Clinical trials impacted by the COVID-19 pandemic: Adaptive designs to the rescue?

Cornelia Ursula Kunz[1]*, Silke Jörgens[2]*, Frank Bretz[3,4], Nigel Stallard[5], Kelly Van Lancker[6], Dong Xi[7], Sarah Zohar[8], Christoph Gerlinger[9,10]*, Tim Friede[11,12]*†

[1]Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany

[2]Janssen-Cilag GmbH, Neuss, Germany

[3]Novartis Pharma AG, Basel, Switzerland

[4]Section for Medical Statistics, Medical University of Vienna, Vienna, Austria

[5]Division of Health Sciences, Warwick Medical School, The University of Warwick, Coventry, UK

[6]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[7]Novartis Pharmaceuticals, East Hanover, New Jersey, USA

[8]INSERM, Centre de Recherche des Cordeliers, Sorbonne Universit, Universit de Paris, Paris, France

[9]Statistics and Data Insights, Bayer AG, Berlin, Germany

[10]Department of Gynecology, Obstetrics and Reproductive Medicine, University Medical School of Saarland, Homburg/Saar, Germany

[11]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[12]DZHK (German Center for Cardiovascular Research), partner site Göttingen, Göttingen, Germany

Abstract

Very recently the new pathogen severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was identified and the coronavirus disease 2019 (COVID-19) declared a pandemic by the World Health Organization. The pandemic has a number of consequences for ongoing clinical trials in non-COVID-19 conditions. Motivated by four current clinical trials in a variety of disease areas we illustrate the challenges faced by the pandemic and sketch out possible solutions including adaptive designs. Guidance is

*Authors contributed equally

†Corresponding author: e-mail: tim.friede@med.uni-goettingen.de, Phone: +49-551-39-4991, Fax: +49-551-39-4995

provided on (i) where blinded adaptations can help; (ii) how to achieve type I error rate control, if required; (iii) how to deal with potential treatment effect heterogeneity; (iv) how to utilize early read-outs; and (v) how to utilize Bayesian techniques. In more detail approaches to resizing a trial affected by the pandemic are developed including considerations to stop a trial early, the use of group-sequential designs or sample size adjustment. All methods considered are implemented in a freely available R shiny app. Furthermore, regulatory and operational issues including the role of data monitoring committees are discussed.

Keywords: SARS-CoV-2; Heterogeneity; Interim analysis; Design changes

1 Introduction

In Wuhan, China pneumonia cases of a new pathogen, which was subsequently named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), were identified in December 2019 (Guan et al 2020). Coronavirus disease 2019 (COVID-19) was declared a pandemic by the World Health Organization (WHO). At the time of writing (end of May 2020), more than 5 million cases were confirmed worldwide according to the COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University (`https://coronavirus.jhu.edu/map.html`). To fight the COVID-19 pandemic, a number of clinical trials were initiated or are being planned to investigate novel therapies, diagnostics and vaccines. Some of these make use of novel, efficient trial designs including platform trials and adaptive group-sequential designs. An overview and recommendations are provided by Stallard et al (2020).

While considerable efforts have been made to set up trials in COVID-19, the vast majority of ongoing trials continue to be in other disease areas. In order to effectively protect patient safety in these trials during the COVID-19 pandemic, across the world, clinical trials answering important healthcare questions were stopped, or temporarily paused to possibly re-start later, some with important modifications. Here we consider the impact of the COVID-19 pandemic on trials in non-COVID-19 indications. The challenges to these trials posed by the pandemic can take various forms including the following: (1) The (amount of) missing data may preclude definite conclusions to be drawn with the original sample size. (2) Incomplete follow-up (possibly not at random) may invalidate the planned analyses. (3) Reduced on-site data monitoring may cast doubt on data quality and integrity. (4) Missed treatments due to the interruptions, but also due to acquiring the SARS-CoV-2 virus may not be random and require a different approach than based on the intention-to-treat principle. (5) Circumstances (in e.g. usual care, trial operations, drug manufacturing) before, during and after the pandemic induced interruptions may differ substantially with impact on interpretability of the clinical trial data, through which the original research question is more difficult or even

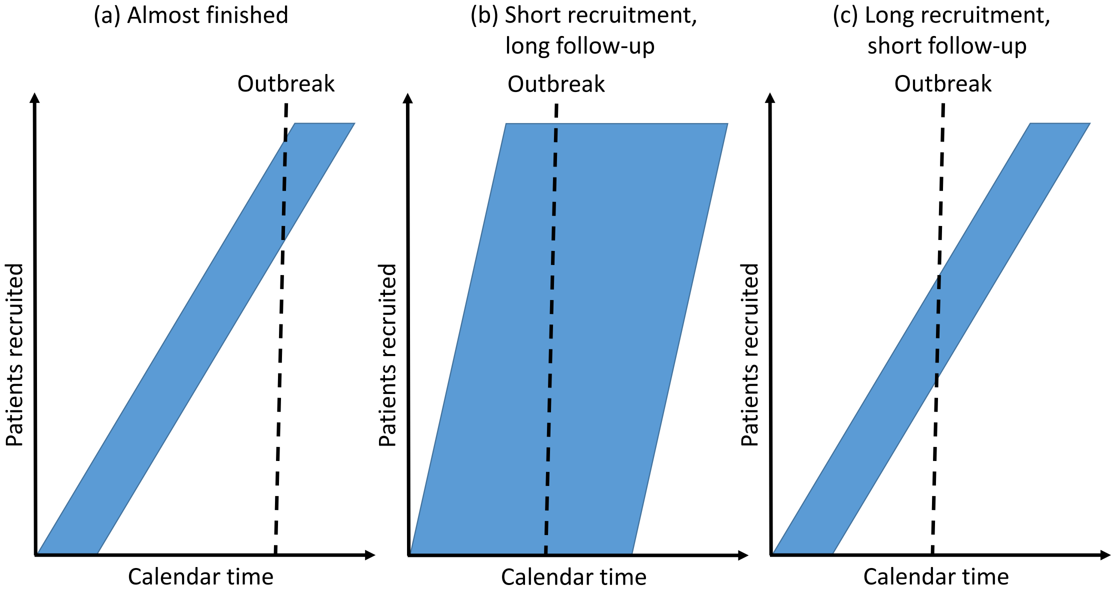


Figure 1: Illustration of how the COVID-19 pandemic impacts clinical trials depending on accrual and follow-up.

impossible to answer. (6) Heterogeneity in patients included in the trial associated with the pandemic may impact results. (7) Potential heterogeneity in included patients may increase for multi-center trials, as the prevalence/incidence of infected patients varies from region to region.

Regulatory authorities have produced guidance on implications of COVID-19 on methodological aspects of ongoing clinical trials (EMA 2020a,b; FDA 2020a,b). The EMA guideline states that the current situation should not automatically encourage unplanned interim or early analyses (EMA 2020b). Despite strong scientific reasons to conduct trials as planned, there may be situations where an unplanned or early analysis may be required to minimize the effect of COVID-19 on the interpretability of the data and results. Potential situations include trials where data collection is nearly finished, an interim analysis is planned in the near future, or when recruitment of new patients is slowing down or interrupted. In particular, the impact of the pandemic depends on the timing of the pandemic compared to the timeline of the trial, the length of follow-up to observe the primary endpoint and the recruitment rate. Figure 1 illustrates three different scenarios, namely (a) an almost finished trial at the time of the outbreak; (b) a trial with relatively short recruitment and long follow-up; and (c) a trial with relatively long recruitment and short follow-up.

For example, when recruitment has been paused and will be restarted after the pandemic, the trial duration will be prolonged. This may be the main impact the outbreak has on a trial such as the one illustrated in panel (c) of Figure 1. A two-stage adaptive design might then be considered for the clinical trial. An interim analysis evaluating the first stage data, which includes participants not affected by COVID-

19, should guide the investigators to decide whether it is worthwhile to restart recruitment after the pandemic and with which sample size. Nevertheless, as any unplanned interim analysis needs to protect the trial integrity (e.g., blinding) and validity (e.g., type I error rate) appropriate statistical methodology for testing and estimation at the end of the trial is an essential aspect. The adaptive design literature offers potential solutions to deal with the concerns in modified trial designs. This has also been recognized by Anker et al (2020) in the context of clinical trials in heart failure, a chronic condition.

The manuscript is organized as follows. In Section 2 four ongoing clinical trials are introduced which are all impacted by the COVID-19 pandemic. These serve as examples and illustrate the many ways trials might be affected by the pandemic. In Section 3 general comments are made on how adaptive designs might be used to overcome the various challenges posed by the pandemic before the issue of resizing trials in terms of trial duration or sample size is considered in more detail in Section 4. In Section 5 other adaptations are briefly touched upon, including blinded and unblinded modifications of the trial design. Regulatory and operational issues including the role of data monitoring committees or data safety monitoring boards are considered in Section 6. In Section 7, we close with a brief discussion.

2 Motivating examples

Clinical trials are affected in many different ways by the COVID-19 pandemic. On the one hand, patients may get infected leading to missed visits, missing data, or even COVID-19 related adverse events. On the other hand, the various lockdown and quarantine measures may disrupt the trial conduct: Patients may be unable to attend their scheduled visits or the study medication cannot be delivered to the patients as planned. While these issues apply to all trials recruiting patients or collecting data during the pandemic, they are affected quite differently depending on the stage the trial was in and also depending on the endpoint of the trial as illustrated by increasing impact in examples 2.1 to 2.4 below.

One important point is still open at the time this paper was written: When and how to restart trials that have had their recruitment interrupted or even study treatment stopped by the onset of the pandemic? The only thing that seems clear is that the conditions under which a trial is restarted will be very trial specific and can be elaborated only provisionally at the end of this paper.

2.1 Long acting reversible contraception: The devil is in the detail

For our first example, consider a study to assess the contraceptive efficacy beyond 5 years up to 8 years of a hormone releasing intrauterine device (IUD) (Jensen et al 2020) (NCT02985541). At the onset of the pandemic all participating women had their IUD in place for more than 6 years but only a few had already completed 8 years of treatment. The primary outcome of the trial is the contraceptive failure rate in years 6 to 8 measured by the Pearl Index (Gerlinger et al 2003). The trial uses a treatment policy estimand, albeit the term *estimand* was not yet common when the contraceptive trial was conceived.

COVID-19 related intercurrent events such as missed or postponed visits to the study center can be ignored for the primary analysis. There will be no interruption of study treatments as the IUD has been in the woman's uterus for 5 years at the beginning of the trial and remains there for up to 8 years in total. Even if the pandemic will last past the scheduled end of the trial, the primary outcome (pregnant yes/no) can still be ascertained even if a woman is not able to attend the final visit in person on time, albeit that according to the statistical analysis plan the continued exposure to the IUD needs to be confirmed by the investigator. Nevertheless, the contraceptive failure rate observed over the whole trial may be impacted not only by a potential loss in confirmed exposure time but also by other COVID-19 related intercurrent events. For instance, a couple who usually commutes long-distance on weekends is not at risk of contraceptive failure during the lockdown if they observe the lockdown living apart, but they are possibly at a higher risk if they observe the lockdown living together. However, given the treatment policy estimand and the very low rate of contraceptive failure with an IUD (Mansour et al 2010) these intercurrent events are not likely to be relevant for the interpretation of the trial's results.

It should be noted that other endpoints of the trial may also be impacted by COVID-19 related intercurrent events. The regular safety assessments planned at the scheduled visits might be at least partially missing if women need to skip the physical visit. While details of some adverse events can be obtained by phone, laboratory values will be definitely missing in such instance. Thus, even for a trial that is very moderately affected by COVID-19, adaptations of the study protocol or the statistical analysis plan might be needed.

2.2 The START:REACTS trial: Change in endpoints due to difficulties in recruitment

Our second example is the Subacromial spacers for Tears Affecting Rotator cuff Tendons: a Randomised, Efficient, Adaptive Clinical Trial in Surgery (START:REACTS) (ISRCTN16912075), an adaptive design multi-center randomized controlled trial conducted in the United Kingdom comparing arthroscopic debridement with the InSpace balloon (Stryker, USA) to arthroscopic debridement alone for people with a symptomatic irreparable rotator cuff tear (Metcalfe et al 2020). Recruitment to the trial started in February 2018, with a planned total sample size of 221 with the potential to stop the study for efficacy or futility at a number of interim analyses. The primary endpoint was shoulder function 12 months after surgery measured using the Constant Shoulder Score (CS) recorded at a hospital out-patient visit, with assessments taken at 3 and 6 months following surgery also used for interim decision-making (Parsons et al 2019).

