The Annals of Statistics
2024, Vol. 52, No. 2, 817–841
https://doi.org/10.1214/24-AOS2370

THE ONLINE CLOSURE PRINCIPLE

BY LASSE FISCHER[1,a], MARTA BOFILL ROIG[2,c] AND WERNER BRANNATH[1,b]

[1]*Competence Center for Clinical Trials Bremen, University of Bremen,* [a]*fischer1@uni-bremen.de,* [b]*brannath@uni-bremen.de*
[2]*Center for Medical Data Science, Medical University of Vienna,* [c]*marta.bofillroig@meduniwien.ac.at*

The closure principle is fundamental in multiple testing and has been used to derive many efficient procedures with familywise error rate control. However, it is often unsuitable for modern research, which involves flexible multiple testing settings where not all hypotheses are known at the beginning of the evaluation. In this paper, we focus on online multiple testing where a possibly infinite sequence of hypotheses is tested over time. At each step, it must be decided on the current hypothesis without having any information about the hypotheses that have not been tested yet. Our main contribution is a general and stringent mathematical definition of online multiple testing and a new online closure principle, which ensures that the resulting closed procedure can be applied in the online setting. We prove that any familywise error rate controlling online procedure can be derived by this online closure principle and provide admissibility results. In addition, we demonstrate how shortcuts of these online closed procedures can be obtained under a suitable consonance property.

1. Introduction. The closure principle by Marcus, Peritz and Gabriel [22] is one of the most fundamental principles in multiple testing, especially when considering familywise error rate (FWER) control. It has been used to derive many popular and efficient multiple testing procedures commonly applied in current practice, for example, gatekeeping procedures [4] and graphical approaches [3]. Indeed, it can be shown that every FWER controlling procedure is also a closed procedure [35]. Furthermore, the closure principle can often be used to improve existing procedures [15]. In order to apply the closure principle, one needs to define the set of all nonempty intersection hypotheses, also called closure set. The idea is then to enforce coherence [10] by rejecting an individual hypothesis, if all intersection hypotheses containing this individual hypothesis are rejected, each at level α. However, many modern applications do not require that all individual hypotheses are known at the beginning of the evaluation. In such cases, we must decide on individual hypotheses without knowing the intersection hypotheses formed with hypotheses added later. This makes the use of the closure principle in the classical sense impossible. One can also think the other way around. Suppose there exists an intersection test for each intersection hypothesis in the full closure set. Which properties must these intersection tests have such that we can decide on a hypothesis H at a given time with the information that is available then? This is the main question we seek to answer in this paper.

In this paper, we focus on the online multiple testing setting, introduced by Foster and Stine [9]. In this setting, the hypotheses arrive sequentially over time and, at each step, it must be decided whether to reject the current hypothesis without having any information about the future ones. Large internet companies, for instance, face this problem when they perform A/B tests during their marketing research [20]. But also in genetics, thousands of tests are carried out in a sequential manner [24]. Javanmard and Montanari [18] even interpret

Received November 2022; revised December 2023.
MSC2020 subject classifications. 62L10.
Key words and phrases. Online multiple testing, closure principle, familywise error rate.

scientific research itself as an online multiple testing problem, since a stream of hypotheses is continuously tested [17].

Since online multiple testing was introduced in [9], a diverse range of online procedures has been proposed [18, 25, 26, 36–38]. Most of these procedures provide false discovery rate (FDR) control [18, 25, 26, 36, 38], where FDR is the expected proportion of true null hypotheses among the rejected hypotheses. Since FDR is less conservative than FWER [1], it is especially useful when testing a large number of hypotheses. This is also the reason why the literature on online multiple testing has been initially focused on FDR control. However, in classical applications, FWER is a very common error rate and there are also online problems where it is necessary to ensure that the probability of committing any type I error is below a certain level. For example, this may be the case in platform trials [28] and the sequential modification of machine learning algorithms [7]. The paper by Tian and Ramdas [37] is the only one fully focused on online control of the FWER so far. They have introduced online versions of popular multiple testing procedures such as the graphical approach by Bretz et al. [3]. However, their most promising method is the ADDIS-Spending which stands for adaptive discarding and combines two powerful concepts used in multiple testing. First, one adapts to the number of false hypotheses, as false hypotheses cannot lead to a type I error [34]. Second, one ignores ("discards") hypotheses with large p-values such that the remaining ones can be rejected with higher probability [39]. In addition, online FWER control was considered in [5] and [28]. Döhler et al. (2021) [5] introduced superuniformity reward (SURE) as an alternative to discarding, which is based on a priori information about the marginal CDF of null p-values, whereas Robertson et al. (2022) [28] describe how to apply online error control in the context of platform trials.

Our main contribution is a novel online closure principle, which ensures that the resulting closed procedure can be applied in the online setting. Note that Tian and Ramdas [37] have already provided an initial attempt for an online extension of the closure principle. However, they did not prove that the resulting closed procedure is indeed an online procedure and have therefore formulated this as an open problem. In Section 4, we show that their approach is actually a special case of a more general online closure principle.

The paper begins with a general and precise definition of online multiple testing, which to the best of our knowledge, has not been introduced in the literature yet (Section 2). Afterwards, we introduce the online closure principle including a so-called predictability condition under which a closed procedure is indeed an online procedure (Section 3). Moreover, we show that every FWER controlling online procedure can be obtained by this online closure principle and provide admissibility results for online closed procedures (Section 3.1). After that, we derive shortcuts of online closed procedures under consonance (Section 3.2). In Section 4, we transfer these general results to a more specific setting in which it is assumed that a p-value is obtained for each individual hypothesis. This is the usual setting considered in online multiple testing literature [9, 18]. We then use the online closure principle to derive new online procedures. Particularly, this gives a uniform improvement of the currently most promising online procedure with FWER control, the ADDIS-Spending under local dependence [37]. We exemplify the usage of the proposed procedure by applying them to simulated data (Section 5) and real data of a large-scale genetic study (Section 6.1) and with an ongoing platform trial (Section 6.2). The paper ends with a discussion in Section 7. The code for the simulations and the analyzes of the real data is provided in the Supplementary Material [8].

2. Online multiple testing. In the literature, *online multiple testing* is described as the setting, where an infinite stream of null hypotheses $\mathcal{H} = (H_i)_{i\in\mathbb{N}}$, indexed by entry order, is tested in a sequential manner. This means that at each step/time $i \in \mathbb{N}$ it needs to be decided whether H_i is rejected without access to any information about the future hypotheses and data [9, 18]. In this section, we define online multiple testing in a more mathematical manner so that it becomes clearer what a multiple testing procedure must satisfy to be termed online.

Let $(\Omega, \mathcal{A})$ be a measurable space and $\mathcal{P}$ some set of probability distributions on $(\Omega, \mathcal{A})$. Note that $(\Omega, \mathcal{A}, \mathcal{P})$ is to be understood in an abstract sense and it is not supposed to be completely known in advance. We assume that the data follows some unknown distribution $\mathbb{P} \in \mathcal{P}$. The hypotheses $(H_i)_{i\in\mathbb{N}}$ can be formally considered as subsets of $\mathcal{P}$ and by testing H_i, we want to examine whether $\mathbb{P} \in H_i$. Unless otherwise stated, equalities and inequalities involving random variables should be understood to hold almost surely for all $\mathbb{P} \in \mathcal{P}$. We further assume that a filtration $\mathbb{F} = (\mathcal{F}_i)_{i\in\mathbb{N}}$ (increasing sequence of σ-fields) is given, where $\mathcal{F}_i \subseteq \mathcal{A}$ defines the information that the test decision for H_i is allowed to depend on. For example, $\mathcal{F}_i$ can be the σ-field that is generated by all observations that are available at time i. However, we may not want to use all observations completely or add external randomization. With this, we can formally define an online multiple testing procedure as follows.

DEFINITION 2.1 (Online multiple testing procedure). An *online multiple testing procedure* (hereinafter referred to as online procedure for short) for $\mathcal{H} = (H_i)_{i\in\mathbb{N}}$ is a sequence of test decisions $\boldsymbol{d} = (d_i)_{i\in\mathbb{N}}$, where each d_i is a random variable with values in $\{0, 1\}$ that is measurable with respect to $\mathcal{F}_i$. If $d_i = 1$, we conclude that H_i is rejected and if $d_i = 0$, that H_i is accepted.

Therefore, in contrast to classical "offline" multiple testing, we have an infinite number of hypotheses to consider. Furthermore, each test decision d_i is only allowed to use some partial information $\mathcal{F}_i$ of the total information $\mathcal{A}$, whereby the partial information $\mathcal{F}_i$ is growing over time $i \in \mathbb{N}$. Note that this setting encompasses the classical setting as a special case, in which $\mathcal{F}_i = \mathcal{A}$ for all $i \in \mathbb{N}$ and the testing process is stopped after $m \in \mathbb{N}$ steps, for example, by choosing $H_i = \mathcal{P}$ and $d_i = 0$ for all $i > m$. In a similar way, *online batch testing* [40] can be embedded in our framework. Even though the theoretical results of this paper apply in this general setting, we focus on the strict online case of $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ...$ when deriving concrete online closed procedures.

