Online control of the False Discovery Rate in group-sequential platform trials

Sonja Zehetmayer, Martin Posch, and Franz Koenig

Abstract
When testing multiple hypotheses, a suitable error rate should be controlled even in exploratory trials. Conventional methods to control the False Discovery Rate assume that all $p$-values are available at the time point of test decision. In platform trials, however, treatment arms enter and leave the trial at different times during its conduct. Therefore, the actual number of treatments and hypothesis tests is not fixed in advance and hypotheses are not tested at once, but sequentially. Recently, for such a setting the concept of online control of the False Discovery Rate was introduced. We propose several heuristic variations of the LOND procedure (significance Levels based On Number of Discoveries) that incorporate interim analyses for platform trials, and study their online False Discovery Rate via simulations. To adjust for the interim looks spending functions are applied with O’Brien-Fleming or Pocock type group-sequential boundaries. The power depends on the prior distribution of effect sizes, for example, whether true alternatives are uniformly distributed over time or not. We consider the choice of design parameters for the LOND procedure to maximize the overall power and investigate the impact on the False Discovery Rate by including both concurrent and non-concurrent control data.

Keywords
Multiple hypothesis testing, interim analysis, spending function, LOND procedure

1 Introduction
Platform trials are an innovative type of study design, where randomized clinical trials with related aims or questions are combined to improve efficiency by reducing costs or saving time.1,2 Treatment arms can enter and leave the trial at different times during its conduct and the actual number of hypothesis tests eventually to be performed in the platform trial is not fixed in advance. One major benefit of platform trials is the comparison of different treatment arms to one shared control arm and a therefore reduced number of control patients. As more than one statistical test is performed, an adjustment for testing multiple hypotheses has been proposed,3,4 for example, control of the Family Wise Error Rate (FWER) or the False Discovery Rate (FDR), defined as the expected proportion of false rejections under all rejections.5 Conventional methods, however, such as the Bonferroni method to control the FWER or the Benjamini-Hochberg (BH) method to control the FDR, require that the number of hypothesis tests is fixed and, for the BH method, that all $p$-values are available at the time point of test decision. These methods are thus not appropriate for the special design of platform trials. Recently, the concept of online control of the FDR6,7 or the FWER8 was introduced where hypothesis tests and test decisions can be performed sequentially while guaranteeing FDR control. At each step a decision has to be performed if the current null hypothesis should be rejected based on previous decisions, but without knowledge on future $p$-values or the number of hypotheses to be tested.

To control the online FDR, several procedures have been proposed for independent test statistics, as for example, the LORD,7 or the SAFFRON method.9 For these, however, it is so far not proven that they control the online FDR for...
positively dependent test statistics as it is the case in platform trials due to a shared control group. The LOND procedure (significance Levels based On Number of Discoveries) is a method to control the online FDR based on the observed p-values and the number of previous rejections. There are modifications of the LOND procedure which allow a specification of a maximum number of hypotheses to be tested within the platform trials, but again the actual number of treatments and corresponding null hypotheses have not to be fixed in the beginning. A proof has been given that the FDR is controlled also for positively dependent test statistics for a design with fixed sample sizes, throughout the paper online control of the FDR will be performed with the LOND procedure. A detailed comparison of online FDR procedures can be found in the literature, for example, Robertson and Wason and Robertson et al. (see also the R-package onlineFDR).

The aim of this paper is to explore the LOND procedure for platform trials: In previous work, only designs with fixed sample sizes for the online control of the FDR were investigated; we now present a framework for incorporating group-sequential hypothesis tests with the option of early efficacy or futility stopping for individual hypotheses and investigate the online FDR of the group-sequential experiments by simulations. A proof sketch is given in the supplemental material that the gsLOND procedure controls the online FDR for independent test statistics (for the updated LOND, following the proof by Znic et al.11 for the fixed sample design). In the next section, we review the LOND procedure and propose an extension for the online FDR control of group-sequential designs (gsLOND) as well as two modifications for gsLOND. The procedures are investigated in simulation studies in the Results section. The average power, defined as the proportion of rejected alternatives among all alternatives will be analysed for different priors for the distribution of the alternative hypotheses or several methods distributing the alpha level among the hypotheses of interest. Scenarios including both concurrent and non-concurrent control data will be investigated.

2 Methods
Consider a sequence of null hypotheses $H_1, H_2, \ldots$ and corresponding p-values $p_1, p_2, \ldots$, where a maximum number of hypotheses is not pre-fixed in advance. At each step $i$, a decision has to be made whether to retain or reject $H_i$ based on information on the p-values of the previous hypotheses $H_1, \ldots, H_{i-1}$, only, but without information on the outcome of the future hypothesis tests or the eventual total number of hypotheses to be tested in the platform. Thus, in online testing, the significance level $\alpha_i$ for hypothesis $i$ is only a function of the decisions for $H_1, \ldots, H_{i-1}$. Online FDR algorithms aim at FDR control at a pre-specified significance level $\alpha$ after each test decision for treatment $i$, that is,

\[
\text{FDR}_i = \mathbb{E}\left( \frac{V(i)}{\max(\sum_{j=1}^{i} R_j, 1)} \right) \leq \alpha
\]

with $V(i)$ denoting the number of incorrectly rejected hypotheses among the first $i$ hypotheses and test decision $R_j = 1$ if $p_j \leq \alpha_i$ (reject $H_j$) and $R_j = 0$ otherwise.

2.1 The LOND procedure – Significance levels based on the number of discoveries
Given an overall significance level $\alpha$, a sequence of non-negative numbers $\beta_i$ is determined before starting the trial, such that $\sum_{i=1}^{\infty} \beta_i = \alpha$.

In the LOND procedure, the values of the nominal significance levels $\alpha_i$ for $H_i$ are given by:

\[
\alpha_i = \beta_i \left( \sum_{j=1}^{i-1} R_j + 1 \right)
\]

Znic et al. showed that for the updated LOND with $\alpha_i = \beta_i \left( \sum_{j=1}^{i-1} R_j \lor 1 \right)$, the online FDR is controlled at level $\alpha$ for positively dependent p-values (i.e. that satisfy positive regression dependency on a subset (PRDS)).