Due to the coronavirus pandemic, recruitment to the study was delayed by the cancellation of elective surgery in UK hospitals. The study team are working closely with the Data Monitoring Committee in reviewing the planned timing of the interim analyses to reflect this, and the resulting change in the anticipated numbers of patients with 3, 6 and 12 month follow-up data at different time-points in the study.

The pandemic also threatened disruption of the collection of follow-up data for patients for whom surgery had already been completed, as even prior to lockdown, many patients in the study, a large proportion of whom are in vulnerable groups, were unwilling to attend planned appointments for assessment. In order to be able to obtain follow-up data from as many patients as possible, the study team decided to change the primary endpoint to be the 12 month measurement of the Oxford Shoulder Score (OSS), as this does not require face-to-face data collection, but can be completed by post or over phone (or app). As this had originally been included as a secondary endpoint in the study, data were available for all completed patients. The OSS is known to be well correlated to the CS, with the same minimum clinically important difference on a standardized scale, so that the power of the trial is maintained and, as the change was made prior to interim data being observed, there is no loss of trial integrity. For other trials similarly affected, a change in endpoint might be required after the analysis of some data on the original endpoint. In this case an adaptive approach such as that proposed by Bretz (2006) or Klinglmüller (2014) might be used.

2.3 The ATALANTE 1 trial: Premature study discontinuation not to endangering sensitive patients during the COVID-19 pandemic

The ATALANTE 1 clinical trial (NCT02654587) aimed at evaluating and comparing the medicinal product tedopi (OSE2101) to standard treatment (docetaxel or pemetrexed) as second and third line therapy in HLA-A2 positive patients with advanced NSCLC after failure of immune checkpoint inhibitor. This clinical trial was planned in two stages (1) randomized controlled trial (RCT) on a small sample of patients estimating overall survival rate at 12 months (with about 100 patients) and (2) a RCT comparing overall survival (with about 363 patients in total). After the first stage, 99 patients were included (63 in the experimental arm and 36 in the standard arm), the overall survival rate at 12 months was 46% (95% confidence interval: 33% - 59%) in the experimental arm and 36% (95% confidence interval: 21% - 54%) in the standard arm (ose-immuno therapeutics 2020). The second stage of the study was supposed to include patients during 2020. However, this trial was stopped because of the COVID-19 pandemic since, as patients were suffering from lung cancer, the DSMB decided that it was too risky to continue. They stated that it was impossible to expose patients suffering from lung cancer to COVID-19 infection, this could endanger them and may end up biasing the results of the trial (ose-immuno therapeutics 2020). As the results of the first stage were promising, the trial stakeholders decided to discuss with the FDA and the EMA asking whether an additional clinical trial would be required, knowing that there are crucial treatment needs in this indication.

2.4 The CAPE-Covid and the CAPE-Cod (Community-Acquired Pneumonia: Evaluation of Corticosteroids) studies: Embedding a COVID-19 trial within an ongoing trial

Our fourth example is the CAPE-Cod trial (NCT02517489), which aims to assess the efficacy of hydrocortisone at ICU on patients suffering of severe community-acquired pneumonia. At the beginning of the COVID-19 pandemic the trial was active and including patients. As SARS-CoV-2 pneumonia was not an exclusion criterion of CAPE-Cod, centers started to include COVID-19 infected patients into the study. The clinical characteristics between the two indications differed, so trial stakeholders have decided to put temporarily on hold the inclusions in CAPE-Cod study and to use the information of COVID-19 patients by embedding a specific study considering COVID-19 indication only. A group-sequential design using the alpha-spending approach by Kim-DeMets (Kim and DeMets 1987a,b) was chosen for the COVID-19 substudy to account for the considerable uncertainty with regard to the treatment effect in this new group of patients. If the CAPE-Covid study does not achieve the required sample size or stop (for efficacy or futility) before next autumn, there will potentially be inclusions of patients into two studies, as community-acquired pneumonia is a seasonal disease and COVID-19 will still be present. Taking into account patients' heterogeneity will be a major methodological challenge for this trial. More details on the revision of the design are provided in Dequin et al (2020).

3 Issues in adapting a running trial in the COVID-19 pandemic

In this section, guidance is provided on (i) how to achieve type I error rate control; (ii) how to deal with issues surrounding the definiton of estimands; (iii) how to deal with potential treatment-effect heterogeneity; (iv) how to utilize early read-outs; and (v) how to utilize Bayesian techniques.

3.1 Type I error rate control

Even in an open-label trial, an adaption could be based on an analysis that is blinded in the sense that it does not compare treatments. Here we use 'blinded data' to refer more generally to non-comparative data, i.e.

data pooled across treatment arms (FDA 2019). Generally speaking, potential inflation of type I error rate is less of a concern when adaptations are informed by blinded data (EMA 2007; FDA 2019). Therefore, they might be considered first before looking into unblinded adaptations with knowledge of treatment effect estimates. In certain circumstances, however, blinded adaptations may lead to some (often modest) inflation of the type I error rate. Here, we mention non-inferiority and equivalence trials as an example (Friede and Kieser 2003; Friede and Stammer 2010).

It is well known that repeated analyses of accumulating clinical trial data can lead to inflation of the type I error rate (Armitage et al 1969) and to estimation bias. For this reason there is generally a reluctance to modify the design of a clinical trial during its conduct for fear that the scientific integrity will be compromised. The necessity of a severe pause in recruitment in many trials due to the current pandemic, however, raises questions of whether additional analyses can be added to an ongoing trial to enable the data obtained so far to be analysed now, with a decision of whether or not to continue with the trial at a later post-COVID-19 time. Although the current situation of clinical trials being conducted in the setting of a global pandemic is without precedent, the particular question of adding interim analyses to a trial is not a new one.

If interest solely concerns adding an early stop for efficacy or futility in a trial planned with a single final analysis, prior to unblinding one should define an alpha-spending function and change to a group-sequential design (GSD). Although this is sufficient to maintain the validity when it is only of interest to add an early stop, this is no longer the case when one wants to make adaptations. In this case, an appropriate method to control the type I error is required.

Proschan and Hunsberger (1995) introduced the concept of a conditional error function specified prior to the first analysis of accumulating data to be a function that gives the conditional probability of a type I error given the stage 1 data, summarized by a standardized normal test statistic, z_1. In order to control the type I error of the test at level α, the conditional error function, $A(z_1)$, with range $[0, 1]$, must have

$$\int_{-\infty}^{\infty} A(z_1)\phi(z_1)dz_1 = \alpha$$

where ϕ is the standard normal density function. Wassmer (1998) and Müller and Schäfer (2001) showed how this approach can be used to change a single-stage trial to have a sequential design equivalent to that obtained using a group-sequential or combination function test.

The conditional error principle thus enables a trial planned with a single final analysis to be modified at any point prior to that analysis to have a sequential design, with this constructed in such that the type I error rate is not inflated. It should be noted, however, that it is necessary to specify how any data before and after the interim analysis are combined before the first interim analysis is conducted. Modification of the design to include initially unplanned interim analyses will also generally lead to a reduction in the power of the trial, as considered in more detail below.

A similar application of the conditional error principle can be used to modify a trial initially planned with interim analyses. For ongoing clinical trials initially planned with interim analyses, the impact of the COVID-19 pandemic may lead to a desire to modify the timing of the planned analyses. Analyses are often taken at times specified in terms of the information available, which may be proportional to the number of patients for a normally distributed endpoint, or to the number of events for a time-to-event endpoint, or given by the number of events for a binary endpoint. Changes to the timing of the interim analyses do not generally lead to an inflation of the type I error rate provided these are not based on the observed treatment difference, and the spending function method (Lan and DeMets 1983) can be used to modify the critical values used to allow for such changes. In general, we do not expect that the timing of interim analyses related to COVID-19 is based on the estimated treatment difference. If, for whatever reason, these would be related, the type I error rate can be inflated using a group-sequential test (see, for example, Proschan et al (1992)). The combination testing or recursive combination testing approach could then be used to control the type I error rate in this setting (see, for example, Wassmer and Brannath (2016) and Brannath et al (2002)).

3.2 Impact on estimands

Care has to be taken when employing an adaptive design methodology to combine e.g. the information before and after the COVID-19 outbreak. This because the different stages may target a different estimand. An estimand provides a precise description of the treatment effect reflecting the clinical question posed by the trial objective (ICH 2019). It summarizes at a population-level what the outcomes would be in the same patients under different treatment conditions being compared. When each stage of an adaptive design is based on a different estimand, the interpretability of the statistical inference may be hampered. If, for example, the pandemic markedly impacts the trial population after the outbreak because elderly and those with underlying conditions such as asthma, diabetes etc. are at higher risk and therefore excluded from the trial, then this would lead to different stagewise estimands (due to the different population attributes) and limit the overall trial interpretation. The situation is different in adaptive designs with a preplanned selection of a population at an interim analysis (as this does not change the estimand), when following the usual recommendations for an adequately planned trial (which includes the need to pre-specify the envisaged adaptation in the study protocol).

Central to the estimand framework introduced in ICH (2019) are intercurrent events, which occur after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest. Generally, the intercurrent events due to COVID-19 can be categorized into those that are of administrative or operational nature (e.g. treatment discontinuation due to drug supply issues), and those that are directly related to the effect of COVID-19 on the health status of subjects (e.g. treatment discontinuation due to COVID-19 symptoms), see Akacha et al (2020) and Meyer et al (2020). However, the additional intercurrent events are introducing ambiguity to the original research question and teams need to discuss how to account for them (Akacha et al 2017a,b; Lipkovich et al 2020). Care therefore also has to be taken if the pattern of intercurrent events is different before and after an interim analysis, in line with the usual recommendations to assess consistency across trial stages in an adaptive design. Generally speaking, as the definition of an adaptive design implies that we are considering a trial design, it needs to be aligned to the estimands that reflect the trial objectives according to ICH (2019). There are similar problems in non-adaptive trial settings, as the data before and after the COVID-19 outbreak will have to be investigated for consistency (e.g. Friede and Henderson (2009), but adaptive designs raise additional uncertainty through the inclusion of interim analyses.

The considerations in the previous paragraphs are closely related to the trial homogeneity issues discussed in Section 3.3. One particular concern is the possible shift in the study population after the onset of the pandemic. At present we see a notable decline in hospital admissions for non COVID-19 related diseases. It can be assumed that patients with less severe problems tend to postpone a hospital stay for fear of an infection in the hospital or for not putting stress on the already overloaded health system in some countries. Although standard trial procedures like randomization assures the validity of the statistical hypothesis test, it is unclear which population's treatment effect is actually being estimated.

3.3 Treatment-effect heterogeneity

Homogeneity over the stages of a multistage design has always been a topic of discussion. Even without pandemic disruptions, there are various reasons why studies could change over time: Some sites may only contribute to part of study, the study population may change over time, e.g. for reasons of a depleted patient pool, and the disease under study itself may vary over time. While many of these reasons also apply to fixed sample size designs, multistage and especially adaptive trials are under obligation to deliver justifications of why the stages can be considered sufficiently homogeneous in order to test a common hypothesis. The EMA reflection paper on adaptive designs states that "Using an adaptive design implies [...] that methods for the assessment of homogeneity of results from different stages are pre-planned" (EMA 2007). One option they give is the use of heterogeneity tests as known from the area of meta-analyses. However, as Friede and Henderson (2009) point out, this can reduce the power of studies substantially even if there is no heterogeneity as such tests are typically carried out at a higher significance level than the standard ones, thus accepting a higher false positive rate. An alternative they propose is searching for timewise cutpoints in the data. Conclusions about the relationship between timing of change and occurrence of interim analyses can then be drawn from the resulting findings. In the current COVID-19 situation, the challenge statisticians face is similar to the general challenge described above. The nature and the severity of the impact will very much depend on the actual situation of the trial and the disease under study. Consequently, the way to deal with them may differ as described elsewhere in this paper. Here, we will focus on the question of whether the COVID-19-related changes are such that a rescue by introducing an adaptive design seems justifiable from the homogeneity aspect. There is one major difference to the situation described in the preceding paragraph: The presence of one or two cutpoints, depending on whether the trial will continue both during and after COVID-19, can be taken as a given. Also the question of whether the changes are due to a possibly performed interim analysis or due to COVID-19 seems moot; the question we need to answer is whether a combination is justified.