We denote by $I_0^{\mathbb{P}} := \{i \in \mathbb{N} : \mathbb{P} \in H_i\}$ and $I_1^{\mathbb{P}} := \mathbb{N} \setminus I_0^{\mathbb{P}}$ the index sets of true and false null hypotheses, if $\mathbb{P}$ was the true distribution, respectively. Furthermore, for all $i \in \mathbb{N}$, we define $v_{\mathbb{P}}(i) := \sum_{j\le i, j\in I_0^{\mathbb{P}}} d_j$ as the number of falsely rejected hypotheses up to step $i \in \mathbb{N}$ and set $v_{\mathbb{P}} := \lim_{i\to\infty} v_{\mathbb{P}}(i)$. With this, we define the *familywise error rate* (FWER) at time $i \in \mathbb{N}$ and overall hypotheses as

$$\text{FWER}_{\mathbb{P}}(i) := \mathbb{P}\big(v_{\mathbb{P}}(i) > 0\big) \quad \text{and} \quad \text{FWER}_{\mathbb{P}} := \mathbb{P}(v_{\mathbb{P}} > 0) \quad (\mathbb{P} \in \mathcal{P}). \tag{1}$$

We aim for strong control of the FWER at each time $i \in \mathbb{N}$, which means that for some prespecified $\alpha \in (0, 1)$, we have $\text{FWER}_{\mathbb{P}}(i) \le \alpha$ for all $i \in \mathbb{N}$ and $\mathbb{P} \in \mathcal{P}$. Note that this is equivalent to requiring $\text{FWER}_{\mathbb{P}} \le \alpha$ for all $\mathbb{P} \in \mathcal{P}$, since $E_i^{\mathbb{P}} := \{v_{\mathbb{P}}(i) > 0\} \in \mathcal{F}_i$ is an increasing sequence $(E_1^{\mathbb{P}} \subseteq E_2^{\mathbb{P}} \subseteq \ldots)$ with $E_i^{\mathbb{P}} \subseteq \{v_{\mathbb{P}} > 0\}$ for all $i \in \mathbb{N}$. Therefore, we drop the index i in the following. In contrast to strong control, weak FWER control only requires that $\text{FWER}_{\mathbb{P}} \le \alpha$ for all distributions contained in the global null hypothesis $\mathbb{P} \in \bigcap_{i\in\mathbb{N}} H_i$. This is of limited use in practice. Hence, when we write control in the remainder of this paper, we always mean strong control.

3. Online closure principle. For a potentially infinite index set $I \subseteq \mathbb{N}$, we denote the corresponding intersection hypothesis and intersection test by $H_I = \bigcap_{i\in I} H_i$ and ϕ_I, respectively. Each ϕ_I is a random variable with values in $\{0, 1\}$ such that H_I is rejected by ϕ_I, if $\phi_I = 1$, and accepted, if $\phi_I = 0$. We say that ϕ_I, $I \subseteq \mathbb{N}$, is an online intersection test, if ϕ_I is measurable with respect to $\mathcal{F}_{\sup(I)}$, where $\mathcal{F}_\infty = \mathcal{A}$. Furthermore, ϕ_I is an α-level intersection test, if $\mathbb{P}(\phi_I = 1) \le \alpha$ for all $\mathbb{P} \in H_I$. For the online closure principle, we need an online α-level intersection test ϕ_I for each $I \subseteq \mathbb{N}$, where we always set $\phi_\varnothing = 0$. We will see that if the family $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ fulfils the following condition, the resulting closed testing procedure is indeed an online procedure.

DEFINITION 3.1 (Predictable family of online intersection tests). A family of online intersection tests $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$ is called *predictable*, if for all $i \in \mathbb{N}$ and $I \subseteq \{1, \dots, i\}$ holds that:

$$\phi_I = 1 \text{ implies } \phi_K = 1 \text{ for all } K = I \cup J \text{ with } J \subseteq \{j \in \mathbb{N} : j > i\}.$$

This predictability condition ensures that if a finite intersection hypothesis H_I, $I \subseteq \{1, \dots, i\}$ is rejected, it remains rejected when future hypotheses H_j, $j > i$, are added. For example, suppose $H_1 \cap H_3$ is rejected, then $H_1 \cap H_3 \cap H_4$ needs to be automatically rejected as well. However, $H_1 \cap H_2 \cap H_3$ does not need to be rejected, as H_2 is not a future hypothesis in that case. Now, we can formulate a closure principle for online multiple testing.

THEOREM 3.2 (Online closure principle). *Let $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$ be an arbitrary family of α-level intersection tests. Then the closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}} = (d_i^{\boldsymbol{\phi}})_{i \in \mathbb{N}}$ based on $\boldsymbol{\phi}$ defined by*

$$d_i^{\boldsymbol{\phi}} = \min\{\phi_I : I \subseteq \mathbb{N} \text{ with } i \in I\}$$

controls the FWER at level α in the strong sense. In addition, if each ϕ_I is an online intersection test and the family of online intersection tests $\boldsymbol{\phi}$ is predictable, then $\boldsymbol{d}^{\boldsymbol{\phi}}$ is an online procedure. We refer to such procedures as online closed procedures.

To prove Theorem 3.2, we first show the following lemma, which states that if the predictability condition is satisfied, only the current and previous hypotheses need to be considered at each step.

LEMMA 3.3. *If $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$ is predictable, then $d_i^{\boldsymbol{\phi}}$ defined in Theorem 3.2 satisfies $d_i^{\boldsymbol{\phi}} = \min\{\phi_I : I \subseteq \{1, \dots, i\} \text{ with } i \in I\}$ for all $i \in \mathbb{N}$.*

PROOF. Let $i \in \mathbb{N}$ and $K \subseteq \mathbb{N}$ with $i \in K$ be arbitrary. Note that K can be written as $K = I \cup J$, where $I = \{k \in K : k \leq i\}$ and $J = \{k \in K : k > i\}$. The predictability of $(\phi_I)_{I \subseteq \mathbb{N}}$ ensures that H_K is rejected by ϕ_K if H_I is rejected by ϕ_I. □

PROOF OF THEOREM 3.2. We first show FWER control. Let $\mathbb{P} \in \mathcal{P}$ be arbitrary. In order to reject any true null hypothesis, it needs to hold for the subset containing the indices of all true hypotheses $I_0^{\mathbb{P}}$ that $\phi_{I_0^{\mathbb{P}}} = 1$. Since $\phi_{I_0^{\mathbb{P}}}$ is an α-level intersection test, we have

$$\mathbb{P}(v_{\mathbb{P}} > 0) \leq \mathbb{P}(\phi_{I_0^{\mathbb{P}}} = 1) \leq \alpha.$$

To show the second assertion, we assume that each ϕ_I is an online intersection test and $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$ is predictable. Lemma 3.3 implies that $d_i^{\boldsymbol{\phi}} = \min\{\phi_I : I \subseteq \{1, \dots, i\} \text{ with } i \in I\}$, $i \in \mathbb{N}$. Since each ϕ_I with $I \subseteq \{1, \dots, i\}$ is measurable with respect to $\mathcal{F}_{\sup(I)} \subseteq \mathcal{F}_i$, $d_i^{\boldsymbol{\phi}}$ is measurable with respect to $\mathcal{F}_i$. □

Note that Lemma 3.3 does not hold in general when the predictability of $(\phi_I)_{I \subseteq \mathbb{N}}$ is violated. For example, let p_1 and p_2 be the p-values for H_1 and H_2 that are measurable with respect to $\mathcal{F}_1$ and $\mathcal{F}_2 \supset \mathcal{F}_1$, respectively. Suppose $(\phi_I)_{I \subseteq \mathbb{N}}$ is a family of online intersection tests such that $\phi_{\{1\}} = 1$ if $p_1 \leq \alpha$ and $\phi_{\{1,2\}} = 1$ if $p_1 \leq \frac{\alpha}{2}$ or $p_2 \leq \frac{\alpha}{2}$. Now assume that $\frac{\alpha}{2} < p_1 \leq \alpha$ and $p_2 > \frac{\alpha}{2}$. Thus, $\phi_{\{1\}} = 1$ but $\phi_{\{1,2\}} = 0$ and rejecting $H_{\{1\}}$ by $\phi_{\{1\}}$ would not be sufficient to reject H_1 by the closure principle. This implies that this closed procedure cannot be an online procedure and that the assumption of ϕ_I, $I \subseteq \mathbb{N}$, being an online intersection test, is insufficient to obtain an online closed procedure. In the following section, we show that predictability is necessary in general.

3.1. *Admissibility of online closed procedures.* There is a large body of literature discussing the admissibility of classical closed procedures [12, 21, 31, 35]. Sonnemann and Finner (1988) [35] showed that every admissible procedure with FWER control can be derived as a closed procedure and Romano et al. (2011) [31] proved that one can further restrict to consonant intersection tests. In this section, we derive similar admissibility results for the online closure principle (Theorem 3.2). We follow with our definition of admissibility the one in [12].

