Collaborative Platform Trials to Fight COVID-19: Methodological and Regulatory Considerations for a Better Societal Outcome

Olivier Collignon[1,*], Carl-Fredrik Burman[2], Martin Posch[3] and Anja Schiel[4]

For the development of coronavirus disease 2019 (COVID-19) drugs during the ongoing pandemic, speed is of essence whereas quality of evidence is of paramount importance. Although thousands of COVID-19 trials were rapidly started, many are unlikely to provide robust statistical evidence and meet regulatory standards (e.g., because of lack of randomization or insufficient power). This has led to an inefficient use of time and resources. With more coordination, the sheer number of patients in these trials might have generated convincing data for several investigational treatments. Collaborative platform trials, comparing several drugs to a shared control arm, are an attractive solution. Those trials can utilize a variety of adaptive design features in order to accelerate the finding of life-saving treatments. In this paper, we discuss several possible designs, illustrate them via simulations, and also discuss challenges, such as the heterogeneity of the target population, time-varying standard of care, and the potentially high number of false hypothesis rejections in phase II and phase III trials. We provide corresponding regulatory perspectives on approval and reimbursement, and note that the optimal design of a platform trial will differ with our societal objective and by stakeholder. Hasty approvals may delay the development of better alternatives, whereas searching relentlessly for the single most efficacious treatment may indirectly diminish the number of lives saved as time is lost. We point out the need for incentivizing developers to participate in collaborative evidence-generation initiatives when a positive return on investment is not met.

As the coronavirus disease 2019 (COVID-19) pandemic spread across the world, a vast amount of clinical trials were initiated to meet the urgent need of finding efficacious treatments. However, these efforts were in general not coordinated and this may have led to a suboptimal use of resources, with potentially many trials providing promising evidence of candidate drugs to develop further, but without the robustness needed for regulatory approval.[1–5]

Moreover, these trials often do not use comparable design features: for example, end point definitions vary and target populations tend not to be tailored to a specific research question. Assumptions on sample sizes, if given at all, are driven by optimistically large target treatment effects resulting in underpowered trials unable to detect small but potentially clinically meaningful treatment differences.[6]

Currently, results published for the majority of drugs are therefore at best conflicting and in most cases statistically weak.[7,8] According to ClinicalTrials.gov,[9] 2,993 COVID-19 trials were initiated between October 1, 2019, and August 15, 2020. After removing duplicates and withdrawn studies, 2,940 actual trials remained, out of which 1,643 were interventional and 1,297 were observational. Of the interventional trials, 1,202 were randomized—out of which 400 trials had a sample size smaller than 100 patients. By now, the large number of patients enrolled in these trials would already have allowed for the generation of robust evidence eliciting the potential efficacy of several drugs.

In this article, we therefore wish to initiate a discussion on the need for collaborative and flexible initiatives aiming at providing larger and more robust datasets. These initiatives should allow different sponsors to investigate several candidates simultaneously in a common trial and to deliver data meeting regulatory standards. Indeed, false regulatory decisions and fast approvals based on weak evidence can be expensive, can indirectly delay the development of more promising treatments, and can even be harmful to patients.

To achieve a better societal outcome, we propose the use of platform trials,[10,11] which have also been recently advocated by the European Medicines Agency (EMA).[6,12] We will, in particular, discuss how the use of adaptive elements can potentially improve drug development, for example, by dropping inefficacious treatments as data accrue. We will also shed light on certain obstacles to design a platform trial for COVID-19, such as the heterogeneity of the target population, the potentially high number of false positive treatments progressed to phase III, and a time-varying standard of care (SOC). The design of a platform trial can be more or less conservative or innovative and will differ with the stakeholders' objectives.

We will stress that there is a need for a change in perspective on how governments and regulators should guide and facilitate drug

[1]GlaxoSmithKline, Stevenage, Hertfordshire, UK; [2]Statistical Innovation, Data Science, and Artificial Intelligence, AstraZeneca R&D, Gothenburg, Sweden; [3]Section for Medical Statistics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria; [4]Norwegian Medicines Agency, Oslo, Norway. *Correspondence: Olivier Collignon (olivier.x.collignon@gsk.com)

Received October 26, 2020; accepted January 19, 2021. doi:10.1002/cpt.2183

development in a situation where there is a stronger than usual societal aspect that demands collaboration.

In case of emergency, the burden of development has to be a collaborative effort rather than a competition, a concept that is not the normal *modus operandi* in drug development.

The structure of this paper is as follows. Section "General principles" outlines some of the basic principles and issues around design of COVID-19 trials and drug regulation. Elements to consider when designing a platform trial for COVID-19 is given in the section "Designing an efficient platform trial to fight the COVID-19 pandemic," which also contains simulated examples. The section "Methodological perspectives: pitfalls and potential improvements" shows how the introduction of adaptive elements and stopping rules can improve platform trials, and also elicits several obstacles, such as a potentially increased risk of false positive drugs and an evolving SOC. We will then further discuss regulatory and societal aspects in the section "Societal and regulatory aspects" and give conclusions in the section "Conclusions."

GENERAL PRINCIPLES

The overall objective of any trial is to provide robust evidence on the safety and efficacy of new treatments. A basic principle is to save future lives of patients without endangering the well-being of trial participants.[13]

Unfortunately, many COVID-19 trials are conducted in a nonrandomized fashion, rendering it impossible to make firm causal inference regarding the efficacy of the treatment. Although nonrandomized experiments and real-world data can potentially help understand certain aspects of a disease, it is generally considered necessary that clinical trials are randomized and controlled in order to be able to attribute any potential clinical benefit to the experimental treatment. When possible, blinding of treatment arms greatly increases the integrity of the trial. In this regard, in the absence of a clearly defined population, a predictable disease trajectory and hard end points, single arm trials are therefore not valid designs to address the COVID-19 pandemic.

In the context of the current COVID-19 pandemic, we moreover argue that resources should be focused on a few large (platform) trials. As mentioned above, many trials have been conducted in an uncoordinated way. At a minimum, some measures should be taken to prospectively help compare results between trials. These include providing clear definitions of the trial populations, using a similar control group, and detailing what is used as SOC. More aligned core protocol elements as well as improved transparency of trial reporting would be of help in this regard.[14] In particular, a core set of end points should be defined and systematically collected and reported in every COVID-19 trial. A global regulatory workshop on COVID-19 was recently conducted in order to seek agreement on such acceptable end points.[15] Although mortality might be one of the preferred end points, some drugs can be expected to help accelerate recovery, ease the symptoms among mildly diseased patients, or reduce transmission rates. For many drugs, other end points may have higher power and sensitivity than mortality itself.[16]

Even if cross trial comparisons are made possible, we believe alternative approaches need to be considered given the finite amount of patients and resources.[14] Platform trials, which use a common control group for several drug candidates, can greatly increase the information obtained and allow direct comparisons of active drugs. We will discuss the design of these trials in the next section.