In some cases, it will be obvious that a combination is not warranted. One example for such a case could be studies in respiratory diseases with hospitalizations included in the endpoint, where a COVID-19 related hospitalization may be an intercurrent event. In other cases, it may not be that obvious and there might be reasons to believe that the pooled patient set is suitable to answer the study hypothesis. Due to the reasons listed above, again a formal heterogeneity test will not be the tool of choice. The EMA Draft Points to Consider on COVID-19 (EMA 2020b) does not make mention of the burden of proving homogeneity; rather it states the need of "additional analyses [...] to investigate the impact of the three phases [...] to understand the treatment effect as estimated in the trial". While this does not give sponsor carte blanche to combine as they wish, it clearly leaves room for a number of approaches of justifying combination, both from a numerical and a medical perspective. The estimand framework will be an important factor in the decision on pooling or not pooling the data as it will make arguments visible in a structured way: If estimands differ between study parts, then no meaningful estimator for them will be obtained from pooled data (see also Section 3.2).

What can statistical methodology contribute if it must be conceded that pooling the patients is not justifiable? In some situations, the number of patients before the COVID-19 impact may already be sufficient to provide reasonable power (see Section 4.1). In this case, patients in the COVID-19 timeframe would also need to be analyzed, but it is unclear how they might be included. General guidance for such patients is given, such as repeating the analysis including all patients and discussing changes in the treatment effect estimate. Medical argumentation will then be needed to underpin the assumptions that changes are due to COVID-19. In some cases, causal inference can help estimate outcomes from those patients under the assumption that COVID-19 had not happened. If interested in the treatment effect in a pandemic free world, it might be worth clarifying the question of interest by relating to the estimand framework (ICH 2019) where COVID-19 is seen as an intercurrent event. Alternatively, one could standardize results from all patients to the subgroup of patients pre-COVID-19 (e.g., Shu and Tan (2018) and Hernan and Robins (2020)). Sometimes, also an artificial censoring at the COVID-19 impaction and the use of short-term information (see Section 3.4) to estimate final outcomes will provide a helpful sensitivity analysis.

If it is not feasible to gain sufficient evidence from the pre-COVID-19 patients and a combination does not seem justifiable, then it may be advisable to pause the trial and to re-start it after the COVID-19 time. The during-COVID-19 patients should be included in supporting analyses, but the main evidence will come from the patient pool not directly affected by the pandemic (see also Anker et al (2020)). Short-term endpoints from during-COVID-19 patients may be used in addition to completed patients to inform decisions on the future sample size.

Possible adaptations to mitigate concerns on misjudged effects and to still get a valid and appropriately powered study, like adaptive sample size increase or group sequential testing, are discussed in Sections 4 and 5.

3.4 Use of early read-outs

The use of short-term follow-up for decision making can be helpful as it is generally expected to lead to more efficient decision making. In particular, this is relevant to studies interrupted by COVID-19 as investigators may wish conduct an early or unplanned interim analysis using the pre-pandemic data. Several proposals have been made to use the information on early read-outs to inform the adaptation decision (e.g. Friede et al 2011; Rufibach et al 2016; Jörgens et al 2019). Although the information is different from the primary outcome with all limitations that this might have, a greater proportion of subjects can contribute to the analysis. This is especially useful in trials where only information about the short-term endpoint would be available at the interim analysis (Friede et al 2011).

If primary endpoint data are available, another approach is to retain the pre-specified long-term endpoint as the primary focus of the interim analysis, but to support it with information on short-term data. In particular, such methodology exploits the possible statistical association between the short- and long-term endpoints to provide information about the long-term primary endpoint on patients who did not reach their primary endpoint yet (e.g. Galbraith and Marschner 2003; Sooriyarachchi et al 2006; Stallard 2010; Niewczas et al 2019; Van Lancker et al 2020). To maintain the type I error – even if all unblinded available first-stage data are used in the adaptation decisions, it is recommended to define the first stage p-value by the cohort of patients included before the interim analysis (e.g. Jenkins et al 2011). In comparison with the other existing methods for binary and continuous endpoints, the method of Van Lancker et al (2020) has the advantage of making fully efficient use of the information in the data by, besides multiple short-term endpoints, also taking into account baseline measurements.

Similarly, methods for applying flexible study designs to time-to-event data have also been developed (Brückner et al 2018; Jörgens et al 2019). When data are separated into stages by the occurrence of the primary event, the type I error will be compromised if information other than the current logrank test statistic is used for interim decisions (Bauer and Posch 2004). If short-term endpoints are to be used, Jenkins et al (2011) proposed to base the separation on patients instead of on events. As for other endpoints, this would mean that the primary event for patients who were included before the COVID-19 impact but did experience their primary event only after that impact, would need to be analysed together with those occurring before the impact. Depending on the actual impact, it may be appropriate to either use these patients as a separate cohort – in which case their short-term endpoint should not be used for decision making – or to artificially censor them at the impact timepoint and use their complete data for supplemental analyses only.

However, one should be cautious when employing short-term and longitudinal measurements in adaptive design methodology for trials impacted by COVID-19. As long as the estimand of interest is the (hypothetical) treatment effect not impacted by COVID-19, these analyses will be unbiased if only pre-COVID-19 data is used. For example, in the situation where the number of patients before the COVID-19 impact may already be sufficient to provide reasonable power (see Section 4.1), the use of short-term information to estimate primary endpoint might lead to an even higher power (e.g. Van Lancker et al 2020). In situations where it is not feasible to obtain sufficient information from the pre-COVID-19 patients, it may be advisable to support the interim analysis with historical data (Van Lancker et al 2019). Similarly, if the trial is paused and will re-start after the COVID-19 time, short-term endpoints from during-COVID-19 patients may be used in addition to completed patients to inform decisions on the future sample size (see Section 3.3). As the main evidence will come from the patient pool not directly affected by the pandemic (see also Anker et al (2020)), the during-COVID-19 patients should be included in supplemental analyses only. However, in studies where the estimand of interest is defined with respect to the combination of pre-, during- and post-COVID-19 patients, predicting the the long-term primary endpoint of patients who are still at risk to be impacted by COVID-19 with prediction models based on pre-COVID-19 data only, will lead to biased estimators. Although the prediction models used in Van Lancker et al (2020) could be adapted to account for the (expected) dilution effect, this would require strong assumptions about the effect of the longitudinal measurements on the dilution; which falls out of the scope of this paper. We therefore recommend using the methods presented in Section 4 to modify the trial. Once the pandemic is over, and the trial is resumed, it seems reasonable to resize the trial. In that case, prediction models based on the pre-COVID-19 data can be used to predict the outcomes for the patients recruited after COVID-19 and not impacted by it. Depending on the estimand of interest the sample size can be adjusted or not for the period with dilution.

Note that the different methods described in this section can also take into account missing data (eg, due to COVID-19-related drop-out) if one can assume that the missing mechanism is missing completely at random. In the Appendix of Van Lancker et al (2020) an extension of their method that allows the weaker assumption that missingness is at random is discussed. An alternative for the other methods is to consider more detailed informative missingness models (e.g., via multiple imputation (Sterne et al 2009)).

3.5 A Bayesian perspective

Inter-patient heterogeneity as well as intra-patients heterogeneity are both very common in clinical trials. In COVID-19, however, several types of heterogeneity might add to the ususal level of variability. These include (1) patients infected or not by COVID-19, especially incidence of COVID-19 variability per country in international multi-center trials, (2) patients' outcomes (in cancer studies is the present mortality due to the disease or immunosuppressed systems), (3) patients' follow-up, and (4) patients' compliance due to missing treatments. One way of considering these types of heterogeneity is to use Bayesian approaches during, if possible at all, or at the end of the trial. Using hierarchical Bayesian methods associated with Bayesian evidence synthesis methods will allow different types of heterogeneity to be accounted for (Friede et al 2017; Röver et al 2016; Thall and Wathen 2008). These approaches take into account uncertainty in estimating the between-trial or subgroup heterogeneity but they can also be used in the setting of within-trial heterogeneity. By using potential variation of the scale parameter of the heterogeneity prior would facilitate sensitivity analyses. Friede et al (2017) proposed a Bayesian random-effects meta-analyses with priors covering plausible heterogeneity values. In the setting of within-trial heterogeneity prior calibration of each source of heterogeneity is at most interest, indeed one should not be limited to methods accounting for only one source of heterogeneity as more than one type can be present. Let ϕ be the within-trial standard deviation, it determines the degree of heterogeneity across patients either included before or after COVID-19 pandemic or patients infected or not by COVID-19 (or any other COVID-19 source of heterogeneity) and μ the parameter of interest. Under Bayesian inference, uncertainty for ϕ is automatically accounted for and inference for μ and ϕ can be captured by the joint posterior distribution of the two parameters. The key point is in the choice of the prior distribution of ϕ, in particular when subgroups are small or unbalanced. In the absence of relevant external data or information about within-trial heterogeneity, the 95% prior interval of ϕ should capture small to large heterogeneity. Moreover, the use of a Bayesian approach entails the question of what constitutes sensible prior information in the context of COVID-19 in which there is a continual updating of information that is still not considered reliable. This may be argued on the basis of the endpoint in question, that is, what is the plausible amount of heterogeneity expected, what constitutes relevant external data, and how this information may be utilized. A relatively simple solution would be the use of weakly informative priors. For priors of effect parameters, adaptive priors using power or commensurate prior approaches have proved to be efficient in updating if, when and how to incorporate external information (Hobbs et al 2012; Ollier et al 2019).

4 When trial duration or sample size should be changed

As previously mentioned, there are a number of reasons for changes to the duration or sample size of a clinical trial affected by the COVID-19 pandemic. In particular, such changes may be appropriate if either the trials' feasibility as a whole is affected or if there are serious concerns that the treatment effect will be diluted due to the pandemic, e.g. due to missing data or missed study drug administration.

In the following we develop approaches which can give guidance as to when each change may be appropriate. Specifically, we look at how to calculate the power of the trial under various assumptions and give guidance about the introduction of sample size reassessments and interim analyses. Many of these methods are implemented in a freely available R shiny app, which is briefly introduced in Section 4.5.

4.1 Changing trial duration: Almost done - to stop or not to stop early

If data collection is nearly finished at the time of the COVID-19 impact – i.e. we find ourselves in the leftmost situation of Figure 1 –, a natural question that comes into mind is whether one should analyze the trial early based on the data collected so far accepting some loss in power. The decision on this question can be informed by calculating the actual power based on the original assumptions.

In the following, we focus on superiority trials comparing a treatment versus placebo (or standard treatment) with allocation ratio $1 : r$ for placebo versus treatment. Let α denote the one-sided significance level and $1 - \beta$ the desired power at the planning stage of the trial. Assume that the endpoint of interest (approximately) follows a normal distribution, as for example difference of means or proportions, log odds ratios or log hazard ratios. Let δ denote the assumed difference between the means under the alternative hypothesis and let σ^2 denote the common variance for both arms. The total sample size N needed to achieve a desired power of $1 - \beta$ is then given by

$$N = (z_{1-\alpha} + z_{1-\beta})^2 \frac{\sigma^2}{\delta^2} \frac{(r+1)^2}{r} \tag{1}$$

with $z_{1-\alpha}$ and $z_{1-\beta}$ denoting the $(1-\alpha)$- and $(1-\beta)$-quantile of the cumulative standard normal distribution. We assume that the trial was originally planned to be analyzed based on N observed patients but so far has only data available for a fraction $n = \tau N$ patients. The actual power $(1 - \beta_\tau)$ based on the data observed can easily be shown to be

$$1 - \beta_\tau = \Phi\left(\frac{\delta}{\sigma}\sqrt{\frac{r}{(r+1)^2}}\sqrt{\tau N} - z_{1-\alpha}\right) \tag{2}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. If the true treatment effect is indeed equal to the assumed effect, Equation (2) reduces further to

$$1 - \beta_\tau = \Phi\left(z_{1-\beta}\sqrt{\tau} - z_{1-\alpha}(1 - \sqrt{\tau})\right). \tag{3}$$

Note that the achieved power for the reduced sample size depends on the original assumptions in this specific situation only through the originally planned power and the significance level.