2.2 The LOND procedure in platform trials
In platform trials treatment arms enter and leave the trial at different times, possibly depending on previous results or available resources. Treatment arms often share one control arm (see Figure 1) and hypothesis testing is performed for each treatment arm against the control arm. Thus for the remainder of the paper $H_i$ denotes the null hypothesis for comparing treatment arm $i$ to the data of the common control arm (indicated by the orange bar in Figure 1). The actual number of
hypotheses to be tested within a platform trial is denoted by $K$. This number, however, has not to be fixed in the beginning. For some modification of the LOND procedure the maximum number of hypotheses $N$ to be tested within the platform trial is specified at the beginning with $K \leq N$.10

In principle the methodology presented would allow that for $H_i$ different endpoints are tested. For simplicity, we assume that the same endpoint is used for each treatment-control comparison. To control the FDR of these trials with the LOND procedure, the sequence in which the hypotheses $H_1, H_2, \ldots$ will be tested has to be predefined, for example, in a platform trial by the order of entering the trial. In the following we assume equal sample sizes $n_i$ per treatment and equal allocation of observations to each treatment and the control arms, thus the sequence of hypotheses entering the study is the same as for data analysis. In case of a fixed sample design with different sample sizes $n_i$ per treatment, the sequence of observing individual $p$-values might be different than the sequence defined by the starting times. To control the online FDR in this scenario it is required to maintain the predefined sequence of the hypotheses for the allocation of the significance levels in equation (1).11 Note that it is possible to fix the order not according to the starting times but according to the anticipated availability of the data for each $H_i$. It is, however, essential that the order of the hypotheses remains independent of the observed data and may thus not be influenced by the outcome. For example, as long as (i) there are no interim analyses, (ii) the sample sizes are fixed and (iii) the allocation ratios do not depend on outcome data, then we can use the order induced by the actual availability of the data for each $H_i$.

For a hypothesis test the control group data can be divided into two parts:

- Concurrent (CC) controls: Control samples which are recruited parallel to the treatment group.
- Non-concurrent (NCC) controls: Control samples which are recruited from the beginning of the platform trial until the beginning of the treatment of interest.

For platform trials running over a long time period, incorporating a pooled control group of NCC and CC controls can substantially improve the power of the experiment, however, NCC controls may affect the FDR negatively in case of time trends.14,15
2.3 Group-sequential designs

It is now assumed that for each treatment an interim analysis is performed after having observed a part of the pre-planned total sample size with the option of early rejection or early stopping for futility. In these group-sequential trials two sources of multiplicity must be considered:

1. Adjust for the number of hypotheses and control the online FDR with the LOND procedure (as for the design with fixed sample sizes).
2. Adjust for the option of early rejection or stopping for futility in the interim analysis. This can be done, for example, with spending functions, where the significance level \( \alpha_i \) is split between the first and the second stage.

For hypothesis \( H_i \), the interim analysis is performed after first stage sample size \( n_i^{(1)} \) for the treatment group and the final analysis after \( n_i = n_i^{(1)} + n_i^{(2)} \) with \( n_i^{(2)} \) denoting the sample size of the treatment group in the second stage. The respective first stage sample size of the control group depends on the chosen strategy of incorporating only CC or also NCC controls and the allocation ratio of controls and treatments. The p-value for the interim analysis of \( H_i \) is denoted by \( p_i^{(1)} \). For the final analysis the data of stage one and stage two are pooled and the p-value is given by \( p_i \). The group-sequential boundaries allowing for interim analyses with the option of early rejection for individual hypotheses are determined using Lan-DeMets spending functions. Let \( t_i^* \) denote the fraction of information, defined as the percentage sample size already observed for \( H_i \) at time \( t_i \), with \( t_i \in [0, T_i] \). An \( \alpha \)-spending function \( \alpha_i^*(\alpha_i) \) is defined for each \( H_i \) and corresponding nominal significance level \( \alpha_i \) such that \( \alpha_i^*(0) = 0 \) and \( \alpha_i^*(\alpha_i) = \alpha_i \). For simplicity we assume only two stages for each \( H_i \) and denote the fraction of information of the interim analysis by \( t_i^* = (1) \) and of the final analysis by \( t_i^* = (2) \). Then the \( \alpha \)-spending function for hypothesis \( i \) for the interim analysis gives the adjusted group-sequential critical boundaries \( \alpha_i^{(1)}(\alpha_i) \) and for the final analysis \( \alpha_i^{(2)}(\alpha_i) \). For the \( \alpha \)-spending function, for example, O’Brien-Fleming (OBF) or Pocock (PO) type boundaries as a function of \( t_i^* \) and \( \alpha_i \) may be considered (for the one-sided case): \( \alpha_i^{\text{OBF}} = 2(1 - \Phi^{-1}(1 - \alpha_i/2)/\sqrt(t_i^*)) \) (approximate O’Brien-Fleming type group-sequential boundary with \( \Phi \) denoting the standard normal cumulative distribution function) and \( \alpha_i^{\text{PO}} = \alpha_i \ln(1 + (e - 1)t_i^*) \) (approximate Pocock type group-sequential boundary) (see, e.g. Wassmer and Brannath, page 75).

\( H_i \) is then rejected in the interim analysis if \( p_i^{(1)} \leq \alpha_i^{(1)}(\alpha_i) \) and the sample size of the treatment group in the second stage, \( n_i^{(2)} \), is saved. If not, a second stage is performed subsequently and \( H_i \) is rejected in the final analysis if \( p_i \leq \alpha_i^{(2)}(\alpha_i) \). If additionally a stopping for futility boundary \( \alpha_i^f \) is introduced, then in case of \( p_i^{(1)} \geq \alpha_i^f \), \( H_i \) is stopped in the interim analysis without rejecting \( H_i \) and again the sample size \( n_i^{(2)} \) is saved. If \( \alpha_i^{(1)}(\alpha_i) < p_i^{(1)} < \alpha_i^f \), a second stage is performed and \( H_i \) is rejected in the final analysis if \( p_i \leq \alpha_i^{(2)}(\alpha_i) \).

To control the online FDR for the group-sequential trial, the LOND procedure is applied for both the interim and the final analyses. \( \alpha_i \) depends on the number of previously rejected hypotheses, thus, it may differ between the interim and final analysis of \( H_i \). In the following three different LOND procedures for group-sequential designs are proposed.

- Group-sequential LOND: gsLOND
- Group-sequential LOND exhausting updated local level alpha: gsLOND.II
- Group-sequential LOND updating all ongoing tests in case of any “in-between” rejections: gsLOND.III

The methods presented are straightforward for binary or continuous endpoints (without delayed response). For time-to-event endpoints in a gsLOND procedure further specifications and investigations are needed, which are not covered here. This includes, for example, timing of interim and final analyses which may be based on events or number of patients.