The pandemic raises many additional ethical and societal questions related to the unprecedented pressure on the healthcare systems. Who should be treated and with what? Should societies focus on maximizing benefits, treating equally, promoting and rewarding instrumental value, or giving priority to the worst off first?[17] In addition, under which theoretical framework do we want to develop drugs, solely demanding very rigorous evidence to approve new drugs or flexibly adjusting the treatment of patients to emerging, yet incomplete data? Which goals we pursue might not be easy to determine and can differ based on the underlying healthcare systems, by region and cultural aspects, or by how health care is financed.

One of the challenges in designing a platform trial for COVID-19 would be to find the right way to handle the heterogeneity of the target population (differences in categorization of patients, regional difference in hospitalization or admission to the intensive care units, existing local preferences or emerging changes in treatment algorithms etc.[18]). Depending on the actual aim of the different treatments (symptomatic, curative, and preventive) and type of drugs, only subgroups of the overall COVID-19 affected patient population might be of interest for any particular intervention. Whether to address the heterogeneity by having several parallel platforms, each one specific for a certain part of the population, or rather aim at having one platform, handling the heterogeneity by selective subanalyses, must be carefully considered. For symptomatic treatments, the whole spectrum of mild to severe patients must be considered, yet many of the treatments might have a narrow target population and efficacy signals might be lost if the wrong patients are treated. On the other hand, drugs are often tested on what is considered the most plausible subgroup of a large population without necessarily having excluded the absence of efficacy in a wider population. Depending on the overarching goal of what is to be achieved by the approach to use a platform, develop drugs to reduce mortality, develop symptomatic treatments, develop drugs quickly, or any combination hereof, the approach and design of the platform including the population has to follow.

Depending on these different scopes it is also important to consider a sequential approach, first focusing broader in terms of patients and drugs with a first aim to identify the best candidates with relaxed criteria and then move into more focused development based on the obtained information. Yet, one can also argue to weed out the poor candidates first by focused approaches in the most likely patient population to show a large benefit, thus reducing the risk and costs of developing drugs that might still fail in a confirmatory phase.

DESIGNING AN EFFICIENT PLATFORM TRIAL TO FIGHT THE COVID-19 PANDEMIC

Platform trials are an attractive solution for drug developers to collaborate, streamline efforts, and test a large number of drugs while using a single control arm. Notable evidence of this is that several

platform trials have been initiated during the pandemic, for example DISCOVERY, RECOVERY, ACTIV, and REMAP-CAP.[19]

Designing such a trial can be a challenging task as many possible options are available to the trialist, including the possibility to discard inefficacious drugs via a futility analysis (section "Sequential designs and futility analysis"). We will start with a numerical illustration, presenting a scenario where promising drugs are first screened in a phase II platform trial before confirming their efficacy in a phase III platform trial. In this example, assuming a certain prior distribution around the treatment efficacy of these drugs, we discuss the predicted percentage of efficacious and inefficacious drugs, respectively, progressed to phase III and subsequently submitted to regulators for approval.

A simple example of a phase II platform trial

To illustrate some of the points made in this paper, we will provide an overarching, simplified model showing how candidate drugs could be tested. Say that a total of 50,000 patients will be available for phase II screening of 100 pharmaceuticals that are suggested to potentially have efficacy in newly hospitalized patients with COVID-19. In the example, these drugs will be tested simultaneously, although in practice they could be prioritized according to their expected level of efficacy, taking into account potential similarities due to a shared class or mechanism of action, for example (other clinical development plans could also be envisaged: certain candidate drugs might be investigated directly in phase III trial, whereas for others, development would have to start from phase I). Following the discussion in the section "General principles," we will use the mortality rate as the primary end point (in practice, one could imagine that the end point could differ with the experimental treatment).

If the 50,000 available trial participants are split on the $k = 100$ drugs to be tested, one possibility is to run 100 different 2-armed trials, each comparing one of the active treatments with the SOC in a 1:1 randomization. That is, the sample size per arm will be 250. Assuming a mortality of 10% under the SOC, such a sample size will have limited power to detect a decrease in expected mortality.

A platform trial can be much more efficient, as a common control arm can be shared by different active drugs. In addition, a direct comparison between active drugs may be possible.

Under the global null hypothesis, it is optimal to take the control group to be $k = \sqrt{100} = 10$ times larger than the individual active arms.[20] We choose to allocate 4,000 of the available 50,000 patients in a platform trial to SOC, leaving 460 patients per active arm. In this way, the SOC mortality rate can be estimated with good precision and the sample size per active drug can still be almost doubled as compared with the case when separate trials are run for each drug. Compared with running 100 trials, this platform design reduces the variance for comparisons vs. SOC by a factor 3.3.[†]

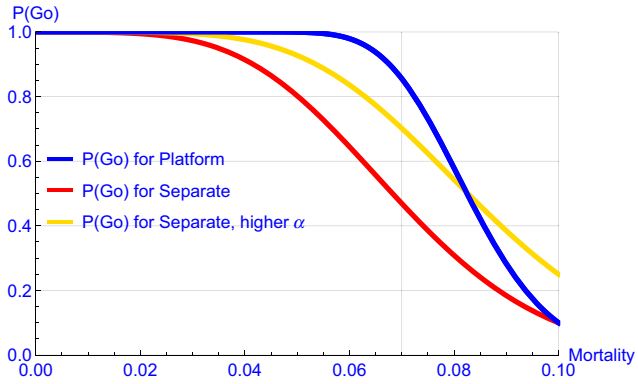

Figure 1 Probability of phase III go (P(Go)) for competing phase II strategies. [Colour figure can be viewed at wileyonlinelibrary.com]

Still, 460 patients is a limited sample for studying mortality and the trialist will be confronted with a trade-off between the risks of progressing inefficacious drugs and the risk of stopping efficacious ones. In this example, we choose to qualify drugs for further testing in phase III if and only if their one-sided P value is lower than $\alpha_2 = 0.10$.[‡] Results in this section are based on a standard normal approximation.

Figure 1 displays the probability of progressing to phase III testing, P(GO), by the expected mortality in the experimental arm. The blue line for the platform trial shows that drugs reducing the mortality from 10% to 5% or less are almost bound to be qualified (P(GO) > 99.9%). If mortality is reduced to 7%, P(GO) for the platform trial is 86% but only 47% if separate trials are conducted (red line). With separate trials, the attrition may be reduced (yellow line) by relaxing the go hurdle ($\alpha_2 = 0.25$) but this could in turn lead to many more inefficacious treatments progressing to phase III.