Resulting values for the power depending on the information fraction τ are shown in the first row of Figure 2 (black dotted line). For a desired power of $1 - \beta = 0.80$, if data is available for about 80% ($\tau = 0.80$) of the planned patients, the absolute loss in power for the fixed design is about 10 percentage points while for a planned power of $1 - \beta = 0.90$, the absolute loss in power is about 7 percentage points. Numerical values are included in comprehensive Table 1 in Section 4.3.

There can be no general guidance on what power might still be acceptable. In each individual trial, the decision will be based on balancing the calculated loss in power against the probability of actually obtaining the originally planned amount of data and also on the degree of belief in the originally planned sample size. If doubts remain then a sample size reestimation might be called for.

4.2 Changing sample size: Without looking at comparative data

In this section we focus on blinded sample size reestimation. More general blinded adaptations will be considered in Section 5.1. Blinded sample size reestimation procedures are well established to account for misspecifications of nuisance parameters in the planning phase of a trial (Friede and Kieser 2013). In this situation considered, namely the impact of the COVID-19 pandemic, a number of circumstances might make a resizing of the trial necessary and, as discussed, could be addressed in a blinded sample size review.

A likely scenario is that due to the COVID-19 outbreak the response to the treatment, and possibly even the response to control, changes. As above, assume that at the time of the outbreak $n = \tau N$ patients have been enrolled into the trial and that it is planned to enroll a total of N patients, randomized to control and treatment in a $1 : r$ ratio.

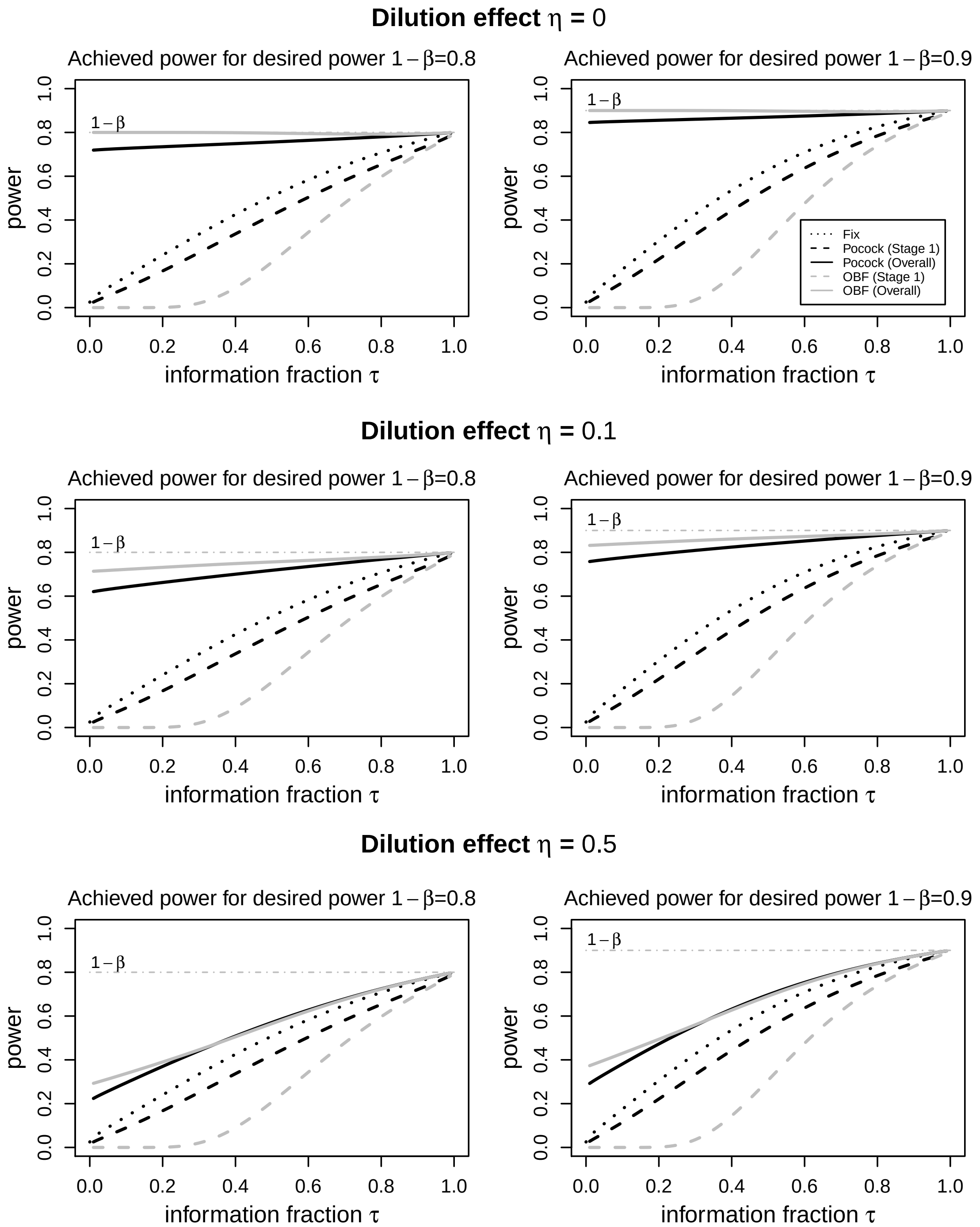


Figure 2: Resulting power depending on the information fraction τ for a dilution effect of $\eta = 0$, $\eta = 0.10$, or $\eta = 0.50$ for the fixed design (black dotted line), the Pocock group sequential design (stage 1: black dashed line, overall: black solid line), and the O'Brien-Fleming group sequential design (stage 1: gray dashed line, overall: gray solid line) for a desired power of either $1 - \beta = 0.80$ or $1 - \beta = 0.90$.

Let μ_{c0} and σ_{c0}^2 denote the mean and the variance for the control group before the outbreak and let μ_{c1} and σ_{c1}^2 denote the mean and variance for the control group after the outbreak. Analogously, the means and variances for the treatment group before and after the outbreak are denoted with μ_{t0}, μ_{t1}, σ_{t0}^2, and σ_{t1}^2. Let $\delta = \mu_{t0} - \mu_{c0}$ denote the treatment effect before the outbreak started. The difference between the means after the outbreak started can then be expressed as a fraction of the difference before the outbreak started, i.e. $\mu_{t1} - \mu_{c1} = (1-\eta)\delta$. While this applies to a relative change in treatment effect, an absolute change in treatment effect can be handled in much the same way as both definitions can be converted into one another. In the following, η will be called the dilution effect.

For the variances, we only consider a relative change of the variance and define $\sigma_{c1}^2 = \psi_c \sigma_{c0}^2$ and $\sigma_{t1}^2 = \psi_t \sigma_{t0}^2$. A common assumption is that the variances for the treatment and the control group are the same. Here, we consider the case of $\sigma_{t0}^2 = \sigma_{c0}^2 = \sigma_0^2 = \sigma^2$ and $\sigma_{t1}^2 = \sigma_{c1}^2 = \sigma_1^2$ with $\psi_t = \psi_c = \psi$ and $\sigma_1^2 = \psi\sigma_0^2$. That is, we assume equal variances for the two arms but not necessarily equal variances before and after the outbreak.

Let t_0 denote the test statistic based on only the patients enrolled before the outbreak and let t_1 denote the test statistic based on only the patients enrolled after the outbreak. Furthermore, let t denote the test statistic based on all enrolled patients. The joint distribution of t_0, t_1, and t is then given by

$$\begin{pmatrix} t_0 \\ t_1 \\ t \end{pmatrix} \sim N\left(\begin{pmatrix} \sqrt{\frac{Nr\tau}{(r+1)^2}} \cdot \frac{\delta}{\sigma} \\ \sqrt{\frac{Nr(1-\tau)}{(r+1)^2}} \cdot \frac{(1-\eta)}{\sqrt{\psi}} \cdot \frac{\delta}{\sigma} \\ \sqrt{\frac{Nr}{(r+1)^2}} \cdot \frac{\tau+(1-\tau)(1-\eta)}{\sqrt{\tau+(1-\tau)\psi}} \cdot \frac{\delta}{\sigma} \end{pmatrix}, \begin{pmatrix} 1 & & \\ 0 & 1 & \\ \sqrt{\frac{\tau}{\tau+(1-\tau)\psi}} & \sqrt{\frac{(1-\tau)\psi}{\tau+(1-\tau)\psi}} & 1 \end{pmatrix}\right). \tag{4}$$

The general solution for the joint distribution can be found in Appendix A.1.

As before, we assume that the original sample size was planned using a one-sided significance level of α to achieve a desired power of $1-\beta$ based on Equation (1). If the true treatment effect is equal to the assumed treatment effect, we can replace N with $((z_{1-\alpha} + z_{1-\beta})^2 \frac{\sigma^2}{\delta^2} \frac{(r+1)^2}{r}$ yielding

$$\begin{pmatrix} t_0 \\ t_1 \\ t \end{pmatrix} \sim N\left(\begin{pmatrix} (z_{1-\alpha} + z_{1-\beta})\sqrt{\tau} \\ (z_{1-\alpha} + z_{1-\beta})\frac{\sqrt{1-\tau}(1-\eta)}{\sqrt{\psi}} \\ (z_{1-\alpha} + z_{1-\beta}) \cdot \frac{\tau+(1-\tau)(1-\eta)}{\sqrt{\tau+(1-\tau)\psi}} \end{pmatrix}, \begin{pmatrix} 1 & & \\ 0 & 1 & \\ \sqrt{\frac{\tau}{\tau+(1-\tau)\psi}} & \sqrt{\frac{(1-\tau)\psi}{\tau+(1-\tau)\psi}} & 1 \end{pmatrix}\right). \tag{5}$$

As shown in Sections 4.1 and later in 4.3, the resulting distribution depends on the values for the significance level α, the desired power $1-\beta$, and the fraction of data available for the outbreak τ. In the case considered here, the only additional variable is the dilution effect η.

Figure 2 shows the resulting values for the power depending on the information fraction τ for a dilution effect of $\eta = 0$, $\eta = 0.10$, and $\eta = 0.20$, assuming a variance inflation/deflation factor of $\psi = 1$. The first row has already been discussed in Section 4.1, $\eta = 0$ corresponding to the original assumption about treatment effect still holding; the middle plots show the power for a dilution effect of $\eta = 0.1$, and the bottom plots for a dilution effect of $\eta = 0.5$, i.e. half the treatment effect getting lost due to the COVID-19 impact.

As above, no general recommendations can be made as every trial is different. However, if the original trial was planned for a power of 90% and at least 85% of the data are available and no considerable dilution effect is expected, then the recommendation could be to stop the trial immediately (if the power loss is offset by a corresponding gain in other regards, e.g. in time to market). In all other scenarios, consequences of any decision would need explored carefully using the approaches developed. As we will see in Section 4.5 below, these are implemented in a R Shiny app to support this process.

If it is considered undesirable to stop the trial immediately, a sample size adjustment in order to restore the desired power of $1 - \beta$ based on the assumed dilution effect can be considered. This does not involve an interim analysis of unblinded data and therefore the type I error rate will be protected. The number of patients which need to be enrolled after the outbreak can be calculated as shown below.

Let n_0 denote the number of patients already enrolled into the trial before the outbreak and let $\tilde{n}_1$ denote the number to be enrolled after the outbreak started. We wish to determine $\tilde{n}_1$ so that the power based on a total of $\tilde{N} = n_0 + \tilde{n}_1$ enrolled patients is $1 - \beta$.

Based on Equation (4), we know that the final test statistic t follows a normal distribution with

$$t \sim N\left(\sqrt{\frac{(n_0+\tilde{n}_1)r}{(r+1)^2}} \cdot \frac{\frac{n_0}{n_0+\tilde{n}_1} + (1-\frac{n_0}{n_0+\tilde{n}_1})\,(1-\eta)}{\sqrt{\frac{n_0}{n_0+\tilde{n}_1} + (1-\frac{n_0}{n_0+\tilde{n}_1})\psi}} \cdot \frac{\delta}{\sigma}, 1\right) \tag{6}$$

Solving

$$\sqrt{\frac{(n_0+\tilde{n}_1)r}{(r+1)^2}} \cdot \frac{\frac{n_0}{n_0+\tilde{n}_1} + (1-\frac{n_0}{n_0+\tilde{n}_1})\,(1-\eta)}{\sqrt{\frac{n_0}{n_0+\tilde{n}_1} + (1-\frac{n_0}{n_0+\tilde{n}_1})\psi}} \cdot \frac{\delta}{\sigma} - z_{1-\alpha} = z_{1-\beta} \tag{7}$$

for $\tilde{n}_1$ yields

$$\tilde{n}_1 = N\tau \frac{\psi - 2 + 2\tau\eta + \sqrt{\psi^2 - 4\tau(1-\eta)(\eta+\psi-1)}}{\psi - 2\tau\eta(1-\eta) - \sqrt{\psi^2 - 4\tau(1-\eta)(\eta+\psi-1)}} \tag{8}$$

The derivations can be found in Appendix A.2.