2.3.1 Group-sequential LOND (gsLOND)

We first illustrate the challenges of a group-sequential design with the LOND procedure based on Figure 1(b): According to LOND, the value of the nominal level for the interim analysis of \( H_3 \) depends on the test decision for \( H_1 \) and the interim decision for \( H_2 \). If \( H_2 \) proceeds to the final analysis, the level additionally may be influenced by the final analysis of \( H_3 \): If either \( H_2 \) was rejected in the interim analysis, or \( H_2 \) was neither rejected in the interim nor the final analysis, the number of rejections does not change and the level for \( H_2 \) remains the same. If, however, \( H_2 \) was not rejected in the interim analysis, but indeed in the final analysis, the number of rejections increases by one and the local level for \( H_2 \) has therefore to be updated and increased.
More formally, the significance level for testing $H_i$ at the interim analysis is defined by

$$\alpha_i^{(1)} = \beta_i \left( \sum_{j \in I_i^{(1)}} R_j^{(1)} + 1 \right),$$  

(2)

and the corresponding group-sequential critical boundary is given by $\alpha_i^{(1)}(\beta_i^{(1)}). I_i^{(1)}$ is the index set of all hypotheses $j, j < i$, for which, before the time of the interim analysis of treatment $i$, a test (either at an interim or final analysis) has been performed. $R_j^{(1)} = 1$ for $j \in I_i^{(1)}$ if $H_j$ has been already rejected (at the interim or the final analysis) before the interim analysis of treatment $i$ and it is zero, if no rejection has been performed yet.

The level of the significance level for hypothesis $i$ may increase from interim analysis to final analysis as the number of rejections may increase. For the final analysis of $H_i$, the corresponding index set is given by $I_i^{(2)}$ including all $j, j < i$, where for $H_j$ already group-sequential tests have been performed before the final analysis of $H_i$ and the significance level is given by

$$\alpha_i^{(2)} = \beta_i \left( \sum_{j \in I_i^{(2)}} R_j^{(2)} + 1 \right).$$

(3)

The group-sequential critical boundary for the final analysis is then $\alpha_i^{(2)}(\beta_i^{(2)}).

If no hypotheses are rejected between interim analysis and final analysis of $H_i$, for all $j < i$, $R_j^{(1)} = R_j^{(2)}$, and $\alpha_i^{(1)} = \alpha_i^{(2)}$.

### 2.3.2 Group-sequential LOND exhausting updated local level alpha (gsLOND.II)

If in the gsLOND procedure the nominal level for the group-sequential test for $H_i$ is increased between the interim analysis and the final analysis from $\alpha_i^{(2)}$ to $\alpha_i^{(2)}$ (with $\alpha_i^{(2)} < \alpha_i^{(2)}$) due to further rejections of some $H_j$ (for $j < i$), then we could actually reject $H_i$ at the final analysis not only if $p_i \leq \alpha_i^{(2)}(\alpha_i^{(2)})$, but also if only $p_i^{(1)} \leq \alpha_i^{(1)}(\alpha_i^{(2)})$ holds. The latter means that we would re-perform the test at interim analysis using an increased interim significance level. As $\alpha_i^{(1)}(\beta_i^{(1)}) < \alpha_i^{(1)}(\beta_i^{(2)})$ when using the same type of spending function with an increased nominal level $\alpha_i$ for the group-sequential test, we can only get additional rejections for $H_i$. However, it might be unusual to reject the null hypothesis using only part of the data in a situation where all data are available (and especially when all data suggest to retain the null hypothesis with $p_i > \alpha_i^{(2)}(\beta_i^{(2)})$). So we consider a different testing strategy utilizing the alpha not used in the group-sequential testing in the situation where the nominal level alpha is increased between interim and final analysis due to other additional in-between rejections.

Instead of using the same type of spending function with an increased nominal level $\alpha_i^{(2)}$, we propose to use the increment $\alpha_i^{(1)} - \alpha_i^{(2)}$ when calculating the final level for $H_i$ in case further rejections of some $H_j$ (for $j < i$) occurred after the interim analysis of $H_i$. $\alpha_i^{(1)}(\beta_i^{(1)})$ corresponds to the alpha already spent at the interim analysis (where the corresponding hypothesis $i$ could not be rejected) and overall the level $\alpha_i^{(1)}$ is exhausted for $H_i$. Thus, if the nominal level $\alpha_i^{(2)}$ is increased, the final group-sequential plan does not correspond to the initially selected type of spending function, but exhausts fully the local nominal level as determined by the LOND procedure.

Thus, for gsLOND.II, we extend the gsLOND procedure by modifying the group-sequential boundaries when updating the local nominal level alpha. $\alpha_i^{(2)}$ and $\alpha_i^{(2)}$ are derived as described for the gsLOND in equations (2) and (3). However, the level for the final analysis is now modified by recalculating the group-sequential boundary based on the increment $\alpha_i^{(2)} - \alpha_i^{(2)}(\alpha_i^{(1)})$.

### 2.3.3 Group-sequential LOND - update of all ongoing tests in case of any “in-between” rejections (gsLOND.III)

In the group-sequential design, rejections for the hypotheses may occur non-consecutively: As illustrated in Figure 1(b), even if $H_2$ is still running, $H_3$ may already have been rejected in the interim analysis. However, for gsLOND, the test decisions for $H_1, H_{i+1}, \ldots$ have no influence on the significance level of $H_{i-1}$.

In this section, we propose a group-sequential LOND modifying the set of hypotheses which will be updated in case of any “in-between” rejections. For gsLOND.III we propose to update the local nominal significance level for $H_i$ (which is still under investigation) not only on test decisions on $H_1, \ldots, H_{i-1}$, but also on test decisions for $H_{i+1}, H_{i+2}, \ldots$, if they were conducted before the (interim or final) analysis of $H_i$. Note that this does not involve re-decision of a hypothesis test if a new test decision is performed.

$$\alpha_i^{(i)} = \beta_i \left( \sum_{j \in I_i^{(i)}} R_j^{(i)} + 1 \right)$$
and

\[ \alpha^\text{gsLOND.II.III}_i = \beta_i \left( \sum_{j \in I^\text{gsLOND.II.III}_i} R^\text{gsLOND.II.III}_j + 1 \right) \]

denote the nominal significance levels to calculate the group-sequential boundaries for the interim and the final test for \( H_i \), respectively. \( J^\text{gsLOND.II.III}_i \) and \( J^\text{gsLOND.II.III}_i \) again are index sets including all hypotheses with a final test decision at the time of interim or final analysis for \( H_i \). The index sets \( J^\text{gsLOND.II.III}_i \) and \( J^\text{gsLOND.II.III}_i \) may include all \( j \) of the set \( \{1, 2, \ldots, i - 1, i + 1, i + 2, \ldots\} \) where for \( H_j \) already group-sequential tests have been performed (allowing now also for \( j > i \) as well). If there were no additional rejections of other hypotheses between the interim and final analysis for \( i \), then \( \alpha^\text{gsLOND.II.III}_i = \alpha^\text{gsLOND}_i \), but the group-sequential boundaries may differ based on the spending function used (for more details see toy example). If the nominal significance level is increased, then the group-sequential boundaries are calculated using the initially fixed spending function as described for gsLOND.