Phase III

Continuing the main example, the qualified drugs from phase II are then tested in a subsequent phase III platform trial, with a total sample size similar to the phase II trial given reasonable attrition in phase II (see section "Overall platform program performance"). We chose 2,000 patients per active arm and 10,000 patients for the SOC control group. In this phase III trial, a drug is declared positive if it has a statistically significant mortality benefit at one-sided level $\alpha_3 = 0.025$ compared to SOC in phase III. This design gives 90% power for a drug reducing true mortality from 10.0% to 7.8%. The power curve is given in black in **Figure 2**. The blue curve is the same P(GO) curve as in **Figure 1**, whereas the green curve gives the overall probability of success for a drug to be successively positive in both the phase II and III trials.

Overall platform program performance

As we are testing a multitude of drugs, it is of interest not only to consider the type 1 and 2 errors for individual drugs, but also to assess the number of correct and incorrect decisions over the entire

† $\frac{\frac{1}{250}+\frac{1}{250}}{\frac{1}{460}+\frac{1}{4000}} = 3.3$

‡ Other go/no go rules are possible. See e.g., refs. 21 and 22

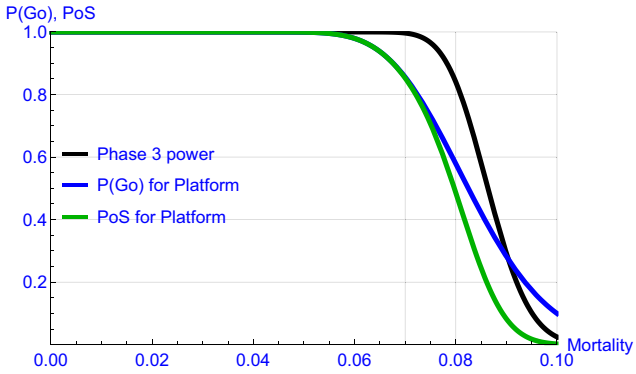

Figure 2 Probability of phase III go (P(Go)), phase III power, and overall probability of success (POS). [Colour figure can be viewed at wileyonlinelibrary.com]

portfolio. To make this concrete, we will assume that the efficacy varies between drugs. More precisely, we assume a prior distribution of the mortality for each of the 100 active drugs (cf.[23]). The possibility of certain drugs increasing mortality is ignored for simplicity. We assume that each drug will have zero efficacy with probability three of four and a certain efficacy with probability one of four. If the drug is efficacious, we assume that a small treatment effect (larger mortality rate under experimental treatment) is more likely than a large one (smaller mortality rate under experimental treatment). Note that this prior should only be viewed as an illustrative example (details about the prior distribution are referred to in the **Supplementary Material**).

The grey line in **Figure 3** shows the assumed prior probability density for the mortality rate, among drugs with efficacy. Using the operating characteristics of phase II (see green curve in **Figure 1**), we can derive the (sub-probability) density of drugs progressing to phase III. This is the pink curve in **Figure 3**. The density of the mortality under the experimental drugs eventually reaching a statistical significance in phase III is shown by the purple line. This is the prior density times the probabilities of going from phase II to III, and then win in phase III.

From **Figure 3**, we can see that almost all drugs with mortality up to 6% succeed in the program. As shown in **Table 1**, the expected number of such drugs, according to the prior, is only 3.2

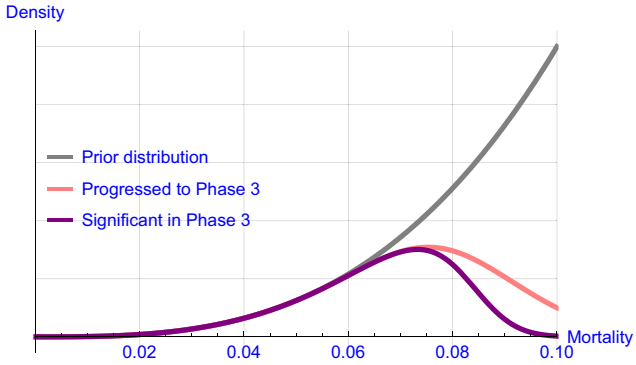

Figure 3 Distribution of mortality rates: before phase II; qualifying from phase II, succeeding in phase III. [Colour figure can be viewed at wileyonlinelibrary.com]

and that attrition among them is ignorable in both phases II and III. Of drugs with mortality between 6 and 8%, there is a 20% risk of stopping in phase II but the drugs proceeding have high power in phase III. Drugs with lower efficacy than that are failing in both phases, leaving only an expected number of 1.7 drugs succeeding out of 15 tested. The figure and table reiterate that many drugs with low efficacy are terminated in phase II, whereas virtually all drugs with high mortality benefit will proceed. Still, according to the prior, of the 19.9 drugs expected to move to phase III, 7.5 have no efficacy at all and the majority has either no or low efficacy. The type 1 error rates α_2 and α_3 in, respectively, phases II and III determine how many inefficacious drugs are propagated. Only $\alpha_2 \cdot \alpha_3 = 0.25\%$ of these drugs get a statistical significance in phase III, the expected number of such outcomes is 0.2. In addition, the familywise error rate for all drugs tested in the program is higher than what would usually be accepted for a single trial.

In this example, we have chosen to focus on a platform trial with as many as 100 drugs in order to push the concept to an extreme (scenario A). We also studied a second scenario (B) including 10 drugs in the platform trial—a more realistic number—and a third scenario (C) in which each of the 10 drugs are studied independently in separate randomized controlled trials with their own control arms. To get the same total sample size per included drug in phase II, we take 400 patients per active arm and 1,000 patients for the control arm in scenario B, and 250 patients on both arms in scenario C. For both scenario B and scenario C, let us assume that 2,500 patients per active and control arm are included in phase III. With other parameters as in the main example, the overall success rate decreases from 10.6% in the main example to 8.9% for scenario B and 6.5% in scenario C, despite the same sample size per drug in phase II and lower expected sample size for the main scenario in phase 3. Relaxing to α_2 to 0.25 (scenario D) increases the overall success rate back to 9.4% but many more drugs are progressed to phase III (graphs and tables are referred to in the **Supplementary Material**).

METHODOLOGICAL PERSPECTIVES: PITFALLS AND POTENTIAL IMPROVEMENTS

The example above used separate platform trials in phase II and phase III, both with fixed designs. Further improvements may be possible by using (group-)sequential designs, potentially stopping early for efficacy or futility; see section "Sequential designs and futility analysis." It is also possible to combine the two phases into a seamless design, as outlined in the section "Seamless design." With many drugs tested simultaneously, there are issues about multiplicity (section "Multiplicity in confirmatory platform trials"). We discuss in the section "Concurrent comparison and changing standard of care" how SOC may change during the course of the trial, and how to handle this in analysis and interpretation of the data. The methodology section concludes in the section "Logistics and infrastructure" with a discussion on logistics and infrastructure.