Note that the dilution effect η cannot be estimated from the data, but needs to be hypothesized. Of course sensitivity analyzes can be conducted based on different assumptions. The R shiny app introduced in Section 4.5 was devised to support such processes.

For time-to-event trials, an additional consideration is that censoring of follow-up might make it necessary to reassess the sample size and length of follow-up. This would be of particular importance in long running trials–as depicted in the middle panel of Figure 1–, particularly prevalent in chronic conditions. In the context of heart failure trials, Anker et al (2020) suggested to censor observations due to regional COVID-19 outbreaks. Such actions would imply a resizing of the trial, potentially in terms of number of patients recruited and length of follow-up, to maintain previously set or in the light of the pandemic revised timelines (Friede et al 2019).

4.3 Possibly changing trial duration based on comparative data: Switching from a fixed to a group-sequential design

If the methods proposed in Section 4.2 suggest that the decision between an immediate stop of the trial and continuation with or without a change in sample size to restore power is not clear, it may be reasonable to include an opportunity for early stopping for efficacy. The trial would then be analyzed as a group sequential design (GSD) using the total sample size from the fixed design. The difference between Sections 4.1 and 4.2 and the situation here is, that we will adjust both the critical value and the sample size to allow for two tests of the null hypothesis. In a first step, we show how to calculate the power for a GSD when the total sample size N is still the one from the fixed design but the critical values are adjusted to account for multiplicity.

Let c_1 and c_2 denote the critical values for a two-stage design and let $\tau = n/N$ denote fraction of data being used for the first stage. Using Φ to denote the cumulative distribution function of the bivariate normal distribution the power is given by (see, for example, Wassmer and Brannath (2016))

$$1-\beta = 1-\Phi\left(\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, \mu = \begin{pmatrix} \sqrt{\tau}\frac{\delta}{\sigma}\sqrt{\frac{r}{(r+1)^2}}\sqrt{N} \\ \frac{\delta}{\sigma}\sqrt{\frac{r}{(r+1)^2}}\sqrt{N} \end{pmatrix}, \sigma^2 = \begin{pmatrix} 1 & \sqrt{\tau} \\ \sqrt{\tau} & 1 \end{pmatrix}\right). \tag{9}$$

Assuming the true effect is equal to the assumed effect at planning stage, we can replace N in Equation (9) by the right-hand side of Equation (1) yielding

$$1-\beta = 1-\Phi\left(\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, \mu = \begin{pmatrix} \sqrt{\tau}\,(z_{1-\alpha}+z_{1-\beta}) \\ (z_{1-\alpha}+z_{1-\beta}) \end{pmatrix}, \sigma^2 = \begin{pmatrix} 1 & \sqrt{\tau} \\ \sqrt{\tau} & 1 \end{pmatrix}\right) \tag{10}$$

As in Section 4.1, the resulting power depends on the originally planned effect (and the allocation ratio) only through the desired power $1-\beta$ at the planning stage. In addition to the dependency on the information fraction τ it depends not only on the significance level α but also on the allocation of the type I error to the stages. The two types of allocation usually used to illustrate the extremes–the O'Brien and Fleming version preserving the major part of the type I error rate to stage 2 and the Pocock GSD boundaries with more type I error allocated to early stages–are included in Figure 2. Each of the designs is represented by two lines, one (the dashed lines) showing the power at the interim analysis and one (the solid lines) showing the power at the final analysis. A selection of values from this figure is included in Table 1 below.

Figure 2 shows the resulting power depending on the information fraction τ for a planned desired power of either $1-\beta = 0.80$ (left-hand panel) or $1-\beta = 0.90$ (right-hand panel) for various values of the dilution effect η. The black dotted line gives the resulting values for the power for the fixed design if analyzed early, the black lines give the resulting power for the Pocock design for the first stage (dashed line) and overall (solid line), and the gray lines give the resulting power for the O'Brien-Fleming design for the first stage (dashed line) and overall (solid line). It should be noted that the power for the fixed design as well as the power for the first stage for the GSDs does not change across different values of η as analyses only use first stage data which was collected before the outbreak.

Table 1 lists the power for some values of τ for a desired power of either 80% oder 90%. The first column gives the value for τ, columns 2 to 6 give the resulting power for the fixed design as well as for both stages the Pocock and the O'Brien-Fleming design for a desired power of $1-\beta = 0.80$, and columns 7 to 11 give the achieved power for a desired power of $1-\beta = 0.90$. The first set of lines assume that there is no dilution effect (see Section 4.2) for patients enrolled into the trial after the COVID-19 outbreak while the second set assumed that the dilution effect is $\eta = 0.10$. For example, if 80% of the planned data has been collected before the COVID-19 outbreak, the resulting power for a fixed design is 0.707 if the planned power is $1-\beta = 0.80$. Using a Pocock GSD, the power for the first stage is 0.653 while the overall power at the end of the second stage is 0.78. For the O'Brien-Fleming GSD, the power for the first stage is 0.597, while the overall power is 0.792.

As the cost of the early efficacy stopping option is paid in terms of power loss, the power for the second stage of such a GSD is lower than the originally planned power even if there is no dilution effect. This loss is generally more pronounced for the Pocock critical boundaries as opposed to the O'Brien and Fleming boundaries (on the other hand, the power for the first stage will generally be higher for Pocock boundaries than for O'Brien and Fleming boundaries).

A sample size reassessment in this case does not require any changes to the design or measures to protect the type I error rate if it is based on methods presented in Section 4.2, i.e. if it uses only the information fraction τ and a guesstimate of the dilution effect η. In order to find the sample size for the second part of the trial for a GSD, a search algorithm based on Equation (4) has to be used.

Table 1: Resulting values for the power based on n patients ($1-\beta_n$) depending on the fraction τ of data already available for an originally planned power of $1-\beta = 80\%$ or $1-\beta = 0.90$ for a dilution effect of $\eta = 0$ or $\eta = 0.10$.

		$\mathbf{1-\beta=0.80}$					$\mathbf{1-\beta=0.90}$			
		Pocock		OBF*			Pocock		OBF*	
		Stage		Stage			Stage		Stage	
τ	fix	1	2	1	2	fix	1	2	1	2
$\boldsymbol{\eta=0,\psi=1}$										
0.50	0.508	0.422	0.756	0.207	0.797	0.630	0.545	0.870	0.307	0.898
0.60	0.583	0.504	0.764	0.344	0.795	0.709	0.637	0.875	0.476	0.896
0.70	0.650	0.581	0.772	0.478	0.793	0.774	0.717	0.880	0.622	0.895
0.80	0.707	0.653	0.780	0.597	0.792	0.826	0.785	0.886	0.739	0.895
0.85	0.733	0.688	0.785	0.650	0.793	0.848	0.815	0.889	0.786	0.895
0.90	0.757	0.721	0.789	0.699	0.794	0.868	0.842	0.892	0.826	0.896
0.95	0.780	0.754	0.794	0.745	0.796	0.885	0.868	0.896	0.862	0.897
0.99	0.796	0.785	0.799	0.783	0.799	0.897	0.890	0.899	0.889	0.899
$\boldsymbol{\eta=0.1,\psi=1}$										
0.50	0.508	0.422	0.718	0.207	0.756	0.630	0.545	0.838	0.307	0.867
0.60	0.583	0.504	0.735	0.344	0.763	0.709	0.637	0.852	0.476	0.872
0.70	0.650	0.581	0.752	0.478	0.770	0.774	0.717	0.864	0.622	0.878
0.80	0.707	0.653	0.768	0.597	0.778	0.826	0.785	0.877	0.739	0.884
0.85	0.733	0.688	0.776	0.650	0.783	0.848	0.815	0.883	0.786	0.887
0.90	0.757	0.721	0.784	0.699	0.788	0.868	0.842	0.888	0.826	0.891
0.95	0.780	0.754	0.792	0.745	0.793	0.885	0.868	0.894	0.862	0.895
0.99	0.796	0.785	0.798	0.783	0.798	0.897	0.890	0.899	0.889	0.899

* OBF: O'Brien-Fleming

While theoretically one could attempt to estimate the dilution effect based on available data, it should be noted that the estimator has a huge variability. For example, assume that the original trial was planned to detect a treatment effect of 0.35 with a desired power of 90% and a one-sided significance level of 0.025. Furthermore, assume that the true treatment effect is indeed as planned and that 70% of the data has been collected before the outbreak while the remaining 30% of the data was collected after the outbreak. For a true dilution effect of 0.1, the lower and upper 5% percentile of the distribution are -2.7453 and 1.2486 with expected value being -0.1011. That is, on average we would conclude that there is no dilution effect but that the treatment effect after the outbreak is even larger than before the outbreak! The distribution for the dilution effect can be derived based on the article by Hinckley (1969). It should be noted that if we would switch to a GSD with Pocock boundaries without adjusting the sample size, the power to reject the null hypothesis at the first stage is 72% while the overall power is 86%.

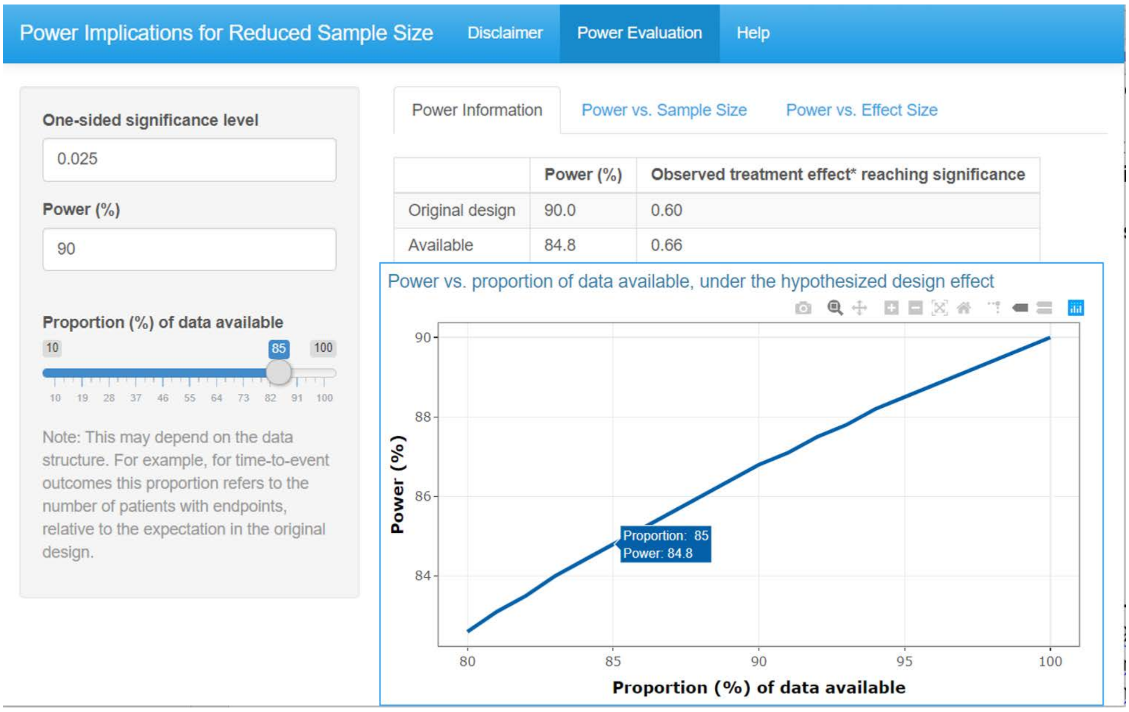
Instead of trying to estimate the dilution effect, we recommend calculating the sample size for a GSD using different values for the dilution effect and then deciding on the maximum sample size still feasible in the specific situation. This approach is similar to calculating the effect that can be detected given a specific sample size instead of calculating the sample size to detect a specific treatment effect. It should be kept in mind that if the dilution effect is too large, even if the trial is significant in the end, the product might not be marketable. If the dilution effect is solely a result of external circumstances, as for example missed treatment cycles, one could also consider stopping recruitment until patients are able to fully comply with the treatment protocol and continue the trial based on the original sample size. Depending on the expected dilution effect, this approach might still lead to a shorter trial length than trying to adjust the sample size. For the example above, the original total sample size is 344 and assuming an accrual rate of 20 patients per months it would have taken approximately 12 month to recruit 70% of the patients. It would require a further 5 months to finish recruitment. If we assume that the dilution effect is 0.25, a GSD with Pocock boundaries requires a further 229 patients which could be recruited in approximately 12 months. If we were able to resolve the issues leading to the dilution effect within 6 months and restart recruitment, the overall trial length would be similar to the GSD with recruitment continued. Overall, we see that there is no simple solution applicable to all trials.