Note that the gsLOND.III gives no formal online FDR control as it does not fulfil the condition that the local significance level \( \alpha \) for \( H_i \) is a function of the decisions for \( H_1, \ldots, H_{i-1} \) only, whereby in the original LOND procedure a pre-fixed order has to be used. But by not exhausting the nominal level alpha when updating the group-sequential boundaries, this might compensate for in principle violating the predefined order in LOND.

2.3.4 gsLOND.II.III

The two modifications gsLOND.II and gsLOND.III may also be combined, for example, if \( H_{i-1} \) and \( H_{i+1} \) both were rejected between the interim and the final analysis of \( H_i \). Thus, for the final analysis of \( H_i \) first the group-sequential boundaries are updated to exhaust the local nominal level (gsLOND.II) due to rejection of \( H_{i-1} \) only after the interim analysis of \( H_i \) and second the updated number of rejections is applied for the calculation of the level (gsLOND.III). We will illustrate this in the toy example.

2.4 Toy example

To illustrate the group-sequential LOND procedures we present a toy example and derive the appropriate critical boundaries: The maximum number of hypotheses to be tested is \( N = 3 \) and the significance level \( \alpha = 0.05 \) (one-sided test) is distributed equally among the three hypotheses, that is, \( \beta_i = 0.0167, i = 1, 2, 3 \) (note that \( \alpha = 0.05 \) in the toy example, whereas in the simulation studies below, \( \alpha = 0.025 \)). For the toy example, we limit the maximum number of hypotheses at the beginning of the trial with \( N = 3 \). This means that we do not fix the actual number of hypotheses which will be eventually tested in the platform trial, but only that not more than \( N \) can be tested. By fixing the maximum with \( N \), we do not give the option of including additional hypotheses later on. This constraint simplifies the illustration for the computation of the critical boundaries. In the toy example the numbers of hypotheses actually being tested and the maximum number are the same (\( N = K \)). An ordering of the interim and final analyses is considered as shown in Figure 1(b). For this scenario, Table 1 shows the critical boundaries for the fixed sample and the group-sequential design for the three hypotheses for the LOND and the gsLOND procedures for the Pocock18 (PO) and O’Brien-Fleming17 (OBF) type \( \alpha \)-spending design in case of 0, 1, or 2 previous rejections. Only for this special case with equal \( \beta_i \), the boundaries for \( H_i \) depend on \( i \) through the number of previous rejections of hypotheses \( H_j \) (with \( j < i \)).

Table 2 shows the group-sequential critical boundaries for gsLOND, gsLOND.II, gsLOND.III, and gsLOND.II.III for \( H_2 \) for all possible outcomes of the tests for \( H_1 \) and \( H_3 \) in the interim or in the final analyses (PO design). Similarly, Table 3

| \( \sum R \) | LOND: \( \alpha \) | \( \alpha \) | \( \sum R \) | LOND: \( \alpha \) | \( \alpha \) |
|---|---|---|---|---|---|
| 0 | 0.0167 | 0.0103 | 0.0089 | 0.0007 | 0.0164 |
| 1 | 0.0334 | 0.0207 | 0.0190 | 0.0026 | 0.0325 |
| 2 | 0.05 | 0.0310 | 0.0297 | 0.0056 | 0.0482 |
shows the boundaries for $H_3$. If the relevant number of rejections does not change between the interim and the final analysis, the gsLOND procedure for the final analysis has the same boundaries as the gsLOND.II and gsLOND.III procedures. With relevant number it is meant, that for updating the level of $H_2$, only rejections of $H_1$ will matter for the procedures gsLOND and gsLOND.II, whereas for gsLOND.III and gsLOND.II.III also rejections of $H_3$ will matter. The group-sequential critical boundaries for gsLOND and gsLOND.III are calculated with the R-package20 ldbounds,21 for the other procedures rpact22 is applied (R-code including a description of the program and examples for the gsLOND, gsLOND.II, and gsLOND.III procedures is available as Supplemental Material).

gsLOND.II
In Table 3, the gsLOND.II procedure can be applied for the calculation of the group-sequential boundaries for $H_3$ if $H_2$ has been rejected in the final but not in the interim analysis: E.g., in line 3, it is assumed that $H_1$ has been retained and $H_3$ has been rejected in the interim analysis and therefore, $\alpha_3^{(1)} = \alpha_3^{(2)} = \beta_3(1 + 1) = 0.0334$, $\alpha_3^{(1)} = 0.02076$, and $\alpha_3^{(2)} = 0.0190$. If, however, $H_2$ is retained in the interim analysis and then rejected in the final analysis, this additional rejection cannot be applied for the calculation of the level for the interim analysis of $H_3$ of gsLOND due to a time overlap (final analysis of $H_2$ after interim analysis of $H_3$) and thus $\alpha_3^{(2)}$ is only set to 0.0167 instead of 0.0334. With the gsLOND.II method this “loss” is translated to the final analysis of $H_3$ by exhausting the local level and increasing $\alpha_3^{(2)}$ to 0.0279 instead of 0.0190.

gsLOND.III
In line 4 in Table 2, it is assumed that $H_1$ is rejected in the interim analysis and thus for the interim analysis of $H_2$, $\alpha_2^{(1)} = \beta_2(1 + 1) = 0.334$. The corresponding group-sequential boundary is $\alpha_2^{(1)} = 0.0207$. If $H_3$ is not rejected, $\alpha_2^{(2)} = \alpha_2^{(1)}$ and $\alpha_2^{(2)} = 0.0190$. If, however, $H_3$ is rejected in the interim analysis (Table 2, line 5) and this decision is made before the final analysis of $H_2$, for the gsLOND.III, $\alpha_2^{(2)} = \beta_2(2 + 1) = 0.05$ and thus the group-sequential level can be increased to $\alpha_2^{(2)} = 0.0279$.
3 Results

3.1 Simulation methods

Simulation studies were performed to compare the LOND procedures for different scenarios for fixed and group-sequential designs with regard to average power, defined as the proportion of rejected alternatives among all alternatives, actual FDR and the average percentage amount of saved sample size of the group-sequential design compared to the fixed sample design. The following procedures are considered:

- Fixed sample design: LOND procedure.
- Group-sequential design: gsLOND, gsLOND.II, gsLOND.III, level-α test, Bonferroni.

For the level-α test, no adjustment is performed for multiple hypothesis testing, only for the interim analyses. Thus for each $H_i$, $\alpha_i = \alpha$ and group-sequential critical boundaries are derived according to the OBF or the PO design, respectively. In contrast, for the group-sequential Bonferroni procedure the significance level $\alpha_i$ for each hypothesis is adjusted according to the Bonferroni method and used as nominal level alpha for the group-sequential plan. Even though the total number of hypotheses $K$ is unknown, in the simulations, we set $\alpha_i = \alpha/K$ (best case scenario).