Sequential designs and futility analysis

Given the large number of candidate drugs currently considered, an important objective of phase II trials is to quickly screen out drugs that show no efficacy. This can be achieved by group

Table 1 Expected number of drugs tested in phase II, proceeding to phase III, and winning in phase III

Mortality	*N* phase II	P(GO)	*N* phase III	P(Win)	*N* wins
< 0.06	3.2	99.6%	3.2	100.0%	3.2
[0.06–0.08)	7.0	80.0%	5.6	97.3%	5.4
[0.08–0.10)	14.8	27.6%	4.1	42.5%	1.7
As SOC	75.0	10.0%	7.5	2.5%	0.2
Total	100		19.9		10.6

Numbers are given by true mortality in the experiemental treatment arms.
P(GO), probability of progressing to phase III testing; P(Win), probability of winning in phase III testing; SOC, standard of care.

sequential designs with repeated interim analyses where the most inefficacious treatments are removed as data accrue. Group sequential designs are most efficient if short-term end points are available, such that at interim analyses their measurement is available for a large fraction of recruited patients. This is the case in COVID-19 trials, because their primary end points can usually be observed within a few weeks. An exception to this is trials with rapid recruitment: in this case, early surrogate end points, if available, can be used to provide a basis for interim decisions and to limit the problem of over-running.[24] For example, if a response biomarker is more informative than the preferred regulatory end point, it can help make faster and more accurate futility decisions.

Although the primary objective is the comparison of each experimental arm to the control group, a platform trial approach also gives the opportunity to make direct comparisons between experimental arms. In addition, it allows one to pool information from several treatment arms for decision making in interim analyses. For example, if the investigated treatments have similar mechanisms of action, then it may be beneficial to also use data from other arms to inform decisions on futility stopping. Note that this will lead to a valid trial, only if the group sequential stopping boundaries have been computed without taking futility stopping into account (e.g., using an α-spending approach). Moreover, the trial will in general become strictly conservative: its actual per comparison error rate may become lower than the nominal significance level. This holds regardless of whether the stopping decision is based on the observed interim treatment efficacy only or also on information from other arms, end points, or external information. On the other hand, if the futility stopping rule for a treatment arm depends only on the interim test statistics corresponding to the primary analysis for that arm, futility stopping can be accounted for in the computation of the group sequential rejection boundaries. Although this can increase the power of the group sequential test, it is in general not recommended, as it entails that futility stopping rules become binding.

Although futility stopping increases the efficiency of trials by stopping nonefficacious treatments at interims, early rejection of null hypotheses at interim analyses can improve the efficiency of trials if treatments are efficacious. This leads to savings in sample size and time gains if treatments can proceed earlier to phase III. Group sequential designs are also an attractive solution for confirmatory phase III trials, as they allow interim looks at the data via an Independent Data Monitoring Committee while controlling the type I error rate using a suitable alpha-spending function. Again, futility stopping and early rejection of the null hypothesis can reduce the required sample size and lead to earlier decisions. This type of design could be particularly suited when implemented in view of a Conditional Marketing Authorization (CMA; see section "Societal and regulatory aspects"). In phase III, however, more than a formal statistical significance for the primary hypothesis may be needed. One also needs to consider whether the available data will provide sufficient evidence in order to change medical practice. Here, considerations with regard to safety analysis, secondary end points, and subgroup analyses may play an important role.

Note that, in general, continuing a group sequential trial, even if a rejection boundary was crossed in an interim analysis, will not lead to an inflation of the type I error rate for the primary hypothesis test. When ethically feasible, continuing to generate additional efficacy information can be very useful. This will, for example, facilitate comparisons of different treatments that all show efficacy vs. old SOC in the trial. Depending on the drugs and their pharmaceutical actions, they may directly compete against each other or could potentially be combined. Further trials will likely be needed to optimize potential combinations.

Seamless design

A further increase in efficiency can be achieved if phases II and III are combined in a single, confirmatory, adaptive, seamless phase II/III trial.[25–28] Seamless designs allow for additional flexibility by, for example, adapting the sample sizes in an interim analysis by reallocating patient numbers from arms that stopped early.[29] However, the selection of promising treatments (or, equivalently, the stopping of apparently ineffective arms) at an interim analysis can introduce a bias in the treatment effect estimates.[27,30] Therefore, appropriate adaptive statistical testing procedures[25,28] that adjust for these biases and guarantee control of the familywise error rate in the strong sense have to be applied. In addition, when integrating phase II and phase III into a single trial, the impact on the overall probability for false positive decisions must be taken into account. In the example, with hypothesis testing hurdles based on mortality in both phase II and phase III, the resulting program-level type 1 error for an inefficacious drug is $\alpha_2 \cdot \alpha_3$. In a standard development program, this could be as low as $0.025^2 = 0.00625$ if conventional $\alpha_i = 0.025$ is used in both trials. Even if the phase II trial efficacy size is not fully predictive (e.g., because it is based on a surrogate end point), phase II will act as a filter and reduce the probability for (inefficacious) treatments to reach phase III. Therefore, to reach the same overall false positive rate in a seamless phase II/III trial as in a classical development program with separate phase II and phase III trials, more stringent significance levels have to be applied in the phase II/III

trial. However, for decision making in a pandemic standard criterion, the required level of evidence may need to be adjusted, balancing the risk of false positive and false negative decisions (see also section "Societal views").

In the example discussed in the section "Designing an efficient platform trial to fight the COVID-19 pandemic," sequential components can bring considerable value, whereas the benefit of a seamless trial may be more limited, as the active arm sample size is considerably lower in phase II than in phase III, so that phase III data will dominate the analysis. Note that the comparison vs. SOC has to be stratified by stage in a platform trial, to reflect the different randomization ratios in the phase II and phase III parts of the trial. A longer trial may be more affected by larger shifts in SOC; see the section "Concurrent comparison and changing standard of care." The gains of the seamless design have to be weighed against any practical advantages in having separate trials. For example, based on learnings from phase II, there could be reasons to adjust the target population, the start of the treatment, the dosing schedule, and perhaps the primary end point. The possibility of adapting within a trial should be contrasted to taking time between phases to optimize phase III.

Important design choices are interim decision boundaries and stage-wise sample sizes in group sequential trials. In a Bayesian framework, these can be optimized (e.g., with the objective to minimize the expected sample size), while controlling the predictive power. As in the example above, given a set of candidate treatments, a prior on the efficacy sizes for these treatments can be specified. Then, for every set of critical boundaries and stage-wise sample sizes, the predictive power of the group sequential trial can be computed for each hypothesis by averaging over the prior and the sampling distribution. Similarly, the expected sample sizes can be evaluated. Now, given an overall number of patients that can be included in the trial, one can search among all decision boundaries and patient allocations that guarantee a given predictive power for the parameters that minimize the expected sample size. This approach can be extended to identify optimal adaptation rules in adaptive clinical trials.[31,32] Adding assumptions on the number of hospitalized patients over time, the expected number of global deaths could in principle be minimized[33] but the optimal solution depends strongly on the assumptions, especially regarding the prior distribution of the treatment efficacy.