4.4 Possibly changing trial duration and sample size: Switching to a group sequential adaptive design

In some cases, it may be desirable to combine the early stopping option with sample size reassessment based on comparative analyses. This may be because the uncertainty seems too big to conduct a sample size reassessment based on an assumption of the dilution effect; or in situations like the right panel in Figure 1, where the study will continue following the pandemic. Two different options already mentioned in Section 3.1 would be candidates for such a design change:

1. A group sequential adaptive design using a combination test with prespecified weights and group sequential boundaries as already used in Section 4.3;
2. A recursive combination test allowing a complete redesign after the interim analysis, including sample size and number and timing of future interim analyses.

While the latter option provides more flexibility, this characteristic may be the very reason not to use it, especially if the trial is located in a later development stage. If the aims of the interim analysis are early efficacy stopping and sample size reassessment only, then the former choice would provide both options and still stay reasonably close to a group sequential designs if the actual changes in sample size are small. The prerequisite, however, is fixing the number of interim analyses and the weight to be used for each stage of the design before the blind is broken. This may of course pose difficulties in the current situation where recruitment during the pandemic and return to projected timelines are hardly predictable. Completely



Power Implications for Reduced Sample Size

One-sided significance level: 0.025

Power (%): 90

Proportion (%) of data available: 85

Note: This may depend on the data structure. For example, for time-to-event outcomes this proportion refers to the number of patients with endpoints, relative to the expectation in the original design.

	Power (%)	Observed treatment effect* reaching significance
Original design	90.0	0.60
Available	84.8	0.66

Power vs. proportion of data available, under the hypothesized design effect
Power (%)
90
88
86
84
Proportion: 85
Power: 84.8
80
85
90
95
100
Proportion (%) of data available

Figure 3: Screenshot of the R shiny app.

overturning the design, while allowed from a methodological point of view, should remain the last resort if is really felt that this is the only way the trial can be salvaged.

The decision to introduce an interim analysis based on comparative data will generally be influenced by operational aspects including the development stage and purpose of the clinical trial. Further advice is given in Section 6.

4.5 Implementation of the resizing approaches in a R shiny app

To facilitate the implementation of the proposed methods, an R shiny app was developed as a simple-to-use web-based application. It provides insights into the power properties on the fly, given user-defined input of design parameters. Specifically, it has a module for the calculations shown in Section 4.1 to answer the following question: If a trial was designed for 90% power for an assumed treatment effect at a significance level $\alpha = 0.025$, what is the power if we conduct the analysis with only 85% of the patient data? By following (3), the app provides the power (84.8%) and a plot for different proportions of data available, in addition to 85%. A screenshot of the app is provided in Figure 3.

The app was originally designed to facilitate the discussion by Akacha et al (2020), where the same calculation as (3) was independently developed. The app is expanded to implement the group sequential design of an interim analysis conducted with data available and a final analysis when the planned data is obtained (see Section 4.3). Two popular group sequential designs are considered which are the Pocock and the O'Brien-Fleming schemes. In addition, the incorporation of dilution effects allows for more general considerations, as demonstrated in Section 4.2. Similar outputs as in displayed Figure 3 are provided with the app for the various scenarios considered above. The app can be accessed at `https://power-implications.shinyapps.io/prod/` and comes with a help tab that contains more information about its usage.

5 Adaptations other than trial duration or sample size

In the previous section we made some detailed comments on changing the trial duration or sample size. Here we consider other adaptations, starting with blinded adaptations and then continuing the discussion considering unblinded adaptations.

5.1 Blinded adaptations

The introduction provided an outline of the potential challenges for clinical trials by the COVID-19 pandemic. In order to assess the extent by which a trial is affected by these, blinded data may be used to investigate baseline patient characteristics, premature study or treatment discontinuations, missing data during follow-up, protocol violations, and nuisance parameters of the outcomes including event rates and variances. The findings may be compared with planning assumptions. Furthermore, time trends can be explored in the blinded data and any changes might be attributed to the COVID-19 if these coincide with the onset of the pandemic (see e.g. Friede and Henderson (2003)).

The findings of such blinded analyses might trigger investigations into resizing the trial. The resizing could be based on blinded or unblinded data; appropriate procedures will be considered in Section 4. However, adaptations are not restricted to sample size reestimation but also include other adaptations such as changes in the statistical model, test statistics or endpoint to be used. For instance, observed changes in baseline characteristics might be reflected in the statistical model by including additional covariates. Similarly, findings regarding missing data, e.g. due to missed visits, might suggest to adopt a more robust analysis approach. Additionally, if the new endpoint is no longer appropriate due to COVID-19 impact, modifying the endpoint midstream in a long-term clinical trial might be desired. One example for such a case could be studies targeting pneumonia as events might require exclusion of COVID-19 related pneumonia, if it becomes apparent that the event rates are severely increased.

5.2 Unblinded adaptations

The methods introduced in Section 3.1 can be used to create quite flexible designs while maintaining type I error rate control. As mentioned in Section 5.1 the need to change the statistical model or test statistics might arise. There we considered this based on blinded data. However, such changes can also be carried out following inspection of unblinded data. This has been considered by e.g. Kieser et al (2002); Friede et al (2003). Although these methods have not been used a lot as the change of a (primary) endpoint is somewhat controversial in confirmatory trials from a regulatory perspective. However, this might be quite different in a pandemic situation as currently experienced with SARS-CoV-2.

In Section 3.3 treatment effect heterogeneity was considered. With the rise of personalized medicine and targeted therapies adaptive enrichment designs starting with a broader population and zooming in on patient subgroups with particularly large benefit or reduced harm following an interim analysis have become popular over the past years Friede et al (2012); Stallard et al (2014); Friede et al (2020). These designs could potentially also be useful to modify eligibility criteria of trials affected by the COVID-19 pandemic.

6 Regulatory and operational aspects

The COVID-19 pandemic affects all clinical trials, with implications for studies intended for drug regulation well beyond statistical aspects (EMA 2020a,b; FDA 2020a,b). For example, on-site monitoring of most trials is suspended during the lockdown and with the interdiction of non-essential travels the recording of adverse events might not be as good if a site visit is replaced by a telephone consultation, or a local laboratory was used instead of the central laboratory. Similarly, the mode of administration of a patient reported outcomes questionnaire might have been changed from an electronic collection at the site on a tablet computer to a paper based version mailed to the patient's home. All these examples may lead to a reduced quality of the trial data which may need to be taken into consideration when interpreting the trial.

As much remains to be learned on the COVID-19 disease manifestations, treatments and pandemic distribution, it appears necessary to monitor the status and integrity of the trial on an ongoing basis. However, it may not be clear in some situations how this can be done in a way that protects the integrity of trial conduct. Care has to be taken that the original responsibilities of an Independent Data Monitoring Committee (IDMC) are not expanded beyond reasonable limits. Many of the responsibilities arising during the pandemic might more naturally seem to belong to trial management personnel, as the associated issues can often be addressed adequately without access to unblinded data; this might involve sponsor personnel, steering committees, etc. If important decisions are advised by unblinded results, then of course this should be done through an IDMC. But many other decisions may not require unblinded access. Some, including initiating a sample size re-assessment or updating a study's final statistical analysis plan (SAP), could be very problematic in terms of validly interpreting final analysis results if initiated by a party with access to unblinded interim results such as an IDMC. In general, IDMCs should not pro-actively initiate a sample size re-assessment scheme, as it has to be specified without knowledge of unblinded results. Likewise, changes to the SAP should not be in the scope of IDMC responsibilities. In current practice and supported by prior regulatory guidance, such decisions are generally initiated by parties remaining blinded. Of course the IDMC should be kept fully aware of any changes implemented in a trial, and should comment if they have any concerns. But for actions taken based upon blinded data, there are generally no confidentiality concerns, and sponsors can enlist any experts who can help arrive at the best decisions.

The introduction of an IDMC into a trial might be warranted if, for example, the trial design is changed from a fixed sample to an adaptive design. Establishing a qualified IDMC when one was not previously felt to be needed can be challenging and time consuming during the pandemic. Attempting to ensure that IDMC members have full understanding of all relevant background for the important tasks they will be assigned to, compared to trial personnel or steering committee members who will already have such perspective, could be risky. Thus, if an unblinded IDMC is felt necessary to be established, given the challenges of identifying and implementing such a group quickly, an internal firewalled group might thus be considered as an option in exceptional situations.

COVID-19 affects ongoing clinical trials in many different ways, which in turns affects many aspects of statistical inference, which are best described in the estimand framework laid out by ICH (2019). Some of complications resulting from the COVID-19 pandemic lead to unforeseen intercurrent events in the sense that they affect either the interpretation or the existence of the measurements associated with the clinical question of interest (ICH 2019), while others prevent relevant data being collected and result in a missing data problem. For pandemic-related unforeseen intercurrent events, we follow Akacha et al (2020) and Meyer et al (2020) in their recommendation to revise existing estimand definitions accordingly. As discussed in Section 3.2, however, care has to be taken when revising the estimand of interest in view of combining the information across different stages: If, for example, the estimands before and after the COVID-19 outbreak are different, this may limit the interpretability of the statistical inference.

The challenges imposed by the pandemic will lead to difficulties in meeting protocol-specified procedures in many instances, thus requiring the need to change aspects of ongoing trials. It is then important to be mindful about the fact that pre-specification of the study protocol and the SAP is the corner stone to avoid operational bias in any clinical trial. Although the ICH (1998) guideline allows changing the SAP even shortly before unblinding a trial, this if often viewed as critical by stakeholders. Changing the characteristics of a trial based on unblinded trial data always requires appropriate measures to control the type I error rate whereas changes triggered by external data are often seen more lenient. In the case of changes to the conduct and/or analysis of a trial caused by the pandemic it is reasonable to assume that such changes are not triggered by the knowledge gained from the ongoing trial. Still, changes will have to be pre-specified and documented, as appropriate. It is recommended to pre-specify key analyses important to interpret the objectives of the trial in the statistical analysis plan, in particular analyses related to the inferential testing strategy. Therefore we suggest to consider first whether different analyses are needed for the primary or key secondary objectives. Other analyses that have a more exploratory character can be included in a separate exploratory analysis planning. If any impact is detected that warrant additional analyses in the clinical study report, then these can be added later.

After the lockdown measures will be eased in future, the medical practice may not return to the state before the onset of the pandemic. Social distancing measures may be kept in place and it is to be expected that the trend of, for example, fewer hospital admissions for minor cases will continue to some extent. Nevertheless, certain trials interrupted by the pandemic will be able to restart, albeit in a possibly changed environment. The trial of the long acting contraceptive (Section 2.1) was largely unaffected by the onset of the pandemic. The START:REACTS trial (Section 2.2) had changed its endpoint to a PRO measure that can be observed remotely if a patient does not wish to come to the clinic. This trial can restart recruitment when elective surgeries will be again possible, albeit with the new endpoint as the original endpoint was not always measured during the lockdown measures. The ATALANTE 1 trial (Section 2.3) was stopped for ethical reasons due to the study population being at high risk of COVID-19. As the trial did not proceed to its second stage, consultations with agencies have started to discuss the partial results in view of the clinical unmet need. Such discussion will be likely focus also on the loss of power even if first stage was promising. This begs the question, how promising the first stage results should have been to provide convincing evidence if a dilution of the treatment effect cannot be excluded a priori and the considerations in Section 4 of this paper may support such discussions. Lastly, the CAPE-Covid and the CAPE-Cod studies (Section 2.4) are both addressing ICU patients with two kind of pneumonia. There is heterogeneity in disease and patients prognostic. For the moment, the CAPE-Cod trial is temporarily stopped but is planned to restart next autumn. As the investigators had no choice than to embed a trial within the other, heterogeneity will need to be addressed at the end of the study in order to preserve both results.