DEFINITION 3.4 (Admissibility of online procedures). A strong FWER controlling online procedure is called *admissible* when it cannot be uniformly improved by another online procedure with strong FWER control, where $\boldsymbol{d} = (d_i)_{i\in\mathbb{N}}$ is uniformly improved by $\tilde{\boldsymbol{d}} = (\tilde{d}_i)_{i\in\mathbb{N}}$, if $\tilde{d}_i \geq d_i$ for all $i \in \mathbb{N}$ and $\mathbb{P}(\tilde{d}_i > d_i) > 0$ for some $i \in \mathbb{N}$ and $\mathbb{P} \in \mathcal{P}$.

In the following theorem, we prove that any online procedure with strong FWER control can be obtained by the online closure principle (Theorem 3.2). This shows that the fundamentality of the classical closure principle can be transferred to the online setting. Furthermore, it implies that the predictability condition (Definition 3.1) is not too strict.

THEOREM 3.5. *Let $\boldsymbol{d} = (d_i)_{i\in\mathbb{N}}$ be an online procedure with strong FWER control. Then $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$, where $\phi_I = \max\{d_i : i \in I\}$, is a predictable family of online α-level intersection tests and $\boldsymbol{d}^{\boldsymbol{\phi}} = \boldsymbol{d}$. Thus, for any online procedure $\boldsymbol{d}$ with FWER control there exists an online closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$ that leads to the same decisions.*

PROOF. Since $\boldsymbol{d} = (d_i)_{i\in\mathbb{N}}$ is an online procedure, $\phi_I = \max\{d_i : i \in I\}$ is measurable with respect to $\mathcal{F}_{\sup(I)}$, and thus defines an online intersection test for all $I \subseteq \mathbb{N}$. Given the strong FWER control of $\boldsymbol{d}$, it follows that ϕ_I is an α-level intersection test. To see this, suppose $I \subseteq I_0^{\mathbb{P}}$ for some $\mathbb{P} \in \mathcal{P}$. Hence, $\mathbb{P}(\phi_I = 1) = \mathbb{P}(\max\{d_i : i \in I\} = 1) \leq \mathbb{P}(\max\{d_i : i \in I_0^{\mathbb{P}}\} = 1) \leq \alpha$. Furthermore, $\phi_I = 1$ implies $\phi_K = 1$ for all $I \subseteq K$, which ensures the predictability of $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$. It remains to show that $\boldsymbol{d}^{\boldsymbol{\phi}} = \boldsymbol{d}$. First, note that for all $i \in \mathbb{N}$, $d_i = 0$ implies $\phi_{\{i\}} = 0$, and thus $d_i^{\boldsymbol{\phi}} = 0$. Second, $d_i = 1$ implies $\phi_I = 1$ for all $I \subseteq \mathbb{N}$ with $i \in I$, and hence $d_i^{\boldsymbol{\phi}} = 1$. □

A family of intersection tests $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ is *consonant* [10], if for all $I \subseteq \mathbb{N}$,

$$\phi_I = 1 \text{ implies } \exists i \in I : \phi_J = 1 \forall J \subseteq I \text{ with } i \in J. \tag{2}$$

If a family of intersection tests is not consonant, it is called *dissonant*. Closed procedures based on consonant intersection tests have the desirable property that the rejection of an intersection hypothesis H_I implies that at least one individual hypothesis H_i with $i \in I$ is rejected. For example, suppose we are testing several treatment arms $T_1, T_2, \ldots$ against a common control (e.g., in a platform trial). Then the rejection of $H_1 \cap H_2$ would imply that at least T_1 or T_2 is efficient. However, if the procedure is dissonant, we might not be able to conclude which of the two treatments is efficient. Romano et al. (2011) [31] showed that every strong FWER controlling online procedure can be written as a closed procedure based on consonant intersection tests. Since the $\boldsymbol{\phi}$ defined in Theorem 3.5 is consonant, it immediately follows that this result also applies in the online setting.

COROLLARY 3.6. *For every online procedure $\boldsymbol{d}$ that controls the FWER strongly, there exists an online closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$ with $\boldsymbol{d}^{\boldsymbol{\phi}} = \boldsymbol{d}$ that is based on a consonant family of intersection tests $\boldsymbol{\phi}$.*

Corollary 3.6 only states that any online procedure can also be written as an online closed procedure based on consonant intersection tests. However, Romano et al. (2011) [31] have shown that "consonantizing" dissonant intersection tests $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ by choosing $\tilde{\phi}_I = \max\{d_i^{\boldsymbol{\phi}} : i \in I\}$, $I \subseteq \mathbb{N}$, often reveals weaknesses of $\boldsymbol{d}^{\boldsymbol{\phi}}$, which can be used to improve $\boldsymbol{d}^{\boldsymbol{\phi}}$ by improving $\tilde{\boldsymbol{\phi}}$. We also illustrate this in Section 4.1 by an example. Note that in the online case one needs to be careful with constructing improvements of intersection tests, as the predictability might get lost and the closed procedure is no longer an online procedure. For this reason, the following result may be helpful.

PROPOSITION 3.7. *Let $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ be a predictable family of online intersection tests. Furthermore, let $\psi_I = \phi_I$ for all finite index sets $I \subseteq \mathbb{N}$ and $\psi_I \geq \phi_I$ for all infinite $I \subseteq \mathbb{N}$. Then $\boldsymbol{\psi} = (\psi_I)_{I\subseteq\mathbb{N}}$ is a predictable family of online intersection tests as well.*

PROOF. Let $I \subseteq \{1, \ldots, i\}$ for some $i \in \mathbb{N}$ and $K = I \cup J$ with $J \subseteq \{j \in \mathbb{N} : j > i\}$. By the predictability of $\boldsymbol{\phi}$, we have $\psi_I = \phi_I \leq \phi_K \leq \psi_K$, which shows the predictability of $\boldsymbol{\psi}$. Furthermore, ψ_I is an online intersection test for each $I \subseteq \mathbb{N}$ by definition. □

The proposition shows that we cannot violate the predictability condition by improving intersection tests ϕ_I for infinite $I \subseteq \mathbb{N}$. Therefore, one approach to improve an existing online procedure $\boldsymbol{d}$ using the online closure principle would be to define $\phi_I = \max\{d_i : i \in I\}$ as in Theorem 3.5. Then, if possible, uniformly improve ϕ_I by ψ_I for infinite $I \subseteq \mathbb{N}$. After that, one might be able to also uniformly improve ϕ_I by ψ_I for finite $I \subseteq \mathbb{N}$ while retaining predictability of $(\psi_I)_{I\subseteq\mathbb{N}}$. Note that an improvement of some or all ϕ_I for infinite I leads to a relaxation of the predictability condition, and thereby could create possibilities to improve ϕ_I for finite I.

For example, for all $i \in \mathbb{N}$ let p_i be a p-value for H_i that is measurable with respect to $\mathcal{F}_i$. Then $\boldsymbol{d} = (d_i)_{i\in\mathbb{N}}$ with $d_i = \mathbb{1}\{p_i \leq \alpha_i\}$, where $\alpha_i > 0$ and $\sum_{i\in\mathbb{N}} \alpha_i = \alpha$, defines an online procedure with FWER control due to Bonferroni's inequality. Since $\sum_{i\in I} \alpha_i < \alpha$ for every $I \subset \mathbb{N}$, the intersection tests $\phi_I = \max\{d_i : i \in I\}$ can be improved. For arbitrary p-values and a finite $I \subseteq \mathbb{N}$ an improvement of ϕ_I leads to a violation of the predictability condition, however, due to Proposition 3.7 we can safely improve ϕ_I for infinite $I \subseteq \mathbb{N}$. For instance, define $\psi_I = \max\{\mathbb{1}\{p_i \leq \alpha_i^I\} : i \in I\}$, where $\alpha_i^I \geq \alpha_i$ and $\sum_{i\in I} \alpha_i^I = \alpha$, for all infinite $I \subseteq \mathbb{N}$. Now, we can also improve ϕ_I for finite I by $\psi_I = \max\{\mathbb{1}\{p_i \leq \alpha_i^I\} : i \in I\}$, where $\alpha_i^I = \inf\{\alpha_i^K : \exists J \subseteq \mathbb{N} \setminus \{1, \ldots, \max(I)\}, J \text{ infinite}, K = I \cup J\}$. Then $\boldsymbol{\psi} = (\psi_I)_{I\subseteq\mathbb{N}}$ is a predictable family of online α-level intersection tests with $\psi_I \geq \phi_I$ for all $I \subseteq \mathbb{N}$. In Sections 4.1 and 4.2, we derive specific improvements of this Alpha-Spending procedure [9].