Multiplicity in confirmatory platform trials

Whereas in phase II it is generally considered sponsor's risk, it is in general a regulatory mandate to control the probability of declaring at least one false positive test (the familywise error rate (FWER)) at 5% in a confirmatory trial.[34] In a setting where multiple experimental treatments are compared with the same SOC within the same COVID-19 platform trial, it is sensible to wonder about the consequences of multiple testing. On one hand, it can be argued that the probability of erroneously declaring at least one treatment to be efficacious should not be adjusted for multiplicity in a confirmatory platform trial. Indeed, should each of these treatments be investigated separately in a specific trial with its own control arm, 5% of type I errors would be allocated to each of these trials (which would be "spent" over different end points, interim analyses, etc.). It could therefore be perceived as unfair to penalize trials which are optimized to recruit less patients by using a common SOC.

Moreover, the risk of erroneously promoting at least one ineffi-cacious treatment would actually be lower within a platform trial as compared with a series of independent randomized controlled trials (if the hypotheses were to be considered within the same family) thanks to the correlation between the test statistics due to the common control arm. These considerations do, however, not apply anymore when different regimens, doses, or combinations involving the same treatment are tested simultaneously, as it would offer multiple chances of success for the same treatment. Although this would not be an issue in an exploratory trial, in a confirmatory platform trial, it would therefore be recommended to control the FWER for each variant of the same drug progressed to phase III. Admittedly, it can be challenging to agree on how different two drugs are, and engaging in early regulatory interactions should help in this regard.

On the other hand, the probability of erroneously declaring multiple treatments simultaneously superior to SOC might

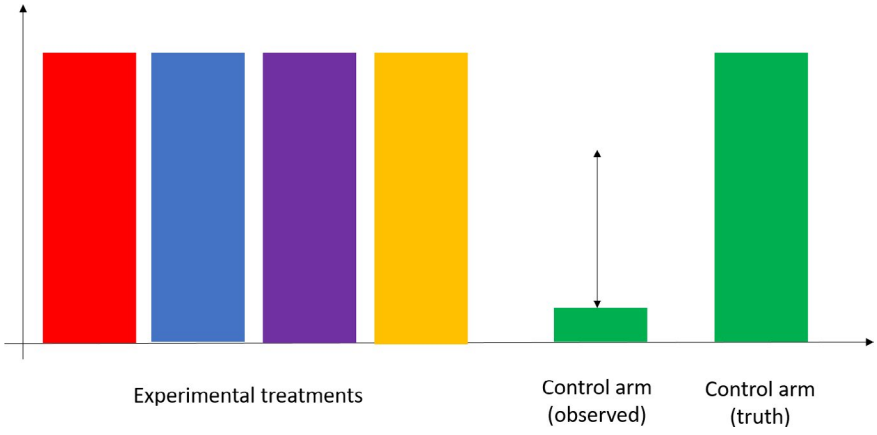

Figure 4 Response rates in the different arms of the platform trial when all treatments are like the standard of care. If in the control arm by chance a low response is observed, for all the experimental treatments an erroneous significant result becomes more likely. [Colour figure can be viewed at wileyonlinelibrary.com]

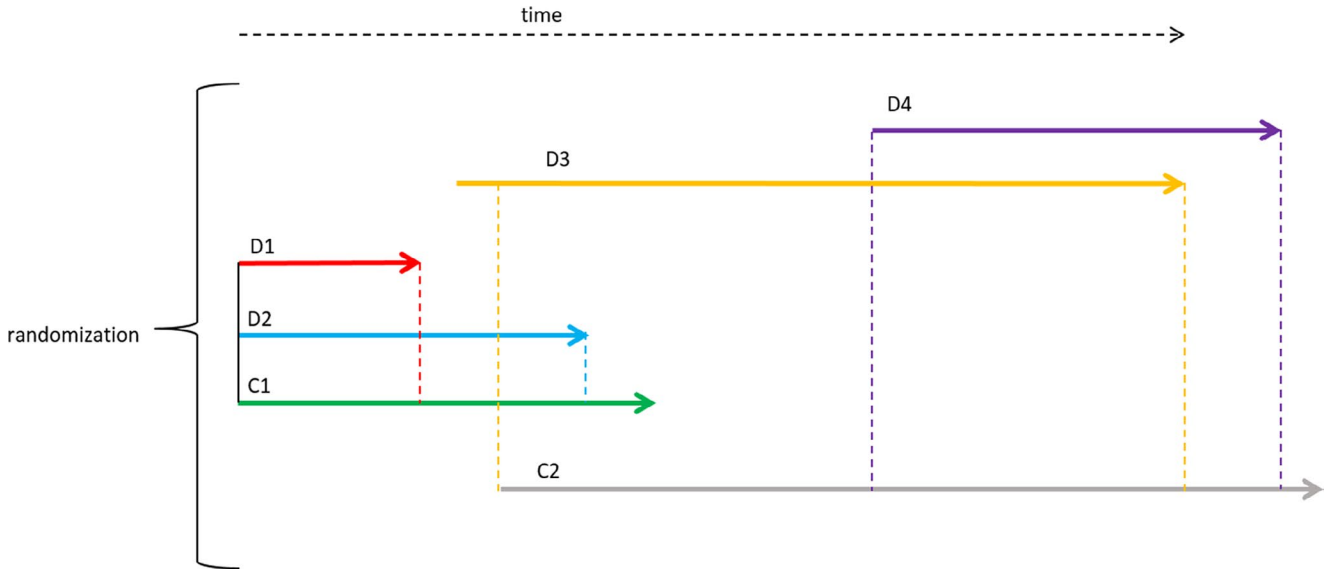

Figure 5 The platform trial starts as a three-arm randomized trial including drugs D_1, D_2, and an active comparator C_1. As data accrue the treatment arms D_3 and D_4 and another active comparator C_2 are added. [Colour figure can be viewed at wileyonlinelibrary.com]

increase with the number of treatments investigated. Indeed, a control arm with an extraordinary low response could by comparison render the investigational treatments more promising (**Figure 4**).

Using a single control arm in a platform trial has many advantages, but it also has its risks. Indeed, it is the regulators' own right to deem a platform trial "uninterpretable" if the response in the control arm is "untrustable" because it seems too low. Equally, a control arm with an extraordinary high response could be signal cancelling. In the context of the current pandemic, it is, however, almost impossible to define what "extraordinary" means, given there is such a short hindsight on the disease trajectory and that new and variable symptoms are described nearly every week. Moreover, there is no acknowledged standard **Figure 4**: response rates in the different arms of the platform trial when all treatments are like the SOC. If in the control arm by chance a low response is observed, for all the experimental treatments, an erroneous significant result becomes more likely of care and the scientific community will be prone to assume that any effect is attributable to the active treatment.

Although controlling this type of error is not governed by a regulatory mandate, this could potentially be costly for sponsors planning a phase II platform trial in which all treatment arms with significant *P*values are progressed to phase III. Indeed, as we have seen in the simulations, a potentially large number of inefficacious drugs could be progressed to phase III. Whereas this could be limited by using smaller significance levels allocated to each treatment arms, this will be at the cost of reducing the percentage of truly efficacious drugs progressed to phase III.

Subpopulations are also a potential source of multiplicity.[35] Indeed, declaring the trial successful if a given (or any) treatment is clinically and statistically significantly superior to control in any subgroup will create a type I error inflation. Although this is acceptable in a nonconfirmatory trial, this is more problematic in the current setting of the pandemic where decisions are to be taken fast, with a smaller hindsight on the biological plausibility of the finding, and with limited supplementary evidence.

A platform trial investigating several drugs with a similar mechanism of action would need to show a certain level of internal consistency. How to explain that out of say 20 drugs with a similar mechanism of action, only one, for example, shows a clinically meaningful and statistically significant efficacy on the primary end point?

For multi-armed clinical trials, especially where a large number of treatments are tested, besides the FWER other measures to quantify the risks of false positives have been proposed. Especially, the false discovery rate, which corresponds to the expected proportion of inefficacious drugs among the drugs that are successful in phase III trials can be relevant in this context.[36]

Concurrent comparison and changing standard of care

In a platform trial, new drugs can also be added dynamically. For example, if a platform phase II trial uses sequential testing, drugs may be qualified for phase III at different points in time. When the trial duration is long or the environment is rapidly evolving, the characteristics of the participants enrolled can change over time. In this situation, the assessment of the efficacy of the new treatment could be restricted to participants enrolled in the control arm after the opening of this new arm, in order to minimize any bias. Alternatively, certain authors have introduced the concept of a time-machine in order to account for these potential differences in characteristics.[37] Others consider the patients enrolled in the control arm prior to the opening of the new arm as a source of external controls, which could be down-weighted to a certain extent in case of data conflict (using, for example, a robustified meta-analytic predictive prior[38]).

In addition, the SOC used as the control arm changes rapidly, especially given the pandemic is lasting. Indeed, the SOC usually refers to the best treatment option available at the current timepoint. In this case, the flexibility of platform trials could prove

useful to integrate a new SOC via a protocol amendment. Once a new SOC, and therefore a new control arm is introduced, the efficacy of each investigated drug will have to be assessed depending on the scenario (**Figure 5**):

1. The treatment was meant to be compared with the initial control arm (C1)
   a The enrollment in the treatment arm (D1) has been completed prior to the addition of the new SOC (C2). In this case (D1), the treatment is compared with the initial control arm (C1), as initially planned.
   b The enrollment in the treatment arms (D2) and (D3) has been completed after the addition of the new SOC (C2). Because the overlap between the treatment (D2) and the new SOC (C2) is short, the treatment should be analyzed as planned and compared with the first comparator (C1). However, via a suitable amendment, the recruitment to the treatment arm could be extended in order to compare it with the new SOC (C2) if needed. On the contrary, the recruitment to the treatment arm (D3) mainly overlaps with the new SOC (C2), the treatment should be compared with the new SOC (C2) by restricting the population to the participants enrolled after the opening of the new SOC. Note that whenever nonconcurrent data are used in the comparisons, the potential bias introduced by time trends needs to be addressed.
2. The treatment (D4) was meant to be compared with the new control arm (C2) (and was therefore initiated after the opening of (C2)). In this case, the comparison is performed as planned.

It is also possible to argue that an add-on drug could, in a stratified analysis, be compared with the series of SOCs. If the drug is shown to have a statistically significant benefit on average, it is reasonable to conclude that it has true efficacy. Further information may be needed to assess the relative merit of this drug as compared with the latest SOC.

If the SOC changes over time and becomes the new drug allocated to the control arm, the sample size initially planned for the trial might become insufficient to demonstrate superiority to this new benchmark for the other treatments in the platform. Planning interim sample size re-assessments at regular intervals could be a suitable solution to this issue.

Logistics and infrastructure

The requirements in terms of logistic and resource to run such platform trials should not be underestimated. The complexity of running platform trials has been discussed in several publications already.[39,40] In particular, initiating or terminating arms is challenging in terms of data handling. Sites need to be informed in a timely fashion of such changes, documentation has to be updated, analyses plans need to be adapted, and appropriate measures have to be in place to ensure robust data handling when multiple developers have competing data management tasks.[41] Last not least, platform trials can be costly from a logistical viewpoint and it remains an unanswered question how and by whom structural costs should be covered.

SOCIETAL AND REGULATORY ASPECTS

Regulatory aspects in Europe

Ethics committees and trial approval bodies are faced with large numbers of trial submissions, some with little scientific rationale. Given the need to protect patients but also avoid exhausting research resources, there is an increasing understanding that proposals with little chance to succeed in providing sufficiently robust evidence should not be approved.

All regulatory agencies have been reacting rapidly to the emerging pandemic. In particular, the EMA has published several living documents in order to support drug developers; the most important being the "Guidance for medicine developers and companies on COVID-19." Furthermore, the EMA has developed several new initiatives. A COVID-19 task force was created to draw on the expertise of the EMA's regulatory network and ensure a fast and coordinated response to the COVID-19 pandemic. A system of rolling review,[42] where emerging evidence is reviewed in cycles until a package is considered complete enough to submit for Marketing Authorization Application, has been offered, considerably shortening the assessment time. Accelerated assessment, a standard regulatory pathway designed to take only 150 days, is also open for products not considered for a rolling review. Furthermore, the EMA now offers rapid scientific advice, a 20-day procedure with no submission deadlines, to allow fast and flexible access to regulatory and scientific advice free of charge.

All of these activities are initiated to invite drug developers to discuss and propose innovative ideas to the regulatory network, to include all relevant expertise, and to ensure that proposals can be assessed and agreed on as fast as possible. The aim is to provide the flexibility needed to move forward fast, while ensuring that no compromises are made on the quality and robustness of the generated evidence. The latter has proven to be of utmost importance as recent discussions around potentially prematurely published or leaked data have shown. Indeed, publishing trial results by "press release" or using unconfirmed sources undermines the public trust and in the worst case makes robust evidence generation impossible. In the unprecedented situation of COVID-19, it must be clear that frequent interactions between developers and regulators or other stakeholders are essential. Any deviations from the normal framework of evidence quality and robustness should ideally be agreed on at an early stage. Failure to do so can indeed lead to an inefficient use of resources and to not approving products based on the uncertainty of their true effectiveness (rather than based on evidence of lack of effectiveness). This must be avoided and requires better and more robust evidence. Similarly, approval in the presence of noncomprehensive data, as is foreseen by the CMA procedure, requires the expectation that uncertainties related to safety and efficacy can be alleviated postapproval. This is based on the concept that additional data will become available to further support the benefit/risk analysis the CMA is based on. This regulatory procedure could be adapted particularly when approval is based on early signs of efficacy at interim analyses. Platform trials can to some extent ensure regulators that early approval will be supported by follow-up data, either from the product in question and/or from a group of products (of the same drug class or mode of action), investigated in the population of interest.

Generation of robust evidence is also a cornerstone for health technology assessments and decisions made by payers. From a societal perspective, investing into collaborative drug development must result in cost-effective drugs based on relevant and robust evidence. Only if drugs are affordable and widely accessible would such a collaborative effort be considered a success.

A way needs to be found to compensate developers of "losing drugs" for their effort and contribution to the pool of evidence that makes a platform trial the most efficient and fast way to develop drugs from a societal perspective. Failing to do so has been the reason why fewer and fewer drug developers engage in the development of antimicrobials, mainly due to the discrepancy between development costs and restrictions to use new antimicrobials and the consequent inability to justify such developments when the expected return does not cover the expenses.[43] We therefore consider it necessary that stakeholders find new ways to provide incentives to steer drug development and contribute to faster and more efficient drug development in general, but urgently in the specific current COVID-19 crisis. We propose that such incentives require more flexibility and a willingness to look into new innovative trial designs.

Societal views

When assessing different drug development strategies, trial designs, and frameworks for drug approval, one first has to define an overall objective, as, for example, the optimization of the outcome. Different stakeholders will without doubt consider the optimal outcome from different viewpoints. In the same way as physicians face ethical challenges when clinical resources are scarce[44] and need to triage,[18] developers need to triage and prioritize which drugs or drug classes to investigate and to do so the ultimate goal of the development program has to be defined *a priori* and agreed on by the different stakeholders. These goals have to guide the choices to be made to identify the most appropriate development strategy, possibly not considering the individual developers' preferences but rather a larger societal perspective. Such societal outcome can be quantified, for example, by a utility function that allows to rank the strategies by the expected overall benefit. Based on such a utility function, one can then aim to optimize different aspects of the drug development strategy and assess the impact of different choices[45]: Should we aim at identifying any treatment substantially better than the SOC or (one of) the best treatments? What level of evidence should be required to license new compounds? How should the speed of development be balanced against the uncertainty of licensing decisions based on limited data? Do we want to reduce the risk of moving too many drugs forward, knowing that many promising candidates will not live up to the early promises? We believe governments and regulators should help drug developers to set priorities and define the utility function with the most societal value.

In a fast-spreading pandemic associated with substantial mortality in higher age groups,[46,47] a main societal objective is the minimization of fatalities. Therefore, a simple utility function could be, for example, the total number of deaths in the pandemic (possibly discounting deaths expected in later time periods). If also other aspects are taken into account as the perspectives of payers and sponsors, the utility function can be adjusted (e.g., for the costs of treatment or the costs of drug development). However, the utility resulting from a specific drug development strategy and licensing policy is unknown in advance. It depends on a range of factors, such as the course of the pandemic, the mortality if no efficacious treatment can be identified, the overall number (and promise) of treatment candidates, the outcome of clinical trials, and the impact of "false positives," (i.e., approving inefficacious drugs). Although the utility cannot directly be predicted, an expected utility of a drug development approach can be computed by specifying prior distributions on the unknown factors (as, e.g., the effect sizes of the candidate treatments as in the section "Overall platform program performance") and by averaging over the sampling distribution of the observations in the clinical trials. Having defined such expected utilities, one can compare different drug development strategies (separate developments and platform trials), trial designs (adaptive, group sequential, and fixed sample), decision rules (e.g., how aggressive futility stopping rules in adaptive designs should be chosen), and licensing rules (i.e., the level of evidence that should be required and the type I error rate applied) for drug licensing.

A major challenge in the application of such a decision theoretic approach is the specification of prior distributions on the many unknowns that determine the expected utility and can have a substantial impact on the outcome. Consider, for example, the dynamics of the pandemics. If only a few future cases are expected (either because of the natural course of the epidemic or because a vaccine becomes available), the optimal level of evidence required for licensing might be lower as only a few future patients will be affected by false positive decisions and delaying a licensing decision would further limit the number of patients that can potentially benefit from the novel treatment. In contrast, for scenarios where the pandemic persists over a longer period and the number of future patients is high, a higher bar for licensing will be optimal, as false positive decisions would affect a large patient population.

CONCLUSION

Basic principles of good clinical trials are highly relevant also in times of an acute pandemic. Clinical trials should be conducted in an ethical way, should preferably be controlled, randomized, and (when possible) blinded. As many treatment candidates are tested, platform trials are an attractive solution. They can be further optimized by integrating adaptive elements to their design. Moreover, stringent futility criteria are needed to prioritize the most promising trial options. Consequently, trial design and decision criteria must be discussed and agreed *a priori* by all stakeholders, as normal regulatory levels of evidence robustness might be questioned. Regulatory requirements may have to change to give optimal benefits to patients and the society, while maintaining a reasonable incentive for drug developers. However, the consequences of deviating from these requirements should be carefully considered and must be fully supported by a firm understanding of the properties of the evidence that underlies such decisions.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENT
The authors would like to thank Antony Sabin (AstraZeneca) for his critical review of the manuscript.

CONFLICT OF INTEREST
The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

1. Coronavirus drugs trials must get bigger and more collaborative. Nature 581(7807), 120 (2020).
2. Janiaud, P. *et al*. The worldwide clinical trial research response to the COVID-19 pandemic-the first 100 days. *F1000Research* **9**, 1193 (2020).
3. London, A.J. & Kimmelman, J. Against pandemic research exceptionalism. *Science* **368**, 476–477 (2020).
4. Noor, N.M. *et al*. Adaptive platform trials using multi-arm, multi-stage protocols: getting fast answers in pandemic settings. *F1000Research* **9**, 1109 (2020).
5. Pundi, K. *et al*. Characteristics and strength of evidence of COVID-19 studies registered on ClinicalTrials.gov. *JAMA Intern. Med.* **180**, 1398–1400 (2020).
6. Eichler, H.-G. *et al*. Clinical trials for COVID-19: can we better use the short window of opportunity? *Clin. Pharmacol. Ther.* **108**, 730–733 (2020).
7. Fairgrieve, D., Feldschreiber, P., Howells, G. & Pilgerstorfer, M. Products in a pandemic: liability for medical products and the fight against COVID-19. *Eur. J. Risk Regulat.* 11(3), 1–53 (2020).
8. Kouzy, R. *et al*. Characteristics of the multiplicity of randomized clinical trials for coronavirus disease 2019 launched during the pandemic. *JAMA Network Open* **3**, e2015100 (2020).
9. ClinicalTrials.gov. <https://clinicaltrials.gov/ct2/results/details?cond=COVID-19> (2020). Accessed August 15, 2020).
10. Collignon, O. *et al*. Current statistical considerations and regulatory perspectives on the planning of confirmatory basket, umbrella, and platform trials. *Clin. Pharmacol. Ther.* **107**, 1059–1067 (2020).
11. Meyer, E.L. *et al*. The evolution of master protocol clinical trial designs: a systematic literature review. *Clin. Ther.* **42**, 1330–1360 (2020).
12. EMA. A call to pool EU resources into large-scale, multi-centre, multi-arm clinical trials against COVID-19 <https://www.EMA.europa.eu/en/documents/other/call-pool-eu-research-resources-large-scale-multi-centremulti-arm-clinical-trials-against-covid-19_en.pdf/> (2020).
13. Burman, C.-F. & Carlberg, A. Future challenges in the design and ethics of clinical trials. In *Clinical Trials Handbook* (ed. Gad, S.) 1173–1200 (Wiley & Sons, Hoboken, NJ, 2009).
14. Wolkewitz, M. & Puljak, L. Methodological challenges of analysing COVID-19 data during the pandemic. *BMC Med. Res. Methodol.* **20**, 81 (2020).
15. EMA. Global regulatory workshop on COVID-19 therapeutics #2: agreement on acceptable endpoints for clinical trials. <https://www.EMA.europa.eu/en/news/global-regulatory-workshop-covid-19-therapeutics-2-agreementacceptable-endpoints-clinical-trials> (2020).
16. Dodd, L.E. *et al*. Endpoints for randomized controlled clinical trials for COVID-19 treatments. *Clin. Trials* **17**(5), 472–482 (2020).
17. Emanuel, E.J. *et al*. Fair allocation of scarce medical resources in the time of COVID-19. *N. Engl. J. Med.* **382**, 2049–2055 (2020).
18. Antommaria, A.H.M. *et al*. Ventilator triage policies during the COVID-19 pandemic at U.S. hospitals associated with members of the association of bioethics program directors. *Ann. Int. Med.* **173**, 188–194 (2020).
19. Mullard, A. COVID umbrella trials multiply. *Nat. Rev. Drug Discov.* **19**, 373 (2020).
20. Dunnett, C. New tables for multiple comparisons with a control. *Biometrics* **20**, 482–491 (1964).
21. Lalonde, R. *et al*. Model-based drug development. *Clin. Pharmacol. Ther.* **82**, 21–32 (2007).
22. Stallard, N. *et al*. Optimal choice of the number of treatments to be included in a clinical trial. *Stat. Med.* **28**, 1321–1338 (2009).
23. Miller, F. & Burman, C.-F. A decision theoretical modeling for phase III investment and drug licensing. *J. Biopharm. Stat.* **28**, 698–721 (2017).
24. Hampson, L.V. & Jennison, C. Group sequential tests for delayed responses (with discussion). *J. R. Stat. Soc. Series B Stat. Methodol.* **75**, 3–54 (2013).
25. Bretz, F. *et al*. Adaptive designs for confirmatory clinical trials. *Stat. Med.* **28**, 1181–1217 (2009).
26. EMA. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design <http://www.EMA.europa.eu/> (2007).
27. Posch, M. *et al*. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Stat. Med.* **24**, 3697–3714 (2005).
28. Stallard, N. *et al*. Efficient adaptive designs for clinical trials of interventions for COVID-19. arXiv preprint arXiv:2005.13309 (2020).
29. Beckman, R. *et al*. Adaptive design for a confirmatory basket trial in multiple tumor types based on a putative predictive biomarker. *Clin. Pharmacol. Ther.* **100**, 617–625 (2016).
30. Bauer, P. *et al*. Selection and bias—two hostile brothers. *Stat. Med.* **29**, 1–13 (2010).
31. Ondra, T. *et al*. Optimized adaptive enrichment designs. *Stat. Methods Med. Res.* **28**, 2096–2111 (2019).
32. Ondra, T. *et al*. Optimizing trial designs for targeted therapies. *PLoS One* **11**, e0163726 (2016).
33. Stallard, N. *et al*. Determination of the optimal sample size for a clinical trial accounting for the population size. *Biom. J.* **59**, 609–625 (2017).
34. EMA. Guideline on Multiplicity Issues in Clinical Trials. <http://www.EMA.europa.eu/> (2017).
35. EMA. Guideline on the investigation of subgroups in confirmatory clinical trials. <http://www.EMA.europa.eu/> (2014).
36. Wason, J.M. & Robertson, D.S. Controlling type I error rates in multi-arm clinical trials: a case for the false discovery rate. *Pharm. Stat.* **20**, 109–116 (2021).
37. Angus, D.C. *et al*. Adaptive platform trials: definition, design, conduct and reporting considerations. *Nat. Rev. Drug Discov.* **18**, 797–808 (2019).
38. Schmidli, H. *et al*. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70**, 1023–1032 (2014).
39. Morrell, L. *et al*. Mind the gap? The platform trial as a working environment. *Trials* **20**, 1–6 (2019).
40. Schiavone, F. *et al*. This is a platform alteration: a trial management perspective on the operational aspects of adaptive and platform and umbrella protocols. *Trials* **20**, 264 (2019).
41. Hague, D. *et al*. Changing platforms without stopping the train: experiences of data management and data management systems when adapting platform protocols by adding and closing comparisons. *Trials* **20**, 294 (2019).
42. EMA. EMA starts rolling review of remdesivir for COVID-19 <https://www.EMA.europa.eu/en/news/ema-starts-rolling-review-remdesivircovid-19> (2020).
43. Shlaes, D.M. The economic conundrum for antibacterial drugs. *Antimicrob. Agents Chemother.* **64**, 1 (2019).
44. McGuire, A.L. *et al*. Ethical challenges arising in the COVID-19 pandemic: an overview from the Association of Bioethics Program Directors (ABPD) task force. *Am J Bioeth.* 20(7), 15–27 (2020).
45. Pearce, M. *et al*. Value of information methods to design a clinical trial in a small population to optimise a health economic utility function. *BMC Med. Res. Methodol.* **18**, 1–9 (2018).
46. Banerjee, A. *et al*. Estimating excess 1-year mortality associated with the COVID19 pandemic according to underlying conditions and age: a population-based cohort study. *Lancet* **395**, 1715–1725 (2020).
47. Posch, M. *et al*. Analysis of Austrian COVID-19 deaths by age and sex. *Wien. Klin. Wochenschr.* **132**, 685–689 (2020).