7 Discussion

The COVID-19 pandemic has not only led to a surge a clinical research activities in developing treatments, diagnostics and vaccines to fight the pandemic, but also impacted in many ways on ongoing trials. Here we illustrated the negative effects the pandemic might have on trials by giving four examples from ongoing studies and describing the considerations and consequences in reaction to the pandemic. Furthermore, we focused here on the role of adaptive designs in mitigating the risks of the pandemic which might result in a large number of inconclusive or misleading trials. Aspects that are of particular importance here are type I error rate control and treatment-effect heterogeneity. When trials are affected, the question to stop the trial early or to continue the trial, possibly with modifications is of particular interest. Considering normally distributed outcomes we developed a range of strategies. We believe that these are transferable to other types of outcomes with only limited modifications.

In Section 3.1 on type I error rate control we mention non-inferiority trials as one example where blinded adaptations may inflate the type I error rate. Otherwise we believe that extensions of the discussions and approaches proposed from superiority trials to non-inferiority trials are straightforward.

A Appendix

A.1 General joint distribution

The joint distribution of t_0, t_1, and t is given by

$$\begin{pmatrix} t_0 \\ t_1 \\ t \end{pmatrix} \sim N\left(\begin{pmatrix} \frac{\mu_{t0}-\mu_{c0}}{\sqrt{\frac{\sigma_{t0}^2}{n_{t0}}+\frac{\sigma_{c0}^2}{n_{c0}}}} \\ \frac{\mu_{t1}-\mu_{c1}}{\sqrt{\frac{\sigma_{t1}^2}{n_{t1}}+\frac{\sigma_{c1}^2}{n_{c1}}}} \\ \frac{\frac{n_{t0}\mu_{t0}+n_{t1}\mu_{t1}}{n_{t0}+n_{t1}}-\frac{n_{c0}\mu_{c0}+n_{c1}\mu_{c1}}{n_{c0}+n_{c1}}}{\sqrt{\frac{n_{t0}\sigma_{t0}^2+n_{t1}\sigma_{t1}^2}{(n_{t0}+n_{t1})^2}+\frac{n_{c0}\sigma_{c0}^2+n_{c1}\sigma_{c1}^2}{(n_{c0}+n_{c1})^2}}} \end{pmatrix},\right.$$

$$\left.\begin{pmatrix} 1 & & \\ 0 & 1 & \\ \frac{\frac{\sigma_{t0}^2}{n_{t0}+n_{t1}}+\frac{\sigma_{c0}^2}{n_{c0}+n_{c1}}}{\sqrt{\left(\frac{\sigma_{t0}^2}{n_{t0}}+\frac{\sigma_{c0}^2}{n_{c0}}\right)\left(\frac{n_{t0}\sigma_{t0}^2+n_{t1}\sigma_{t1}^2}{(n_{t0}+n_{t1})^2}+\frac{n_{c0}\sigma_{c0}^2+n_{c1}\sigma_{c1}^2}{(n_{c0}+n_{c1})^2}\right)}} & \frac{\frac{\sigma_{t1}^2}{n_{t0}+n_{t1}}+\frac{\sigma_{c1}^2}{n_{c0}+n_{c1}}}{\sqrt{\left(\frac{\sigma_{t1}^2}{n_{t1}}+\frac{\sigma_{c1}^2}{n_{c1}}\right)\left(\frac{n_{t0}\sigma_{t0}^2+n_{t1}\sigma_{t1}^2}{(n_{t0}+n_{t1})^2}+\frac{n_{c0}\sigma_{c0}^2+n_{c1}\sigma_{c1}^2}{(n_{c0}+n_{c1})^2}\right)}} & 1 \end{pmatrix}\right). \tag{11}$$

Equation (11) can then be rewritten as

$$\begin{pmatrix} t_0 \\ t_1 \\ t \end{pmatrix} \sim N \left(\begin{pmatrix} \sqrt{\frac{N\tau}{r+1}} \frac{\delta}{\sqrt{\frac{\sigma_{t0}^2}{r}+\sigma_{c0}^2}} \\ \sqrt{\frac{N(1-\tau)}{r+1}} \frac{(1-\eta)\delta}{\sqrt{\frac{\psi_t \sigma_{t0}^2}{r}+\psi_c \sigma_{c0}^2}} \\ \sqrt{\frac{N}{r+1}} \frac{\delta(\tau+(1-\tau)(1-\eta))}{\sqrt{\sigma_{t0}^2 \frac{\tau+(1-\tau)\psi_t}{r}+\sigma_{c0}^2(\tau+(1-\tau)\psi_c)}} \end{pmatrix}, \right.$$

$$\left. \begin{pmatrix} 1 & & \\ 0 & 1 & \\ \frac{\sqrt{\tau}\left(\frac{\sigma_{t0}^2}{r}+\sigma_{c0}^2\right)}{\sqrt{\left(\frac{\sigma_{t0}^2}{r}+\sigma_{c0}^2\right)\left(\sigma_{t0}^2 \frac{\tau+(1-\tau)\psi_t}{r}+\sigma_{c0}^2(\tau+(1-\tau)\psi_c)\right)}} & \frac{\sqrt{1-\tau}\left(\frac{\psi_t \sigma_{t0}^2}{r}+\psi_c \sigma_{c0}^2\right)}{\sqrt{\left(\frac{\psi_t \sigma_{t0}^2}{r}+\psi_c \sigma_{c0}^2\right)\left(\sigma_{t0}^2 \frac{\tau+(1-\tau)\psi_t}{r}+\sigma_{c0}^2(\tau+(1-\tau)\psi_c)\right)}} & 1 \end{pmatrix} \right) \tag{12}$$

A.2 Sample Size Adjustment

Substituting $n_0/(n_0+\tilde{n}_1)$ with $\xi = n_0/(n_0+\tilde{n}_1)$, we can rewrite Equation (7) as follows:

$$\begin{aligned} & \sqrt{\frac{(n_0+\tilde{n}_1)r}{(r+1)^2}} \cdot \frac{\frac{n_0}{n_0+\tilde{n}_1} + (1-\frac{n_0}{n_0+\tilde{n}_1})\,(1-\eta)}{\sqrt{\frac{n_0}{n_0+\tilde{n}_1} + (1-\frac{n_0}{n_0+\tilde{n}_1})\psi}} \cdot \frac{\delta}{\sigma} - z_{1-\alpha} &&= z_{1-\beta} \\ \Leftrightarrow \quad & \sqrt{\frac{\frac{n_0}{\xi} r}{(r+1)^2}} \cdot \frac{\xi + (1-\xi)\,(1-\eta)}{\sqrt{\xi+(1-\xi)\psi}} \cdot \frac{\delta}{\sigma} - z_{1-\alpha} &&= z_{1-\beta} \\ \Leftrightarrow \quad & \frac{n_0}{\xi} \cdot \frac{\left(\xi + (1-\xi)\,(1-\eta)\right)^2}{\xi+(1-\xi)\psi} &&= \left(z_{1-\alpha}+z_{1-\beta}\right)^2 \frac{\sigma^2}{\delta^2} \frac{(r+1)^2}{r} \end{aligned} \tag{13}$$

Replacing $n_0 = N\tau$ and also noticing that the right-hand side of the equation also equals N, we obtain

$$\begin{aligned} \Leftrightarrow \quad & \frac{\tau}{\xi} \cdot \frac{\left(\xi + (1-\xi)\,(1-\eta)\right)^2}{\xi+(1-\xi)\psi} &&= 1 \\ \Leftrightarrow \quad & \xi^2 \left(\tau\eta^2 - 1 + \psi\right) + \xi \left(2\tau\eta\,(1-\eta) - \psi\right) + \tau\,(1-\eta)^2 &&= 0. \end{aligned} \tag{14}$$

Now, if $\tau\eta^2 - 1 + \psi = 0$ and replacing $\psi = 1 - \tau\eta^2$, we get

$$\xi = \frac{\tau\,(1-\eta)^2}{1-\tau\eta\,(2-\eta)}. \tag{15}$$

For $\tau\eta^2 - 1 + \psi \neq 0$, we obtain

$$\xi = \frac{\psi - 2\tau\eta(1-\eta) \pm \sqrt{\psi^2 - 4\tau(1-\eta)(\eta+\psi-1)}}{2\,(\psi - 1 + \tau\eta^2)}. \tag{16}$$

Re-substitution of ξ finally yields

$$\begin{aligned}\tilde{n}_1 =& N\tau\frac{1-\xi}{\xi}\\ =& N\tau\frac{\psi - 2 + 2\tau\eta \mp \sqrt{\psi^2 - 4\tau(1-\eta)(\eta+\psi-1)}}{\psi - 2\tau\eta(1-\eta) \pm \sqrt{\psi^2 - 4\tau(1-\eta)(\eta+\psi-1)}}\end{aligned} \tag{17}$$

As can be seen from Equations (16) and (17), ξ and hence $\tilde{n}_1$ have two different solutions due to the square root. Evaluating both solutions for different values of τ, η, and ψ show that only the second solution $(+\sqrt{\ }$ in the numerator and $-\sqrt{\ }$ in the denominator lead to a positive number for the sample size.

Acknowledgment

We are grateful to the Editor and the Reviewers for their detailed comments. This work was inspired by discussions led by Lisa Hampson and Werner Brannath in the joint working group "Adaptive Designs and Multiple Testing Procedures" of the German Region (DR) and the Austro-Swiss Region (ROeS) of the International Biometric Society. The authors thank Andrea Schulze and Andy Metcalfe for their contributions to Examples 2.1 and 2.2, respectively. Furthermore, the authors are grateful to Gernot Wassmer for comments on the R code used in the R shiny app and to Werner Brannath for comments on an earlier version of this manuscript. Kelly Van Lancker was funded by VLAIO (Flemish Innovation and Entrepreneurship) under the Baekeland grant agreement HBC.2017.0219.

Conflict of Interest

The authors have declared no conflict of interest.

References

Akacha, M., Branson, J., Bretz, F., Dharan, B., Gallo, P., Gathmann, I., Hemmings, R., Jones, J., Xi, D., and Zuber, E. (2020), "Challenges in Assessing the Impact of the COVID-19 Pandemic on the Integrity and Interpretability of Clinical Trials," *Statistics in Biopharmaceutical Research*, DOI: 10.1080/19466315.2020.1788984.

Akacha, M., Bretz, F., Ohlssen, D., Rosenkranz, G., and Schmidli, H. (2017), “Estimands and their role in clinical trials,” *Statistics in Biopharmaceutical Research*, 9, 268–271.

Akacha, M., Bretz, F., and Ruberg, S. (2017), “Estimands in clinical trials-broadening the perspective,” *Statistics in Medicine*, 36, 5–19.

Anker, S. D., Butler, J., Khan, M. S., Abraham, W. T., Bauersachs, J., Bocchi, E., Bozkurt, B., Braunwald, E., Chopra, V. K., Cleland, J. G., Ezekowitz, J., Filippatos, G., Friede, T., Hernandez, A. F., Lam, C. S. P., Lindenfeld, J., McMurray, J. J. V., Mehra, M., Metra, M., Packer, M., Pieske, B., Pocock, S. J., Ponikowski, P., Rosano, G. M. C., Teerlink, J. R., Tsutsui, H., Van Veldhuisen, D. J., Verma, S., Voors, A. A., Wittes, J., Zannad, F., Zhang, J., Seferovic, P., and Coats, A. J. S. (2020), “Conducting clinical trials in heart failure during (and after) the COVID-19 pandemic: an Expert Consensus Position Paper from the Heart Failure Association (HFA) of the European Society of Cardiology (ESC),” *European Heart Journal*, 41(22), 2109-2117.

Armitage, P., McPherson, C.K., and Rowe, B.C. (1969), “Repeated Significance Tests on Accumulating Data,” *Journal of the Royal Statistical Society, Series A*, 132, 235-244.