Theorem 3.5 and Corollary 3.6 show that predictability and consonance of the family of intersection tests are necessary conditions for admissibility of an online procedure. Furthermore, Proposition 3.7 implies that if admissible intersection tests exist, admissible intersection tests for infinite index sets are also necessary for admissibility of an online procedure. Analogously to Definition 3.4, a single α-level test δ for a hypothesis $H \subseteq \mathcal{P}$ is admissible, if there exists no other α-level test $\tilde{\delta}$ for H with $\tilde{\delta} \geq \delta$ and $\mathbb{P}(\tilde{\delta} > \delta) > 0$ for some $\mathbb{P} \in \mathcal{P}$ [12, 21]. But, as in classical multiple testing, it is difficult (or even impossible) to find nontrivial sufficient conditions for admissibility. For example, it is not ensured that a closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$ is admissible, if ϕ_I is admissible for all $I \subseteq \mathbb{N}$ [2, 12]. As pointed out by Goeman et al. (2021) [12], showing admissibility for multiple testing procedures can be resolved by consideration of a monotone family of procedures $(\boldsymbol{d}^I)_{I\subseteq\mathbb{N}}$ that defines a multiple testing procedure for each subset of hypotheses, which we not do in this paper. However, we can prove a condition under which the event of rejecting any hypothesis cannot be enlarged without violating FWER control.

PROPOSITION 3.8. *Let $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ be a consonant family of intersection tests. If $\phi_{\mathbb{N}}$ is admissible, there does not exist a strong FWER controlling procedure $\mathbf{d} = (d_i)_{i\in\mathbb{N}}$ with $\max\{d_i : i \in \mathbb{N}\} \geq \max\{d_i^{\boldsymbol{\phi}} : i \in \mathbb{N}\}$ and $\mathbb{P}(\max\{d_i : i \in \mathbb{N}\} > \max\{d_i^{\boldsymbol{\phi}} : i \in \mathbb{N}\}) > 0$ for some $\mathbb{P} \in \mathcal{P}$.*

PROOF. Suppose there exists a $\mathbf{d}$ with the property stated in the theorem and define $E := \{\max\{d_i : i \in \mathbb{N}\} > \max\{d_i^{\boldsymbol{\phi}} : i \in \mathbb{N}\}\}$. Further, let $\tilde{\phi}_I = \max\{d_i : i \in I\}$, $I \subseteq \mathbb{N}$, be the intersection test defined in Theorem 3.5. Since $\boldsymbol{\phi}$ is consonant, we have $\phi_{\mathbb{N}} = \max\{d_i^{\boldsymbol{\phi}} : i \in \mathbb{N}\} \leq \max\{d_i : i \in \mathbb{N}\} = \tilde{\phi}_{\mathbb{N}}$ with a strict inequality if E happens. This contradicts the admissibility of $\phi_{\mathbb{N}}$. □

REMARK. Proposition 3.8 is inspired by a result shown in [31]. They considered the case, where the global test $\phi_{\mathbb{N}}$ maximizes the minimum probability of rejecting $H_{\mathbb{N}}$ over some set of alternative distributions $H_A \subseteq \mathcal{P} \setminus H_{\mathbb{N}}$ and showed that any consonant closed procedure based on this $\phi_{\mathbb{N}}$ also maximizes the minimum probability of rejecting any hypothesis over H_A. We think our result fits the online setting better, since one does usually not consider a fixed set of alternatives. Furthermore, the conditions of our proposition are also easy to meet in the online case as shown by a simple example in the following.

Proposition 3.8 is not restricted to online procedures, and thus it might seem that the sufficient condition is difficult to achieve in the online case. However, we already showed that making a predictable family of online intersection tests consonant (Corollary 3.6) or uniformly improving its infinite intersection tests (Proposition 3.7) will always lead to a predictable family of online intersection tests again. Therefore, it should not be too hard to meet these conditions. For example, suppose we have independent p-values $(p_i)_{i\in\mathbb{N}}$ that are uniformly distributed under the null hypothesis. It can then be shown that using the Online-Šidák procedure [37], which is defined by $d_i = \mathbb{1}\{p_i \leq \alpha_i\}$ with $\alpha_i = 1-(1-\alpha)^{\gamma_i}$ and $\sum_{i\in\mathbb{N}} \gamma_i = 1$, the probability of rejecting any hypothesis is exactly α under the global null hypothesis. Thus, if we define $\boldsymbol{\phi}$ as in Theorem 3.5 for the Online-Šidák, we obtain a consonant and predictable family of online intersection tests, where the test $\phi_{\mathbb{N}}$ has exact size α. Under mild assumptions, for example, that the collection of null sets is the same for all distributions $\mathbb{P} \in \mathcal{P}$ [12], this implies that $\phi_{\mathbb{N}}$ is admissible. Thus, in this setting the event of rejecting any hypothesis by Online-Šidák cannot be enlarged without violating FWER control.

3.2. *Shortcuts under consonance.* By Theorem 3.5, we can focus completely on online closed procedures when constructing new online procedures with FWER control, and by Corollary 3.6, we may even restrict to consonant intersection tests. Usually, at each step $i \in \mathbb{N}$ we need to consider up to 2^{i-1} intersection hypotheses H_I, $I \subseteq \mathbb{N}$, with $i \in I$. Since i tends to infinity, it is unrealizable to test all of these intersection hypotheses in practice. Even if the testing process stops at some point, it is computationally intensive and difficult to communicate. In the offline setting, the same problems occur when a large number of hypotheses is tested, which led to the establishment of shortcut procedures [15]. The objective of shortcut procedures is to find decisions for the individual hypotheses without testing every intersection hypothesis. In this way, the number of operations should be reduced to the number of individual hypotheses while the decisions coincide with those of a closed procedure. We now want to apply this approach to the online case. To formulate a shortcut, additional assumptions toward the family of intersection tests $\boldsymbol{\phi}$ are required. In this paper, we focus on shortcuts based on consonance (2).

When constructing consonance-based shortcuts in offline multiple testing, one would usually start with the global hypothesis [16], which is the intersection of all individual hypotheses. If the global hypothesis is rejected, there exists an index i satisfying the consonance

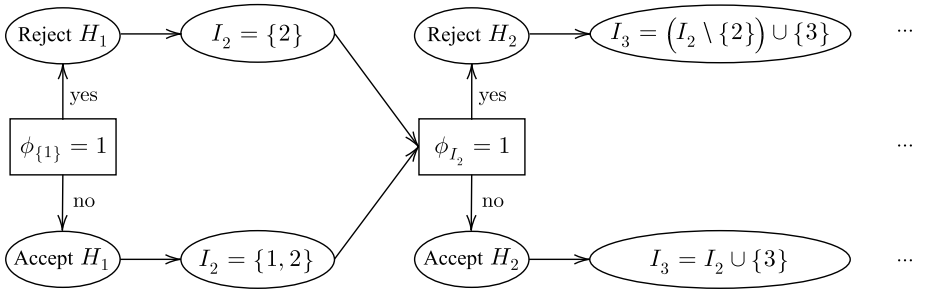

FIG. 1. *Shortcut of an online closed procedure based on consonance.*

property (2), which implies that the individual hypothesis H_i can be rejected by the closure principle. In the next step, the intersection of all hypotheses except H_i is considered and the testing step is repeated. This can be continued until the intersection of the remaining hypotheses cannot be rejected. In online multiple testing, this proceeding is not possible as the global hypothesis is not known at the beginning of the evaluation. However, the predictability condition makes it possible to formulate a shortcut anyway.

Assume $(\phi_I)_{I\subseteq\mathbb{N}}$ is a predictable family of online intersection tests with the consonance property and consider the intersection hypothesis $H_{\{1\}}$. When $H_{\{1\}}$ is rejected by its online intersection test $\phi_{\{1\}}$, the predictability of $(\phi_I)_{I\subseteq\mathbb{N}}$ ensures that H_1 is rejected by the online closure principle (Lemma 3.3). Now, set $I_2 = \{2\}$ if $\phi_{\{1\}} = 1$, and $I_2 = \{1, 2\}$ otherwise. Suppose $\phi_{I_2} = 1$. In case of $I_2 = \{2\}$, it holds $\phi_{\{1\}} = 1$ and due to the predictability of $(\phi_I)_{I\subseteq\mathbb{N}}$, it also holds $\phi_{\{1,2\}} = 1$. Hence, $\phi_{\{1,2\}} = 1$ and $\phi_{\{2\}} = 1$ implying that H_2 is rejected by the closure principle (Lemma 3.3). If $I_2 = \{1, 2\}$, the consonance property implies that $\phi_{\{2\}} = 1$ as well and again H_2 is rejected by the closure principle. This can be continued and a shortcut of the closed procedure is obtained, meaning that only one intersection hypothesis is tested for each individual hypothesis H_i, $i \in \mathbb{N}$. An illustration of the shortcut can be found in Figure 1. A formal description is given in the next theorem, whose proof can be found in Appendix C.

THEOREM 3.9. *Assume $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ is a predictable family of online α-level intersection tests with the consonance property. Let us recursively define $I_1 = \{1\}$ and $I_i = \{j \in \mathbb{N} : j < i, \phi_{I_j} = 0\} \cup \{i\}$ for all $i \geq 2$. Then the following procedures lead to the same decisions*:

1. *The online closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$.*
2. *The shortcut $\boldsymbol{d}^{\boldsymbol{\phi},s} = (d_i^{\boldsymbol{\phi},s})_{i\in\mathbb{N}}$, where $d_i^{\boldsymbol{\phi},s} = \phi_{I_i}$ for all $i \in \mathbb{N}$.*

Note that if an intersection hypothesis H_{I_i}, $i \in \mathbb{N}$, is rejected, it is also uniquely determined which individual hypothesis, namely H_i, is to be rejected, whereas in the classical case, the index satisfying the consonance property must first be determined. In the next section, we will use this fact to calculate individual significance levels for the shortcut. By means of such a result, we will obtain new and more powerful online procedures. For this purpose, we consider in the next section a more specific online multiple testing setting that is based on p-values.

4. Online closed testing with $\boldsymbol{\alpha}$-adjustments. Most existing online multiple testing procedures are defined based on p-values $(p_i)_{i\in\mathbb{N}}$ for the individual hypotheses $(H_i)_{i\in\mathbb{N}}$ [9, 18, 37]. Each p-value p_i can be considered as a random variable with values in $[0, 1]$ that is measurable with respect to $\mathcal{F}_i$. It is assumed that all p-values are valid, which means $\mathbb{P}(p_i \leq x) \leq x$ for all $\mathbb{P} \in H_i$ and $x \in [0, 1]$. Using p-values, a null hypothesis H_i is rejected if $p_i \leq \alpha_i$, where $\alpha_i \in [0, 1)$ is the individual significance level for H_i. We call a sequence

$(\alpha_i)_{i\in\mathbb{N}}$ *α-adjustment* and the multiple testing procedure $\boldsymbol{d} = (d_i)_{i\in\mathbb{N}}$ with $d_i = \mathbb{1}\{p_i \le \alpha_i\}$ *α-adjustment procedure*. An α-adjustment defines an online procedure, if α_i is measurable with respect to $\mathcal{F}_i$. In this case, we refer to it as *online α-adjustment*. It should be noted that in order to obtain online procedures with the required error rate control, p_i and α_i are often chosen to be measurable with respect to smaller σ-fields $\mathcal{X}_i \subseteq \mathcal{F}_i$ and $\mathcal{Y}_i \subseteq \mathcal{F}_i$, respectively, such that $\mathbb{P}(p_i \le \alpha_i | \mathcal{Y}_i) \le \alpha_i$ for all $\mathbb{P} \in H_i$. However, this is contained in our more general setting. For example, the literature often considers the case where solely the p-values $p_1, \dots, p_i$ are available at time $i \in \mathbb{N}$. Thus, we have $\mathcal{F}_i = \sigma(p_1, \dots, p_i)$. The individual significance levels $(\alpha_i)_{i\in\mathbb{N}}$ are then usually chosen as nonrandom functions of indicators (e.g., rejections) of the previous p-values, which ensures that α_i is measurable with respect to $\mathcal{F}_{i-1} \subseteq \mathcal{F}_i$. If the p-values are independent, we have $\mathbb{P}(p_i \le \alpha_i | \mathcal{F}_{i-1}) \le \alpha_i$ for all $\mathbb{P} \in H_i$.

One can also use α-adjustments to test intersection hypotheses H_I, $I \subseteq \mathbb{N}$. In this case, we only require an individual significance level α_i^I for each H_i with $i \in I$. We define $\boldsymbol{\alpha}_I = (\alpha_i^I)_{i\in I}$ as online sub α-adjustment, if α_i^I is measurable regarding $\mathcal{F}_i$ for all $i \in I$. With this, each $\boldsymbol{\alpha}_I$ defines an online intersection test ϕ_I by

$$\phi_I = \mathbb{1}\{\exists i \in I : p_i \le \alpha_i^I\}. \tag{3}$$

In what follows, we introduce the predictability condition for the family of online sub α-adjustments $(\boldsymbol{\alpha}_I)_{I\subseteq\mathbb{N}}$.

DEFINITION 4.1 (Predictable family of online sub α-adjustments). A family of online sub α-adjustments $(\boldsymbol{\alpha}_I)_{I\subseteq\mathbb{N}}$ is called *predictable*, if for all $I \subseteq \{1, \dots, i\}$ and $K = I \cup J$ with $J \subseteq \{k \in \mathbb{N} : k > i\}$ for some $i \in \mathbb{N}$ it holds that $\alpha_j^I = \alpha_j^K$ for all $j \in I$.

Note that this condition implies that the corresponding family of online intersection tests $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ defined by (3) is predictable as well (Definition 3.1), and hence the resulting closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$ is an online procedure (Theorem 3.2). However, it is not necessary for predictability of $\boldsymbol{\phi}$, as one could also choose $\alpha_j^I < \alpha_j^K$, where I and K are defined as in Definition 4.1. We use Definition 4.1, since we think it is the usual way of constructing predictable intersection tests based on α-adjustments, as we will also illustrate in Sections 4.1–4.3. Furthermore, in case of $\alpha_j^I < \alpha_j^K$, the significance levels are not monotone [16, 32], which is often assumed to obtain consonant intersection tests. Excluding this case helps to define the shortcut in Theorem 3.9 as an α-adjustment procedure, which we will show in the remainder of this section.

REMARK. Many online α-adjustments $(\alpha_i)_{i\in\mathbb{N}}$ in current literature can be defined by a fixed algorithm A that takes a finite vector of p-values as input and outputs an individual significance level such that $\alpha_i = A(p_1, \dots, p_{i-1})$ for all $i \in \mathbb{N}$. This ensures that α_i is measurable with respect to $\sigma(p_1, \dots, p_{i-1}) \subseteq \mathcal{F}_i$. Note that Definition 4.1 is always fulfilled if the same algorithm A is used for each online sub α-adjustment procedure, meaning that $\alpha_i^I = A((p_j)_{j\in I, j<i})$ for all $i \in I \subseteq \mathbb{N}$. To see this, let $I \subseteq \{1, \dots, i\}$ and $K = I \cup J$ with $J \subseteq \{j \in \mathbb{N} : j > i\}$. Then we have $\alpha_l^I = A((p_k)_{k\in I, k<l}) = A((p_k)_{k\in K, k<l}) = \alpha_l^K$ for all $l \in I$. This is also what Tian and Ramdas (2021) [37] formulated as initial attempt for extending the closure principle to the online setting. However, they did not show that this indeed leads to an online procedure. Furthermore, it is only a special case of our more general online closure principle, as we allow to use different algorithms for each intersection hypothesis and consider general online sub α-adjustments or even general online intersection tests.

If the family of online intersection tests $(\phi_I)_{I\subseteq\mathbb{N}}$ defined by (3) additionally satisfies the consonance property (2), the shortcut described in Theorem 3.9 can be expressed as an online α-adjustment procedure (see Appendix C for the proof).

THEOREM 4.2. *Assume* $(\boldsymbol{\alpha}_I)_{I\subseteq\mathbb{N}}$, *where* $\boldsymbol{\alpha}_I = (\alpha_i^I)_{i\in I}$, *is a predictable family of online sub* α*-adjustments such that* $\boldsymbol{\phi} = (\phi_I)_{I\subseteq N}$ *defined by* (3) *is a family of* α*-level intersection tests with the consonance property. Let us recursively define* $I_1 = \{1\}$ *and* $I_i = \{j \in \mathbb{N} : j < i, p_j > \alpha_j^{I_j}\} \cup \{i\}$ *for all* $i \geq 2$. *Then the following three procedures lead to the same decisions*:

1. *The online closed procedure* $\boldsymbol{d}^{\boldsymbol{\phi}}$.
2. *The shortcut* $\boldsymbol{d}^{\boldsymbol{\phi},s} = (d_i^{\boldsymbol{\phi},s})_{i\in\mathbb{N}}$, *where* $d_i^{\boldsymbol{\phi},s} = \phi_{I_i}$ *for all* $i \in \mathbb{N}$.
3. *The online* α*-adjustment procedure* $\boldsymbol{d} = (d_i)_{i\in\mathbb{N}}$, *where* $d_i = \mathbb{1}\{p_i \leq \alpha_i^{I_i}\}$ *for all* $i \in \mathbb{N}$.

In the remaining of this section, we apply Theorem 4.2 to construct new online α-adjustment procedures based on existing ones. We start with the Alpha-Spending [9], deriving a simple improvement of it in Section 4.1, which serves to exemplify how to use the proposed shortcuts. In Section 4.2, we show how an online version of the graphical procedure by Bretz et al. [3] can be obtained by the new online closure principle. Finally, we derive an improvement of the ADDIS-Spending under local dependence [37] (Section 4.3).

4.1. *Closed alpha-spending.* The Alpha-Spending is an online version of the weighted Bonferroni, meaning that the overall significance level α is split between the individual hypotheses $(H_i)_{i\in\mathbb{N}}$ according to some weights.

DEFINITION 4.3 (Alpha-Spending [9]). Let $(\gamma_i)_{i\in\mathbb{N}}$ be a nonnegative sequence of real numbers with $\sum_{i=1}^{\infty}\gamma_i \leq 1$. For a given FWER level α, *Alpha-Spending* tests for every $i \in \mathbb{N}$ the hypothesis H_i at the individual level

$$\alpha_i = \alpha\gamma_i.$$

The strong FWER control of the Alpha-Spending follows by the Bonferroni inequality [9]. However, as the Bonferroni, the Alpha-Spending is generally a conservative procedure, meaning that uniform improvements exist that could possibly be obtained by the closure principle. In order to derive a closure of the Alpha-Spending, we first have to formulate an online intersection test based on the Alpha-Spending. Here, we just apply the Alpha-Spending on a subsequence by ignoring the p-values that are not contained in it.

DEFINITION 4.4 (Alpha-Spending intersection test). Let $(\gamma_i)_{i\in\mathbb{N}}$ be as in Alpha-Spending. The *Alpha-Spending intersection test* ϕ_I is defined by (3), where $\alpha_i^I = \alpha\gamma_{t_I(i)}$ with $t_I(i) = |\{j \in I : j \leq i\}|$, for all $I \subseteq \mathbb{N}$.

We assume that the same $(\gamma_i)_{i\in\mathbb{N}}$ is used for all intersection tests ϕ_I. Note that for determining α_i^I with $i \in I \subseteq \mathbb{N}$ it is only important how many indices $j < i$ are included in I, but we do not need information about the indices that are greater than i. This ensures the predictability (Definition 4.1) of $(\boldsymbol{\alpha}_I)_{I\subseteq\mathbb{N}}$, where $\boldsymbol{\alpha}_I = (\alpha_i^I)_{i\in I}$. However, in general, $(\phi_I)_{I\subseteq\mathbb{N}}$ does not have the consonance property, and thus we cannot apply Theorem 4.2. For example, consider $(\gamma_i)_{i\in\mathbb{N}} = (0, 1, 0, 0, \ldots)$. If $p_2 \leq \alpha$, we have $\phi_{\{1,2\}} = 1$ but $\phi_{\{1\}} = 0$ and $\phi_{\{2\}} = 0$. Hence, the consonance property is not satisfied. Also note that for the "consonantized" intersection tests (see Section 3.1), we have $\tilde{\phi}_I = \max\{d_i^{\phi} : i \in I\} = 0$ for all $I \subseteq \mathbb{N}$, since $\phi_{\{i\}} = 0$ for all $i \in \mathbb{N}$. This exemplifies how requiring consonance can help to identify the inefficiency of closed procedures [31]. We can ensure to obtain consonant Alpha-Spending intersection tests by choosing $(\gamma_i)_{i\in\mathbb{N}}$ to be nonincreasing. To see this, consider $I \subseteq \mathbb{N}$ with $\phi_I = 1$. Then there exists an $i \in I$ such that $p_i \leq \alpha_i^I = \alpha\gamma_{t_I(i)}$. Now for $J \subseteq I$ with $i \in J$,

it holds that $t_J(i) \leq t_I(i)$. If $(\gamma_i)_{i\in\mathbb{N}}$ is nonincreasing, this implies $p_i \leq \alpha\gamma_{t_J(i)} = \alpha_i^J$, and hence $\phi_J = 1$. Note that it is fairly common to choose $(\gamma_i)_{i\in\mathbb{N}}$ to be nonincreasing, since it needs to converge to 0 anyway. This leads to the following new online closed α-adjustment procedure.

DEFINITION 4.5 (Closed Alpha-Spending). Let $(\gamma_i)_{i\in\mathbb{N}}$ be as in Alpha-Spending but nonincreasing. *Closed Alpha-Spending* updates the individual significance levels as follows:

$$\alpha_i = \alpha\gamma_{t(i)},$$

where $t(i) = 1 + \sum_{j=1}^{i-1}(1 - d_j)$ and $d_j = \mathbb{1}\{p_j \leq \alpha_j\}$.

PROPOSITION 4.6. *Closed Alpha-Spending controls the FWER in the strong sense.*

PROOF. Let $(\phi_I)_{I\subseteq\mathbb{N}}$ be a family of Alpha-Spending intersection tests based on the same nonincreasing $(\gamma_i)_{i\in\mathbb{N}}$. Due to Theorem 4.2, the individual significance levels of the resulting closed procedure are given by $\alpha_i^{I_i} = \alpha\gamma_{t_{I_i}(i)}$, where

$$t_{I_i}(i) = 1 + \left|\left\{j < i : p_j > \alpha_j^{I_j}\right\}\right| = 1 + \sum_{j=1}^{i-1}\left(1 - \mathbb{1}\left\{p_j \leq \alpha_j^{I_j}\right\}\right).$$

Hence, the FWER control follows by the online closure principle (Theorem 3.2). □

It is easy to verify that the closed Alpha-Spending (Definition 4.5) is a uniform improvement of the Alpha-Spending (Definition 4.3). Alternative closures of the Alpha-Spending procedure can be derived, which are often online variants of the Bonferroni-based closed procedures [16].

REMARK. By applying the Alpha-Spending intersection test (Definition 4.4) to every intersection hypothesis H_I, the significance level might not be fully exhausted when I is finite. However, this is inevitable to obtain a predictable family of online intersection tests. Suppose a closed procedure where for each intersection hypothesis H_I an online sub α-adjustment $(\alpha_i^I)_{i\in I}$ is chosen such that $\sum_{i\in I}\alpha_i^I = \alpha$. Then the intersection hypothesis $H_{\{i\}}$ would be rejected if $p_i \leq \alpha$. Now consider $J \subseteq \mathbb{N}$ with $\min(J) = i$. The only option such that $\phi_{\{i\}} = 1$ implies $\phi_J = 1$ is to choose $\alpha_i^J = \alpha$ and, therefore, $\alpha_j^J = 0$ for all $j \in J \setminus \{i\}$. The resulting online closed procedure is the online analogue of the fixed sequence procedure [23]: reject H_i if $p_1 \leq \alpha, \ldots, p_i \leq \alpha$. Thus, if $p_j > \alpha$ for a $j \in \mathbb{N}$, no hypothesis H_k with $k \geq j$ can be rejected. This is unfavorable in most online scenarios.

4.2. *Online-graph.* In classical multiple testing, the graphical procedure by Bretz et al. [3] has grown in popularity over the last years due to its easy interpretability and high power. Its FWER control is shown by the closure principle and by applying a weighted Bonferroni test to each intersection hypothesis. Tian and Ramdas [37] have built on the result and extended the graphical procedure to the online setting, which led to an Online-Graph (they termed it Online-Fallback in [37]). In this subsection, we give the formal description of the aforementioned graphical procedure and afterwards sketch how the Online-Graph can be obtained by the proposed online closure principle.

DEFINITION 4.7 (Graphical procedure [3]). Let $\mathcal{H} = \{H_1, \ldots, H_m\}$ be the set of hypotheses to be tested, $\{p_1, \ldots, p_m\}$ the corresponding p-values and $\{\alpha_1, \ldots, \alpha_m\}$ the initial

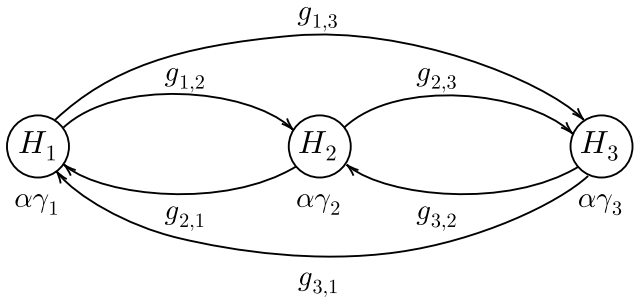

FIG. 2. *Illustration of the graphical procedure by* [3] *for* $m = 3$ *hypotheses.*

allocation of the overall significance level, where for all $i \in \{1, \dots, m\}$: $\alpha_i = \alpha\gamma_i$ such that $\sum_{i=1}^m \gamma_i \leq 1$. In addition, let $\boldsymbol{G} = (g_{i,j})_{i,j\in\{1,\dots,m\}}$ be a matrix containing nonnegative weights such that $g_{i,i} = 0$ and $\sum_{j=1}^m g_{i,j} \leq 1$ for all $i \in \{1, \dots, m\}$. Then the graphical α-adjustment is defined by the following stepwise algorithm:

0. Set $I = \{1, \dots, m\}$.
1. Let $i = \arg\min_{j\in I} \frac{p_j}{\alpha_j}$. If $p_i > \alpha_i$, stop and accept all hypotheses that have not been rejected yet.
2. Reject H_i and update I, the individual significance levels and the weights as follows:

$$I \to I \setminus \{i\}, \qquad \alpha_j \to \begin{cases} \alpha_j + \alpha_i g_{i,j} & j \in I, \\ 0 & \text{otherwise}, \end{cases}$$

$$g_{j,k} \to \begin{cases} \dfrac{g_{j,k} + g_{j,i} g_{i,k}}{1 - g_{j,i} g_{i,j}} & j, k \in I,\ j \neq k, \\ 0 & \text{otherwise}. \end{cases}$$

3. If $|I| \geq 1$, go to step 1. Else, stop.

In Figure 2, the graphical procedure (Definition 4.7) is illustrated for $m = 3$ hypotheses. Below each hypothesis is the initial individual significance level. The arrows represent the allocation of the weights after rejecting a hypothesis. After each rejection, the graph is updated according to Step 2 of the graphical procedure.