In each scenario, we compared either $K = 10$ or $K = 100$ treatment groups with one single control group. As shown in Figure 1, recruitment of the control group starts at the beginning of the platform trial and patients are recruited during the whole observation period. Treatments start at different times and run parallel to the control group, depending on the individual start. We interpret the index of the control patients as the measure for time and assume that control patients enter the trial consecutively. We further assume that the distribution of patients to control and treatments running in parallel is equal for all arms, that is, if two treatment arms 1 and 2 are running, the distribution is 1:1:1 for control, treatment 1, and treatment 2. The sample size for each treatment is $n = 50$ and in case of an interim analysis it is equally divided between two stages of the treatment group, $n^{(1)} = n^{(2)} = 25$ (which can easily be extended to varying per-hypothesis sample sizes $n_i$, $n_i^{(1)}$, $n_i^{(2)}$). For the first part of simulations we assume that after every $n_0 = 20$ control patients, a new treatment arm starts. In case of CC controls only, for each hypothesis test, the observations for the control and the corresponding treatment run parallel and both groups have equal sample sizes. In the case of NCC+CC controls, the control group additionally contains all control observations collected so far and thus has a potentially larger first stage sample size. The observations for a treatment group are the same in the CC and in the NCC+CC scenario.

Normally distributed observations with $N(\Delta, 1)$ were simulated with $\Delta = 0$ for the control group and treatment data from the set of true null hypotheses and $\Delta = 0.6$ for treatments of true alternative hypotheses with a proportion of true null hypotheses of $\pi_0$. For each treatment we considered the one-sided null hypothesis $H_{0i}: \mu_i \leq 0$ versus $H_{1i}: \mu_i > 0$ for the mean of the observations. Two-sample t-tests were performed and corresponding one-sided $p$-values were derived. The significance level for the online FDR procedures was set to $\alpha = 0.025$ and stopping for futility was performed in the interim analysis if $p_{ij}^{(1)} > \alpha^2 = 0.5$. The values of $\beta_j$ for the LOND procedure were first calculated as proposed by, for example, Javanmard and Montanari$^7$:

$$\beta_j = Ca^{-\log(\max(j, 2))}/j^{\exp(\sqrt{\log(j)})}$$

with $C = 0.07720838$ resulting in a decreasing sequence of numbers with, for example, $\beta_1 = 0.00134$, $\beta_2 = 0.00029$, $\beta_3 = 0.00025$. Setting an a-priori upper bound $N$ on the total number of hypotheses as suggested by Robertson and Wason$^10$ such that $\sum_{i=1}^N \beta_i = \alpha$, leads to larger values of $\beta_i$ and consequently larger significance levels, for example, for $N = 1000$, $\beta_1 = 0.00446$, $\beta_2 = 0.00099$, and $\beta_3 = 0.00083$.

Order of alternatives

In the simulations we considered three different orders of alternative hypotheses (see Supplemental Figure 1).

- Random order: In the simulations for each hypothesis the probability for a null hypothesis is $\pi_0$. Thus, for individual simulation steps, the number of alternatives may differ.
- Alternatives first: The simulated platform trial starts with $m_1 = K - \pi_0 K$ alternatives followed by $K - m_1$ true null hypotheses.
• Alternatives last: The simulated platform trial starts with $K - m_1$ true null hypotheses followed by $m_1$ alternative hypotheses.

If not specified otherwise, group-sequential critical boundaries were derived according to the OBF type $\alpha$-spending design (R-packages\textsuperscript{20} ldbounds,\textsuperscript{21} or rpact\textsuperscript{22}). Note that in the simulations we used the same weights for the spending function for NCC+CC controls as for CC controls and thus the same group-sequential boundaries. Thus the procedure for NCC+CC controls becomes conservative as it does not fully use the actual correlation. The calculation of the efficacy critical boundaries does not incorporate the futility stopping, thus the futility threshold is non-binding.\textsuperscript{23} All simulations were conducted using the R program,\textsuperscript{20} for each scenario at least 5000 simulation runs were performed. Note that in all considered simulation scenarios, the FDR is maintained at one-sided level $\alpha = 0.025$ (exception: level-\textalpha{} test). Details can be found in the Supplemental Material.

### 3.2 Comparison of LOND methods

#### 3.2.1 CC controls

Overall power values as a function of $\pi_0$ for a simulation trial with CC control data for the level-\textalpha{} test, the Bonferroni procedure and the four LOND procedures for unlimited number of hypotheses $N$ are shown in Figure 2 (rows 1 and 2). Additionally, gsLOND for $N \in \{10, 100, 1000\}$ (for $K = 100$, only $N \in \{100, 1000\}$) was considered. No differences in the power curves can be observed between the four LOND procedures with unlimited $N$. The reason is that for the gsLOND procedures the OBF group-sequential boundaries for interim analyses are rather low, early rejections thus only occur for large effects. However, for these cases, also the fixed sample LOND procedure leads to a rejection and the advantage of the gsLOND is only comprehensible in the amount of saved sample size. Below we will assume a simulation scenario, where additional treatments are included in case of early stopping in the interim analysis and we will show that the number of rejected alternatives increases for gsLOND procedures compared to the fixed sample design.

Setting $N = 1000$, the power of gsLOND increases by more than 10 percentage points. It further increases for $N = 100$ or $N = 10$ (for $K = 10$ only), but the improvement is less pronounced. Note that again nearly equal power values are observed for LOND, gsLOND, gsLOND.II, and gsLOND.III with $N \in \{10, 100, 1000\}$ (data not shown).

The power of the LOND procedures depends on the order of alternatives: For “alternatives first,” the power increases with the proportion of null hypotheses $\pi_0$, the first alternative has the largest level and thus the largest power, additional alternatives only decrease the power on average. For “alternatives last” the situation is reversed, LOND procedures have decreasing power values for increasing $\pi_0$, because for “late” alternatives with large $i$ the values of $\beta_i$ are only low and hardly any rejections have been performed before. Here the Bonferroni procedure (best case scenario) has even higher power than the comparable gsLOND procedure with $N = K$. The level-\textalpha{} and Bonferroni procedure are not influenced by the order of alternatives as the significance levels are constant.

#### 3.2.2 NCC+CC controls

The same simulations as described above were repeated for NCC+CC controls (Figure 2, rows 3 and 4). As for the scenarios with CC controls, no differences in the power curves can be observed between the four LOND procedures and the power values very much depend on the value of $N$. For the scenario $K = 10$ and “alternatives first,” the power is not monotonous in $\pi_0$. The reason is the trade off between decreasing level and increasing sample size for the NCC+CC control data.