Bauer, P., and Posch, M. (2004), “Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections,” by H. Schäfer and H.-H. Müller, Statistics in Medicine 2001; 20: 3741–3751, *Statistics in Medicine*, 23, 1333–1334.

Brannath, W., Posch, M., and Bauer, P. (2002), “Recursive combination tests,” *Journal of the American Statistical Association*, 97, 236–244.

Bretz, F., Schmidli, H., König, F., Racine, A. and Maurer, W. (2006), “Confirmatory seamless phase II/III clinical trials with hypothesis selection at interim: general concepts,” *Biometrical Journal*, 48, 623–634.

Brückner, M., Burger U., and Brannath W. (2018), “Non-parametric adaptive enrichment designs using categorical surrogate data,” *Statistics in Medicine*, 39, 4507–4524.

Dequin, P.-F., Le Gouge, A., Tavernier, E., Giraudeau, B., and Zohar, S. (2020), “Embedding a COVID-19 group sequential clinical trial within an ongoing trial: lessons from an unusual experience,” *(Submitted to Statistics in Biopharmaceutical Research).*

European Medicines Agency (2007), “Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design,” October 2007.

European Medicines Agency (2020a), “Guidance on the management of clinical trials during the COVID-19 (coronavirus) pandemic,” Version 3, 28 April 2020.

European Medicines Agency (2020b), “Points to consider on implications of Coronavirus disease (Covid-19) on methodological aspects of ongoing clinical trials,” June 2020.

Food and Drug Administration (2019), “ Guidance for industry: Adaptive design clinical trials for drugs and biologics,” 2019.

Food and Drug Administration (2020a), “FDA Guidance on Conduct of Clinical Trials of Medical Products during COVID-19 Public Health Emergency: Guidance for Industry, Investigators, and Institutional Review Boards,” May 2020.

Food and Drug Administration (2020b), “Statistical Considerations for Clinical Trials During the COVID-19 Public Health Emergency: Guidance for Industry,” June 2020.

Friede, T., and Henderson, R. (2003). “Intervention effects in observational survival studies with an application in total hip replacements,” *Statistics Medicine*, 22, 3725–3737.

Friede, T., and Henderson, R. (2009), “Exploring changes in treatment effects across design stages in adaptive trials,” *Pharmaceutical Statistics*, 8, 62–72.

Friede, T., and Kieser, M. (2003), “Blinded sample size reassessment in non-inferiority and equivalence trials,” *Statistics in Medicine*, 22, 995–1007.

Friede, T., and Kieser, M. (2013), “Blinded sample size re-estimation in superiority and non-inferiority trials: Bias versus variance in variance estimation,” *Pharmaceutical Statistics*, 12, 141–146.

Friede, T., Kieser, M., Neuhäuser, M., and Büning, H. (2003), “A comparison of procedures for adaptive choice of location tests in flexible two-stage designs,” *Biometrical Journal*, 45, 292–310.

Friede, T., Pohlmann, H., and Schmidli, H. (2019), “Blinded sample size reestimation in event-driven clinical trials: Methods and an application in multiple sclerosis,” *Pharmaceutical Statistics*, 18, 351–365.

Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes-Marquez, E., Chataway, J., and Nicholas, R. (2011), “Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: an application in multiple sclerosis,” *Statistics in Medicine*, 30, 1528–1540.

Friede, T., Parsons, N., and Stallard, N. (2012), “A conditional error function approach for subgroup selection in adaptive clinical trials,” *Statistics in Medicine*, 31, 4309–4320.

Friede, T., Röver, C., Wandel, S., and Neuenschwander, B. (2017), “Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases,” *Biometrical Journal*, 59, 658–671.

Friede, T., Stallard, N., and Parsons, N. (2020), “Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R,” *Biometrical Journal*, DOI: 10.1002/bimj.201900020.

Friede, T. and Stammer, H. (2010), “Blinded sample size recalculation in non-inferiority trials: A case study in dermatology,” *Drug Information Journal*, 44, 599–607.

Galbraith, S. and Marschner, I.C. (2003), “Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes,” *Statistics in Medicine*, 22, 1787–1805.

Gerlinger, C., Endrikat, J., van der Meulen, E.A., Dieben, T. O. M., and Düsterberg, B. (2003), “Recommendation for confidence interval and sample size calculation for the Pearl Index,” *The European Journal of Contraception & Reproductive Health Care*, 8, 87–92.

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei,C., Hui, D.S.C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., Li, S., Wang, J.-l., Liang, Z., Peng, Y., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J., Chen, Z., Li, G., Zheng, Z., Qiu, S., Luo, J., Ye, C., Zhu, S., and Zhong, N. for the China Medical Treatment Expert Group for Covid-19 (2020), “Clinical Characteristics of Coronavirus Disease 2019 in China.,” *New England Journal of Medicine*, 382, 1708–1720.

ICH (1998), “ICH E9 guideline on statistical principles for clinical trials,” 1998.

ICH (2019). “ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials,” November 2019.

Jenkins, M., Stone, A., and Jennison, C. (2011). “An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints,” *Pharmaceutical Statistics*, 10, 347–356.

Jensen, JT., Kroll, R., Lynen, RF., Schulze, A., and Lukkari-Lax, E. (2020), “Contraceptive Efficacy and Safety of 52 mg LNG-IUS for up to Eight Years: Year 6 Data From the Mirena Extension Trial,” *Obstetrics & Gynecology*, 135, 6S.

Jörgens, S., Wassmer, G., König, F., and Posch, M. (2019), “Nested combination tests with a time-to-event endpoint using a short-term endpoint for design adaptations,” *Pharmaceutical Statistics*, 18, 329–350.

Hernan, M.A., and Robins, J.M. (2020). *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC.

Hobbs, B.P., Sargent, D.J., and Carlin, B.P. (2012), “Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models,” *Bayesian Analysis*, 7, 639–674.

Hinckley, D.V. (1969), “On the Ratio of Two Correlated Normal Random Variables,” *Biometrika*, 56, 635–639.

Kieser, M., Schneider, B., and Friede, T.(2002), “A bootstrap procedure for adaptive selection of the test statistic in flexible two-stage designs,” *Biometrical Journal*, 44, 641–652.

Kim, K., and DeMets, D.L. (1987), “Design and analysis of group sequential tests based on the type I error spending rate function,” *Biometrika*, 74, 149–154.

Kim, K., and DeMets, D.L. (1987), “Confidence Intervals Following Group Sequential Tests in Clinical Trials,” *Biometrics*, 43, 857–864.

Klinglmüller, F., Posch, M., and Koenig, F. (2014), “Adaptive graph-based multiple testing procedures,” *Pharmaceutical Statistics*, 13, 345–356.

Lan, K.K.G., and DeMets, D.L. (1983), “Discrete sequential boundaries for clinical trials,” *Biometrika*, 70, 659–663.

Lipkovich, I., Ratitch, B., and Mallinckrodt, C.H. (2020), “Causal inference and estimands in clinical trials,” *Statistics in Biopharmaceutical Research*, 12, 54–67.

Mansour, D., Inki, P., and Gemzell-Danielsson, K. (2010), “Efficacy of contraceptive methods: A review of the literature,” *The European Journal of Contraception & Reproductive Health Care*, 15, S19–S31.

Metcalfe, A., Gemperle Mannion, E., Parsons, H., Brown, J., Parsons, N., Fox, J., Kearney, R., Lawrence, T., Bush, H., McGowan, K., Khan, I., Mason, J., Hutchinson, C., Gates, S., Stallard, N., Underwood, M., and Drew, S. (2020), “Protocol for a randomised controlled trial of Subacromial spacers for Tears Affecting Rotator cuff Tendons: a Randomised, Efficient, Adaptive Clinical Trial in Surgery (START:REACTS),” *BMJ Open*, 10, e036829, DOI:10.1136/bmjopen-2020-036829.

Meyer, R.D., Ratitch, B., Wolbers, M., Marchenko, O., Quan, H., Li, D., Fletcher, C., Li, X., Wright, D., Shentu, Y., Englert, S., Shen, W., Dey, J., Liu, T., Zhou, M., Bohidar, N., Zhao, P.-L., and Hale, M. (2020), “Statistical issues and recommendations for clinical trials conducted during the COVID-19 pandemic,” *Statistics in Biopharmaceutical Research*, DOI: 10.1080/19466315.2020.1779122.

Müller, H.-H. and Schäfer, H. (2001), “Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches,” *Biometrics*, 57, 886–891.

Niewczas, J., Kunz, C.U., and König, F. (2019), "Interim analysis incorporating short- and long-term binary endpoints," *Biometrical Journal*, 61, 665–687.

Ollier, A., Morita, S., Ursino, M., and Zohar, S. (2019), "An adaptive power prior for sequential clinical trials: Application to bridging studies," *Statistical Methods in Medical Research*, DOI: 10.1177/0962280219886609.

ose-immuno therapeutics, press release April 1st 2020, `https://ose-immuno.com/wp-content/uploads/2020/04/FR_200401_IDMC-Atalante_VF.pdf`.

ose-immuno therapeutics, press release April 4th 2020, `https://ose-immuno.com/wp-content/uploads/2020/04/OSE_BIOTECHFINANCES_6-avril-2020.pdf`.

Parsons, N., Stallard, N., Parsons, H., Wells, P., Underwood, M., Mason, J., and Metcalfe, A. (2019), "An adaptive two-arm clinical trial using early endpoints to inform decision making: design for a study of sub-acromial spacers for repair of rotator cuff tendon tear," *Trials*, 20, 694.

Proschan, M.A., Follmann, D.A., and Waclawiw, M.A. (1992), "Effects of assumption violations on type I error rate in group sequential monitoring," *Biometrics*, 48, 1131–1143.

Proschan, M.A. and Hunsberger, S.A. (1995), "Designed extension of studies based on conditional power," *Biometrics*, 51, 1315–1324.

Röver, C., Andreas, S., and Friede, T. (2016), "Evidence synthesis for count distributions based on heterogeneous and incomplete aggregated data," *Biometrical Journal*, 58, 170–185.

Rufibach, K., Chen, M., and Nguyen, H. (2016), "Comparison of different clinical development plans for confirmatory subpopulation selection," *Contemporary Clinical Trials*, 47, 78–84.

Shu, H. and Tan, Z. (2018), "Improved Estimation of Average Treatment Effects on the Treated: Local Efficiency, Double Robustness, and Beyond," arXiv e-prints, arXiv:1808.01408.

Sooriyarachchi, M.R., Whitehead, J., Whitehead, A., and Bolland, K. (2006), "The sequential analysis of repeated binary responses: a score test for the case of three time points," *Statistics in Medicine*, 25, 2196–2214.

Stallard, N. (2010), "A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information," *Statistics in Medicine*, 29, 959–971.

Stallard, N., Hamborg, T., Parsons, N., and Friede, T. (2014), "Adaptive designs for confirmatory clinical trials with subgroup selection," *Journal of Biopharmaceutical Statistics*, 24, 168–187.

Stallard, N., Hampson, L., Benda, N., Brannath, W., Burnett, T., Friede, T., Kimani, P.K., Koenig, F, Krisam, J., Mozgunov, P., Posch, M., Wason, J., Wassmer, G., Whitehead, J., Williamson, S.F., Zohar, S., and Jaki, T. (2020), "Efficient adaptive designs for clinical trials of interventions for COVID-19," *Statistics in Biopharmaceutical Research*, DOI: 10.1080/19466315.2020.1790415.

Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., and Carpenter, J.R. (2009), "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, 338, b2393.

Thall, P.F., and Wathen, J.K. (2008), "Bayesian designs to account for patient heterogeneity in phase II clinical trials," *Current Opinion in Oncology*, 20, 407–411.

Van Lancker, K., Vandebosch, A., Vansteelandt, S., and De Ridder, F. (2019), "Evaluating futility of a binary clinical endpoint using early read-outs," *Statistics in Medicine*, 38, 5361–5375.

Van Lancker, K., Vandebosch, A., and Vansteelandt, S. (2020), "Improving interim decisions in randomized trials by exploiting information on short-term endpoints and prognostic baseline covariates," *Pharmaceutical Statistics*, DOI: 10.1002/pst.2014.

Wassmer, G. (1998), "A comparison of two methods for adaptive interim analyses in clinical trials," *Biometrics*, 54, 696–705.

Wassmer, G., and Brannath, W. (2016), *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer.