. Neuenschwander B, Branson M and Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. *Stat Med* 2008; **27**: 2420–2439.
20. Kalbfleisch JD. Nonparametric Bayesian analysis of survival time data. *J R Stat Soc Ser B* 1978; **40**: 214–221.
21. Sinha D, Ibrahim J and Chen MH. A Bayesian justification of Cox's partial likelihood. *Biometrika* 2003; **90**: 629–641.

22. Schmidli H, Gsteiger S, Roychoudhury S et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014; **70**: 1023–1032.

23. Marshall JC, Murthy S, Diaz J et al. A minimal common outcome measure set for COVID-19 clinical research. *Lancet Infect Dis* 2020; **20**(8): e192–e197.

24. Mozgunov P, Jaki T and Paoletti X. A benchmark for dose finding studies with continuous outcomes. *Biostatistics* 2020; **21**: 189–201.

25. Lee SM and Cheung YK. Model calibration in the continual reassessment method. *Clin Trials* 2009; **6**: 227–238.

26. Bailey S, Neuenschwander B, Laird G et al. A Bayesian case study in oncology phase I combination dose-finding using logistic regression with covariates. *J Biopharm Stat* 2009; **19**: 469–484.

27. Neuenschwander B, Matano A, Tang Z et al. A Bayesian industry approach to phase I combination trials in oncology. In: Zhao W and Yang H (eds) *Statistical methods in drug combination studies*. Boca Raton, FL: Chapman & Hall/CRC Press, 2015, chapter 6, pp.95–135.

28. Yeung WY, Whitehead J, Reigner B et al. Bayesian adaptive dose-escalation procedures for binary and continuous responses utilizing a gain function. *Pharm Stat* 2015; **14**: 479–487.

29. Cunanan K and Koopmeiners JS. Evaluating the performance of copula models in phase I–II clinical trials under model misspecification. *BMC Med Res Methodol* 2014; **14**: 1–11.

30. Lindborg SR, Persinger CC, Sashegyi A et al. Statistical refocusing in the design of phase II trials offers promise of increased r&d productivity. *Nat Rev Drug Discov* 2014; **13**: 638–640.