The power for the level-\textalpha{} and the Bonferroni procedure now also depends on the distribution of alternatives due to the composition of the control group and is therefore not constant: for example, comparing scenarios “alternatives first” and “alternatives last” for large $\pi_0$, the number of observations in the NCC+CC control group for the “early” true alternative(s) is much lower and thus also power is decreased.

The advantage of the group-sequential LOND procedures in comparison to the fixed sample LOND can be seen in the average percentage amount of saved sample size in Figure 3 (NCC+CC controls). The average percentage saved sample size is defined as the proportion of actually saved treatment observations in stage 2 due to early stopping for efficacy or futility in the interim analysis. For $K = 10$, the maximum total number of treatment observations in stage 2 is given by $Kn_{\text{II}} = 250$ and for $K = 100$ by 2500. For small $\pi_0$, the percentage saved sample size is low as only hypotheses with a very small $p$-value are rejected in the interim analysis, but it increases for decreasing $N$. With increasing $\pi_0$, more sample size can be saved due to early stopping for futility in the interim analyses. For $\pi_0 \approx 1$, on average 25% can be saved for $\alpha^* = 0.5$. Simulated FDR values and % saved sample size for CC controls can be found in the Supplemental Material (Figures 2 to 6).
Figure 2. Power for scenarios with CC (rows 1 and 2) and NCC+CC controls (rows 3 and 4) as a function of $\pi_0$ for the level-$\alpha$ and the Bonferroni procedure (with $N = K$) and the four LOND procedures. OBF design, $N = \{10, 100, 1000, \infty\}$, $\Delta = 0.6$, $\alpha = 0.025$, $\sigma^2 = 0.5$. The four LOND procedures can hardly be distinguished as the power values are very similar. Thus, for $N < \infty$, only the power for gsLOND is depicted.
3.2.3 Direct comparison of concurrent and non-concurrent controls

Figure 4 compares the NCC+CC controls versus CC controls for LOND and gsLOND procedures with upper bound \( N = 100 \). As expected, depending on the distribution of the alternatives, the inclusion of all controls recruited so far increases the power values, particularly if true alternatives arise at the end of a platform trial and the sample size of the control group is large. Note again that the power values of gsLOND and LOND are equal. The values of the actual FDR for the CC and the NCC+CC control data are comparable (see Figures 7 and 8 in the Supplemental Material).

3.3 Distribution of significance level

In the previous simulations the values of \( \beta_i \) were calculated according to equation (4) and possibly adjusted by the upper bound \( N \) and thus descending with increasing number of hypotheses (“descending \( \beta_i \)”). We additionally consider that significance level \( \alpha \) is equally distributed among all (potential) hypotheses (“equal \( \beta_i \)”), that is, \( N \) has to be prefixed and \( \beta_i = \alpha / N \). In Figure 5, the impact of “descending \( \beta_i \)” and “equal \( \beta_i \)” on the power of the group-sequential LOND is investigated as a function of \( N \) (for \( K = 100, \pi_0 = 0.5, NCC+CC \) controls, OBF design; results for CC controls can be found in Figure 9 in the Supplemental Material). The results of the procedures depend strongly on the value of \( N \): In case of random order, only for \( N \) very close to \( K \) “equal \( \beta_i \)” is superior, in case of alternatives last, “equal \( \beta_i \)” is superior for \( N < 500 \). If, for example, \( N = 1000 \) and \( K = 100 \), the significance level for “equal \( \beta_i \)” is \( \beta_i = 0.000025 \) for each hypothesis. Until \( H_{138} \), the nominal level of the “descending \( \beta_i \)” method and thus the power is higher, only as of \( \beta_{138} \) the descending level is lower. However, for \( N = 100 \), as of \( \beta_{14} \) the “equal \( \beta_i \)” level is larger. The Bonferroni procedure, adjusted by the true number of hypotheses \( K \) has a lower power for most scenarios, only for large \( N \) it has similar power values as “equal \( \beta_i \)”.

3.4 OBF versus PO design

The OBF design only distributes a small amount of the significance level to the interim analysis and retains a large amount for the final analysis, whereas the PO type \( \alpha \)-spending design more or less distributes the level equally between interim and
final analysis. Figure 6 shows the power values for the two designs as a function of the effect size for \( \pi_0 = \{0, 0.5, 0.9\} \) and \( K = 10 \), NCC+CC controls (for CC controls see Figure 8 in the Supplemental Material). In terms of power, the OBF design of the gsLOND procedure is always superior compared to the PO design when using the same sample sizes in both designs. This is a well-known feature in group-sequential designs. To achieve similar power, Jennison and Turnbull\(^24\) have shown that when testing a single hypothesis, a larger maximum sample size would be needed for PO designs compared to fixed sample or OBF designs. Within the PO scenarios, the gsLOND procedure has slightly less power than gsLOND.II or seq.LOND.III (see Figure 11 in the Supplemental Material). For larger \( K \) the difference becomes negligible. However, due to spending more alpha at interim, the comparison of the saved sample size reveals that the amount of saved

---

**Figure 4.** Power for CC versus all (NCC+CC) controls as a function of \( \pi_0 \) for LOND and gsLOND (results for gsLOND.II and gsLOND.III are not depicted due to nearly identical power values); OBF design, \( N = 100 \), \( \Delta = 0.6 \), \( \alpha = 0.025 \), \( \alpha_F = 0.5 \).

**Figure 5.** Power of gsLOND (NCC+CC controls) for two distributions of the significance level as a function of \( N \) and Bonferroni for \( N = K \). OBF design, \( \Delta = 0.6 \), \( \pi_0 = 0.5 \), \( K = 100 \), \( \alpha = 0.025 \), \( \alpha_F = 0.5 \).
sample size of the PO design is always higher than for the OBF design, for low $\pi_0$, this difference in some scenarios is more than doubled.

### 3.5 Inclusion of additional treatments for stopping in the interim analysis

Now we assume that the budget of the whole platform trial, that is, the total number of observations, is fixed with budget $B = \sum_{i=1}^{K_0} n_i + C$ where $K_0$ denotes the number of initially planned trials and $C$ the number of pre-planned controls. If a treatment arm is dropped early due to efficacy or futility and sample size may be saved, an additional treatment arm can be included immediately if the total budget $B$ is not yet exhausted. The ordering of the hypotheses is based on the order of entrance into the platform. For the effect size $\Delta$ of the alternatives and the value of $\pi_0$ several assumptions are made. First, it is assumed that $\pi_0$ and the effect size for the alternatives remain constant through the trial, also for added treatments (scenario 1). Second, distributed effect sizes are assumed with $\Delta = \{0.4, 0.8, 1.2\}$ randomly selected (scenario 2). In scenario 3, $\pi_0$ decreases by 1/80 for each new treatment and therefor alternative hypotheses become more likely, whereas in scenario 4 $\pi_0$ remains constant, but the value of the effect size $\Delta$ increases to $\Delta = 1$ for the additional alternative hypothesis.

In the simulations, we compared the LOND with the gsLOND procedure regarding the total number of rejected alternatives with $\alpha = 0.05$, $\alpha^F = 0.5$, $K = 10$, $N = 100$, NCC+CC controls and OBF design. Results are shown in Figure 7. In scenario 1, for small effect size $\Delta = 0.6$, the procedures are rather similar, only for larger $\Delta = 0.8$ and smaller $\pi_0$ the gsLOND procedures are superior compared to the LOND procedures. This effect is even more pronounced for the PO design (see Supplemental Material, Figure 13). For Scenarios 2 to 4 gsLOND is superior to LOND, only for Scenario 4 with low $\Delta = 0.6$ the LOND procedure rejects a larger number of alternatives if $\pi_0$ is large.

### 4 Discussion and conclusions

In this manuscript, the LOND procedure to control the online FDR in platform trials was examined. We showed how group-sequential methods have to be modified to allow for interim analyses with the option to stop for efficacy or futility. For each treatment arm we allow for a single interim analysis, but the methods could be extended allowing for more than one interim analysis per treatment. Extensive simulation studies were performed and we observed that the proposed gsLOND, gsLOND.II, gsLOND.III and LOND with fixed sample sizes have nearly identical power values, however, a large amount of the total sample size may be saved with group-sequential approaches either due to early stopping for futility or for efficacy in the interim analyses. In comparison with a group-sequential Bonferroni approach, however, depending on the value of $N$, the online FDR procedures are more powerful in many scenarios and save more sample size in most scenarios.

Setting an a-priori upper bound on the number $N$ of hypotheses has a substantial impact on the power of the LOND procedures and is thus recommended. In the simulations, setting a large bound of $N = 1000$ for an actual number of,
for example, $K = 10$ hypotheses increases the power strikingly compared to a procedure with unlimited $N$; if the bound is further decreased, the power increases even more.

In the simulation studies, we found no notable differences between the procedures gsLOND, gsLOND.II, and gsLOND.III. However, there might be scenarios where gsLOND.II and gsLOND.III have potentially higher number of rejections and thus higher power values than gsLOND (for a overview, see Table 5 in the Supplemental Material). For simulation studies to assess the operating characteristics of a platform trial in the planning phase we would suggest to use the gsLOND procedure. For the gsLOND.II procedure, the update of the significance level $\alpha_i$ is computationally more exhaustive and as we have not seen a substantial difference in power in the simulations this further technicalities might not be needed for simulations. For the analysis, we would propose the gsLOND.II procedure as it can result in further rejections compared to gsLOND. As noted before, gsLOND, gsLOND.II, and gsLOND.III are considered as heuristic procedures. Assuming independence between the test statistics, we give a proof sketch that gsLOND and gsLOND.II control the online FDR (for the updated LOND procedure, according to the proof of Zrnic et al.\textsuperscript{11} for the fixed sample design, see Supplemental Material, Section 5). Note that for $\beta_{DEP_i} = \beta_i / (\sum_{j=1}^1 1/j)$ instead of $\beta_i$ in the LOND procedure, LOND also controls the FDR for general dependency of $p$-values\textsuperscript{10} (according to the Benjamini-Yekutieli procedure\textsuperscript{25}) and thus also for the group-sequential design. However, we did not consider this modification in our simulations due to its conservative performance.

For all simulation scenarios, we considered positively dependent test statistics and we found no substantial inflation of the FDR level for gsLOND, gsLOND.II, gsLOND.III, and, as expected, for LOND with fixed sample design, also under the global null hypothesis for $\pi_0 = 1$ (see Supplemental Material). A key aspect of all LOND procedures is that the hypotheses have to be ordered independently of the data to ensure a proper control of the FDR when updating the testing procedures due to rejections. In platform trials a natural ordering could be based on the order of entrance into the platform. However, as shown in the paper this ordering might be violated when we allow for unequal sample sizes and interim analyses.

Simulations were performed for both using CC controls only and for NCC+CC controls. Also when including NCC control data, we found no inflation of the FDR level. As expected the inclusion of all so far observed controls leads to higher power values in the simulation studies, particularly for “late” hypotheses in platform trials running for a long time period (“alternatives last”). Nevertheless, the decision for or against using NCC controls must be reached for each platform trial individually, as bias may be introduced in case of time trends and the FDR control of the procedure may be negatively affected by the inclusion of NCC controls. Time trends may be caused by changes in study population over time, for example, due to a change in the standard of care.\textsuperscript{14,15} Platform trials are a new concept for clinical study design with potentially exploratory and confirmatory aims. Since the start of the COVID-19 pandemic, the popularity of platform trials seemed to increase.\textsuperscript{26} Examples are the REMAP-CAP trial\textsuperscript{27} which started in 2016 for

Figure 7. Comparison of numbers of rejected alternatives for platform trials with a fixed budget (NCC+CC controls) as a function of $\pi_0$ for LOND and gsLOND. If a treatment arm is stopped early, an additional treatment is included. Scenario 1: $\pi_0$ and effect size remain constant, $\Delta = 0.6$ and 0.8. Scenario 2: distributed effect sizes of alternatives, $\Delta = \{0.4, 0.8, 1.2\}$. Scenario 3: $\pi_0$ decreases by $1/80$ for each new treatment. Scenario 4: $\Delta$ increases to $\Delta = 1$ for additional alternative hypotheses. $K_0 = 10$, $N = 100$, $\alpha = 0.025$, $\alpha^2 = 0.5$, OBF design.
community-acquired pneumonia and was later adapted for COVID-19, or the RECOVERY trial starting in March 2020.\textsuperscript{28} The RECOVERY trial is a randomized, controlled platform trial which investigates the effects of several treatments in patients admitted to hospital with COVID-19. Up to approximately 200 hospitals are involved with more than 45,000 patients (https://www.recoverytrial.net/). For more details see the Supplemental Material, where results of a detailed (hypothetical) reanalysis of two trial endpoints with the proposed gsLOND procedure are reported. Another prominent, prepandemic platform trial is the Stampede trial\textsuperscript{29} for therapies of prostate cancer, which started in 2005 (see, e.g. Meyer et al.\textsuperscript{1} for an overview of other pre-pandemic platform trials).

Currently, there exists no consensus and many open issues on the type of multiplicity control\textsuperscript{3,30,14} in platform trials. On the one hand, a rather strict adjustment is postulated by control of the FWER, defined as the probability of at least one Type I error, to prevent false positive decisions. On the other hand, no adjustment for multiplicity is recommended as platform trials are considered as a collection of independent aims. The control of the (online) FDR reflects a compromise between no adjustment and the conservative FWER adjustment.\textsuperscript{3,12} The FDR is equivalent to the FWER in case all null hypotheses are true, however, it is less conservative for a positive number of false null hypotheses as false rejections are allowed as long as its expected proportion among all rejections is maintained at the pre-specified level. Thus the power of the platform trial can be increased if FDR instead of FWER control is applied. Note, however, that the advantage of the online FDR procedure also relies on the timing in the assessment of outcomes. In our simulation no delay was assumed. For delayed outcomes other group-sequential testing procedures have been suggested.\textsuperscript{31}

The $\alpha$-spending functions allow to extend the proposed group-sequential approach for a platform trial controlling the online FDR to more than two stages for a treatment. Depending on the stopping time the amount of saved sample size might increase. On the other hand a price has to be paid for the additional looks, the group-sequential boundaries decrease for a larger number of interim analyses and the power might decrease if the maximum sample sizes per treatment are not increased accordingly depending on the spending function and assumed effect sizes. Another complexity with more stages is that the pre-specified order of hypothesis may not relate to the order of testing when different number and timings for the interim analyses are chosen for the various treatments in the platform trial. Depending on the setting the selected testing procedures will become more conservative. To optimize $\alpha$-spending functions for group-sequential designs with more than one interim analysis per treatment will be subject of future research.

Part of future work is also how to use the already accumulated data in an on-going platform trial to specify design aspects of new treatment arms. For example, to reassess power and sample size based on the nominal level actually available at the start of a new arm. Furthermore, conditional power arguments may be used for sample size reassessment at interim analyses.\textsuperscript{32}

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: EU-PEARL (EU Patient-cEntric clinicAl tRial pLatforms) project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking (JU) under grant agreement No 853966. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation programme and EFPIA and Children’s Tumor Foundation, Global Alliance for TB Drug Development non-profit organisation, Springworks Therapeutics Inc. This publication reflects the authors’ views. Neither IMI nor the European Union, EFPIA, or any Associated Partners are responsible for any use that may be made of the information contained herein.

**ORCID iD**

Sonja Zehetmayer https://orcid.org/0000-0001-6863-7997

**Supplemental material**

Supplementary material for this article is available online.

**References**

1. Meyer EL, Mesenbrink P, Dunger-Baldauf C et al. The evolution of master protocol clinical trial designs: a systematic literature review. *Clin Ther* 2020; 42: 1330–1360.
2. Berry SM, Connor JT and Lewis RJ. The platform trial: an efficient strategy for evaluating multiple treatments. *Jama* 2015; 313: 1619–1620.
3. Wason J and Robertson D. Controlling type one error rates in multi-arm clinical trials: a case for the false discovery rate. *Pharm Stat* 2021; 20: 109–116.
4. Bai X, Deng Q and Liu D. Multiplicity issues for platform trials with a shared control arm. *J Biopharm Stat* 2020; 30: 1077–1090.
5. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 1995; 57: 289–300.
6. Javanmard A and Montanari A. On online control of the false discovery rate. *arxiv preprint arxiv: 50206197*, 2015.
7. Javanmard A and Montanari A. Online rules for control of false discovery rate and false discovery exceedance. *Ann Stat* 2018; 46: 526–554.
8. Tian J and Ramdas D. Online control of the familywise error rate. *Stat Methods Med Res* 2021; 30: 976–993.
9. Ramdas A, Znic T, Wainwright M et al. Saffron: an adaptive algorithm for online control of the false discovery rate. *Proceedings of the 35th International Conference on Machine Learning* 2018; 80: 4286–4294.
10. Robertson D and Wason J. Online control of the false discovery rate in biomedical research. *arXiv preprint arxiv: 180907292v1*, 2018.
11. Znic T, Ramdas A and Jordan M. Asynchronous online testing of multiple hypotheses. *J Mach Learn Res* 2021; 22: 1–39.
12. Robertson D, Wason J, Koenig F et al. Online error control for platform trials. *Stat Med* 2022. To Appear.
13. Robertson D, Wildenhain J, Javanmard A et al. onlinefdr: an R package to control the false discovery rate for growing data repositories. *Bioinformatics* 2019; 35: 4196–4199.
14. Collignon O, Gartner C, Haidich A et al. Current statistical considerations and regulatory perspectives on the planning of confirmatory basket, umbrella, and platform trials. *Clin Pharm Ther* 2020; 107: 1059–1067.
15. Bofill Roig M, Krotka P, Burman CF et al. On model-based time trend adjustments in platform trials with non-concurrent controls. *BMC Med Res Methodol* (Accepted) *ArXiv preprint arXiv:211206574* 2022.
16. DeMets DL and Lan KKG. Interim analysis: the alpha spending approach. *Stat Med* 1994; 13: 1341–1352.
17. O’Brien PC and Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; 5: 549–556.
18. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; 64: 191–199.
19. Wassmer G and Brannath W. *Group sequential and confirmatory adaptive designs in clinical trials*. Cham, Switzerland: Springer International Publishing. 2016.
20. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. http://www.R-project.org.
21. Casper C and Perez OA. *ldbounds: Lan-DeMets Method for Group Sequential Boundaries*, 2018. https://CRAN.R-project.org/package=ldbounds. R package version 1.1.1.1. Based on FORTRAN program ld98.
22. Wassmer G and Pahlke F. *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*, 2021. https://CRAN.R-project.org/package=rpact. R package version 3.1.1.
23. Bretz F, Koenig F, Brannath W et al. Adaptive designs for confirmatory clinical trials. *Stat Med* 2009; 28: 1181–1217.
24. Jennison C and Turnbull BW. *Group sequential methods with applications to clinical trials*. Boca Raton: Chapman & Hall/CRC, 2000.
25. Benjamini Y and Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 2001; 29: 1165–1188.
26. Vanderbeek A, Bliss J, Yin Z et al. Implementation of platform trials in the COVID-19 pandemic: a rapid review. *Contemp Clin Trials* 2022; 112: 1–12.
27. Victor B, Talisa V, Yende S et al. Arguing for adaptive clinical trials in sepsis. *Front Immunol* 2018; 9: 1–8.
28. Group RC. Dexamethasone in hospitalized patients with COVID-19. *N Engl J Med* 2021; 384: 693–704.
29. James N, Sydes M, Clarke N et al. Systemic therapy for advancing or metastatic prostate cancer (stampede): a multi-arm, multistage randomized controlled trial. *BJU Int* 2008; 103: 464–469.
30. Bretz F and Koenig F. Commentary on Parker and Weir. *Clinical Trials* 2020; 17: 567–569.
31. Hampson LV and Jennison C. Group sequential tests for delayed responses (with discussion). *J R Stat Soc B* 2013; 75: 3–54.
32. Bauer P and Koenig F. The reassessment of trial perspectives from interim data’a critical view. *Stat Med* 2006; 25: 23–36.