In online multiple testing, only future hypotheses are allowed to benefit from a rejection. Since there are no arrows pointing back, the weights $g_{j,k}$ do not need to be updated at any step. To derive the Online-Graph by the online closure principle, we first define online intersection tests for each $I \subseteq \mathbb{N}$ that are based on Alpha-Spending (Definition 4.3). The idea is to start with the same initial allocation of the significance level as in Alpha-Spending, which is given by $(\gamma_i)_{i\in\mathbb{N}}$. That means $H_{\mathbb{N}}$ is rejected if $p_i \leq \alpha_i^{\mathbb{N}} = \alpha\gamma_i$ for at least one $i \in \mathbb{N}$. Now consider H_I for some index set $I \subseteq \mathbb{N}$. If $j \notin I$, then the individual significance level of H_j is distributed to the future hypotheses according to weights $(g_{j,i})_{i=j+1}^\infty$ such that $\sum_{i=j+1}^\infty g_{j,i} \leq 1$. Since significance level is only assigned to future hypotheses, we have $\alpha_1^I = \alpha\gamma_1$ for all $I \subseteq \mathbb{N}$ with $1 \in I$ and the other levels can be defined recursively by

$$\alpha_i^I = \alpha\left(\gamma_i + \sum_{j<i,\, j\notin I} g_{j,i}\alpha_j^{(I\cup j)}\right) \quad (i \in I).$$

The above derivation implies that $\sum_{i\in I} \alpha_i^I \leq \alpha \sum_{i\in\mathbb{N}} \gamma_i \leq \alpha$, such that each of these online sub α-adjustments $\boldsymbol{\alpha}_I = (\alpha_i^I)_{i\in I}$ defines an online α-level intersection test ϕ_I by (3). In addition, the predictability of $(\boldsymbol{\alpha}_I)_{I\subseteq\mathbb{N}}$ and consonance of $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ can easily be verified.

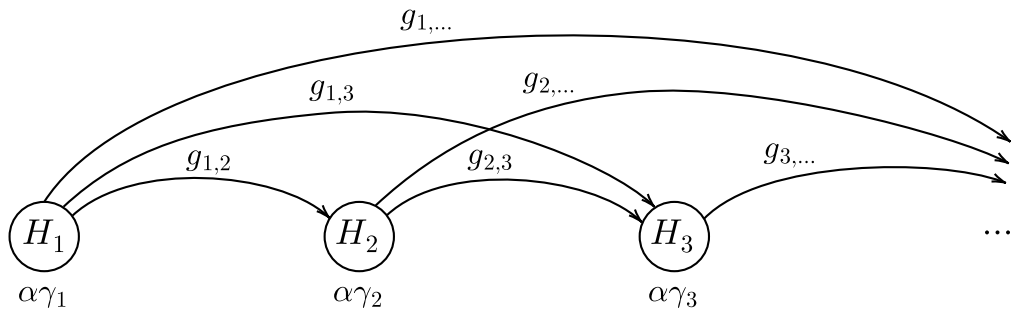

FIG. 3. *Illustration of the Online-Graph.*

The Online-Graph is then obtained as the shortcut of the online closed procedure $\boldsymbol{d^{\phi}}$, and thus is defined by

$$\alpha_i = \alpha_i^{I_i} = \alpha\Big(\gamma_i + \sum_{j<i} g_{j,i}\alpha_j d_j\Big) \quad (i \in \mathbb{N}),$$

where $d_j = \mathbb{1}\{p_j \le \alpha_j\}$ (see Theorem 4.2).

Figure 3 shows the illustration of the Online-Graph. Compared to Figure 2, the arrows only point to future hypotheses. As already mentioned, this means that only future significance levels are updated after a rejection. In addition, the weights $(g_{j,i})_{j\in\mathbb{N},i>j}$ remain the same over the entire testing process and do not have to be updated. The points at the end indicate that there is an infinite number of future hypotheses.

REMARK.

1. Note that not all graphs with an infinite number of hypotheses are special cases of this Online-Graph. One could also think of graphical procedures that allocate significance levels to previous hypotheses. For example, suppose that after each rejection the significance level of the rejected hypothesis is distributed to the first hypothesis. This could be the case, for instance, if the hypothesis H_1 is of main interest and the rejection of one of the future hypotheses should increase the probability of rejecting H_1. Obviously, this does not define an online procedure. Nevertheless, it can be written as a closed procedure resulting from the following Alpha-Spending based online intersection test: $\phi_I = 1$, if $p_1 \le \alpha(\gamma_1 + \sum_{i\notin I}\gamma_i)\mathbb{1}\{1 \in I\}$ or $p_i \le \alpha\gamma_i$ for at least one $i \in I$. This again clarifies that the predictability of $(\phi_I)_{I\subseteq\mathbb{N}}$ (Definition 3.1) is needed in order to obtain online closed procedures.

2. There is a strong connection between the Online-Graph and closed Alpha-Spending (Definition 4.5). If we would choose $\tilde{g}_{j,i} = (\gamma_{t(j)+i-j-1} - \gamma_{t(j)+i-j})/\gamma_{t(j)}$ as weights for the Online-Graph, where $t(j) = 1 + \sum_{k=1}^{j-1}(1-d_k)$, both procedures would be the same. However, in general the weights are not allowed to depend on the previous rejections, and thus we would not consider closed Alpha-Spending as a special case of the Online-Graph. Nevertheless, in some cases both procedures collapse. For example, if $\gamma_i = q^i(1-q)/q$ for some $q \in (0,1)$, we have $\tilde{g}_{j,i} = q^{i-j}(1-q)/q = \gamma_{i-j}$, which is independent of the data.

4.3. *Closed ADDIS-spending.* Although the Online-Graph is a uniform improvement of the Alpha-Spending and its offline version is one of the most popular procedures in classical multiple testing, in particular for clinical trials, Tian and Ramdas (2021) [37] claimed that their proposed ADDIS-Spending is to be preferred in the online setting.

ADDIS-Spending combines the multiple testing approaches of discarding large p-values using a parameter τ [39] and adapting to the number of false hypotheses using a parameter λ

[34]. Suppose that the p-values are uniformly valid, which means $\mathbb{P}(p_i \leq x\tau | p_i \leq \tau) \leq x$ for all $x, \tau \in [0,1]$, $\mathbb{P} \in H_i$, and let $\tau_i \in (0,1]$, $\lambda_i \in [0, \tau_i)$ and $\alpha_i \in [0, \tau_i)$ be fixed parameters for all $i \in \mathbb{N}$. Then, by the Bonferroni inequality, for all $\mathbb{P} \in \mathcal{P}$, we have

$$\begin{aligned}
\mathrm{FWER}_{\mathbb{P}} &\leq \sum_{i \in I_0^{\mathbb{P}}} \mathbb{P}(p_i \leq \alpha_i) \\
&= \sum_{i \in I_0^{\mathbb{P}}} \mathbb{P}(p_i \leq \alpha_i | p_i \leq \tau_i) \mathbb{P}(p_i \leq \tau_i) \\
&\leq \sum_{i \in I_0^{\mathbb{P}}} \frac{\alpha_i}{\tau_i} \mathbb{P}(p_i \leq \tau_i) \\
&\leq \sum_{i \in I_0^{\mathbb{P}}} \frac{\alpha_i}{\tau_i} \mathbb{P}(p_i \leq \tau_i) \frac{\mathbb{P}(p_i > \lambda_i | p_i \leq \tau_i)}{1 - \lambda_i/\tau_i} \\
&\leq \mathbb{E}_{\mathbb{P}}\left[\sum_{i \in \mathbb{N}} \frac{\alpha_i}{\tau_i - \lambda_i} \mathbb{1}\{\lambda_i < p_i \leq \tau_i\} \right],
\end{aligned}$$

where the second and third inequality follow from the uniform validity of the p-values. We restricted to fixed parameters for simplicity. However, the same calculation can be done when τ_i, λ_i and α_i only depend on information that is independent of p_i, by conditioning on this information [37]. Also, note that in the case of $\tau_i = 1$, the uniform validity would not be needed, and thus this assumption is only required for the discarding and not the adaptive part [37]. Anyway, uniform validity is fulfilled in many settings [37, 39]. The above calculation shows that in order to control the FWER, it is sufficient to ensure

$$\sum_{i \in \mathbb{N}} \frac{\alpha_i}{\tau_i - \lambda_i} \mathbb{1}\{\lambda_i < p_i \leq \tau_i\} \leq \alpha. \tag{4}$$

The idea is that we can reuse the significance level if $P_i \leq \lambda_i$ or $P_i > \tau_i$, but need to subtract the larger significance level $\alpha_i/(\tau_i - \lambda_i)$ if $\lambda_i < P_i \leq \tau_i$. Since p-values corresponding to true hypotheses tend to be large and p-values corresponding to false hypotheses tend to be small, we expect this tradeoff to be useful. In order to exploit this condition, the individual significance levels need to depend on information about the p-values observed so far. Since τ_i, λ_i and α_i need to be independent of p_i, more assumptions on the dependence structure of the p-values are needed.

An example is local dependence [41], which allows p-values that are close together in time to depend on each other, while p-values that are further apart are independent. This is an intuitive condition, because one would think that p-values that are closer in time are stronger related than those with a large time gap. For example, local dependence encompasses batch dependence, where p-values within one batch may depend on each other but p-values from different batches are independent. This is the case in practice, for instance, if the used data is replaced by independent data after a period of time. Mathematically, local dependence is defined as follows.

DEFINITION 4.8 (Local dependence [41]). Let $(l_i)_{i \in \mathbb{N}}$ be a fixed sequence of parameters such that $l_i \in \{0, 1, \ldots, i-1\}$ and $l_{i+1} \leq l_i + 1$ for all $i \in \mathbb{N}$. A sequence of p-values $(p_i)_{i \in \mathbb{N}}$ is called *locally dependent* with the lags $(l_i)_{i \in \mathbb{N}}$, if $\forall i \in \mathbb{N}$ holds:

$$p_i \perp p_{i-l_i-1}, p_{i-l_i-2}, \ldots, p_1.$$

If $l_i = 0$ for all $i \in \mathbb{N}$, all p-values are independent. In the other extreme case, $l_i = i - 1$ for all $i \in \mathbb{N}$, all p-values are dependent. Although it is assumed that the lags are constant parameters, in practice, one does not have to know all l_i at the beginning of the evaluation. However, one must determine l_i before testing hypothesis H_i without using the data itself. Tian and Ramdas (2021) [37] proposed the following online procedure that satisfies condition (4) under local dependence, and thus controls the FWER in the strong sense when the p-values are uniformly valid.

DEFINITION 4.9 (ADDIS-Spending under local dependence). Assume local dependence with lags $(l_i)_{i\in\mathbb{N}}$ and let $(\gamma_i)_{i\in\mathbb{N}}$ be a nonincreasing sequence of weights for an Alpha-Spending (Definition 4.3). In addition, let $(\tau_i)_{i\in\mathbb{N}}$ and $(\lambda_i)_{i\in\mathbb{N}}$ be sequences of random variables such that $\tau_i \in (0, 1]$ and $\lambda_i \in [0, \tau_i)$ are measurable with respect to $\mathcal{G}_{i-l_i-1}$ for all $i \in \mathbb{N}$, where $\mathcal{G}_{i-l_i-1} = \sigma(\{p_1, \dots, p_{i-l_i-1}\})$. The *ADDIS-Spending* under local dependence updates the individual significance levels as follows:

$$\alpha_i = \alpha(\tau_i - \lambda_i)\gamma_{t(i)},$$

where $t(i) = 1 + l_i + \sum_{j=1}^{i-l_i-1}(s_j - c_j)$, $s_j = \mathbb{1}\{p_j \le \tau_j\}$ and $c_j = \mathbb{1}\{p_j \le \lambda_j\}$.

We investigate next, whether the above ADDIS-Spending procedure can be improved by an online closed procedure. For this closed procedure, we first define an ADDIS-Spending intersection test. To this end, we define the index set $L_i := \{i - l_i, \dots, i - 1\}$ of previous p-values that depend on p_i.

DEFINITION 4.10 (ADDIS-Spending intersection test). Assume local dependence with lags $(l_i)_{i\in\mathbb{N}}$ and let $(\gamma_i)_{i\in\mathbb{N}}$, $(\tau_i)_{i\in\mathbb{N}}$ and $(\lambda_i)_{i\in\mathbb{N}}$ be as in ADDIS-Spending (Definition 4.9). The *ADDIS-Spending intersection test* ϕ_I is defined by (3), where

$$\alpha_i^I = \alpha(\tau_i - \lambda_i)\gamma_{t_I(i)}$$

with $t_I(i) = 1 + |L_i \cap I| + \sum_{j\le i-l_i-1,\, j\in I}(s_j - c_j)$ for all $i \in I$.

FWER control of the ADDIS-Spending directly implies that the ADDIS-Spending intersection test is an online α-level intersection test under local dependence and uniform validity of p-values. In addition, as with the Alpha-Spending, it can easily be verified that the family of online sub α-adjustments $(\boldsymbol{\alpha}_I)_{I\subseteq\mathbb{N}}$ is predictable and the corresponding family of ADDIS-Spending intersection tests $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ satisfies the consonance property when the same parameters $(\gamma_i)_{i\in\mathbb{N}}$, $(\tau_i)_{i\in\mathbb{N}}$ and $(\lambda_i)_{i\in\mathbb{N}}$ are used for each intersection test. With that, the shortcut of the closed procedure can be obtained by Theorem 4.2.

DEFINITION 4.11 (Closed ADDIS-Spending). Assume local dependence with lags $(l_i)_{i\in\mathbb{N}}$ and let $(\gamma_i)_{i\in\mathbb{N}}$, $(\tau_i)_{i\in\mathbb{N}}$ and $(\lambda_i)_{i\in\mathbb{N}}$ be as in ADDIS-Spending (Definition 4.9). *Closed ADDIS-Spending* updates the individual significance levels as follows:

$$\alpha_i = \alpha(\tau_i - \lambda_i)\gamma_{t(i)},$$

where $t(i) = 1 + \sum_{j=1}^{i-l_i-1}(s_j - \max\{c_j, d_j\}) + \sum_{j=i-l_i}^{i-1}(1 - d_j)$ with $d_j = \mathbb{1}\{p_j \le \alpha_j\}$.

PROPOSITION 4.12. *Closed ADDIS-Spending controls the FWER in the strong sense under local dependence when the p-values are uniformly valid.*

PROOF. Let $(\phi_I)_{I\subseteq\mathbb{N}}$ be a family of ADDIS-Spending intersection tests based on the same parameters $(\gamma_i)_{i\in\mathbb{N}}$, $(\tau_i)_{i\in\mathbb{N}}$ and $(\lambda_i)_{i\in\mathbb{N}}$. Due to Theorem 4.2, the individual significance levels of the resulting closed procedure are given by $\alpha_i^{I_i} = \alpha(\tau_i - \lambda_i)\gamma_{t_{I_i}(i)}$, where

$$t_{I_i}(i) = 1 + \big|L_i \cap \{j < i : \alpha_j^{I_j} > p_j\}\big| + \sum_{j\le i-l_i-1,\, p_j>\alpha_j^{I_j}} (s_j - c_j)$$

$$= 1 + \sum_{j=i-l_i}^{i-1} \big(1 - \mathbb{1}\{p_j \le \alpha_j^{I_j}\}\big) + \sum_{j=1}^{i-l_i-1} (s_j - c_j)\big(1 - \mathbb{1}\{p_j \le \alpha_j^{I_j}\}\big).$$

Hence, the FWER control follows by the online closure principle (Theorem 3.2). □

Note that $\sum_{j=i-l_i}^{i-1}(1-d_j) \le l_i$ and, therefore, $t(i)$ in Definition 4.11 is never larger than $t(i)$ in Definition 4.9. Since $(\gamma_i)_{i\in\mathbb{N}}$ is nonincreasing, closed ADDIS-Spending never rejects less hypotheses than ADDIS-Spending. Furthermore, if there are dependent p-values, which means $l_i > 0$ for some $i \in \mathbb{N}$, closed ADDIS-Spending is a real uniform improvement of ADDIS-Spending. In this case, ADDIS-Spending can only gain significance level from independent p-values, while closed ADDIS-Spending additionally allows to gain from rejections of dependent hypotheses. For example, let p_1 and p_2 depend on each other and $p_1 \le \alpha_1$. Then H_2 is tested at level $\alpha(\tau_2 - \lambda_2)\gamma_2$ using ADDIS-Spending and at level $\alpha(\tau_2 - \lambda_2)\gamma_1$ using closed ADDIS-Spending. If $\lambda_i < \alpha_i$ for some $i \in \mathbb{N}$, we have an additional improvement, since $\max\{c_i, d_i\} > c_i$ if $\lambda_i < P_i \le \alpha_i$. In particular, for $\lambda_i = 0$ closed ADDIS-Spending provides a uniform improvement of Discard-Spending [37] under independent p-values.

5. Simulation study. In this section, we aim to quantify the gain in power when using closed ADDIS-Spending instead of ADDIS-Spending and show its FWER control by means of simulations. For this purpose, we first describe the simulation design and then show the results. We considered similar simulation scenarios as in [37], but generated locally dependent p-values instead of independent p-values.

5.1. *Simulation design*. We simulated trials where $n = 1000$ null hypotheses $(H_i)_{i\in\{1,\dots,n\}}$ are tested sequentially. We assume that the local dependence structure of the p-values is given by finite batches B_i, $i \in \mathbb{N}$, with a fixed batch-size $b \in \mathbb{N}$. That means we have batches $B_1 = \{p_1, \dots, p_b\}$, $B_2 = \{p_{b+1}, \dots, p_{2b}\}$ and so forth, and the p-values within one batch depend on each other while p-values from different batches are independent. For this simulation, we considered $b \in \{1, 10, 25, 100\}$.

Let $X^{1:b}, X^{(b+1):2b}, \dots, X^{(n-b+1):n} \overset{\text{i.i.d.}}{\sim} N_b(\mu_0, \Sigma)$, where $X^{j:i} = (X_j, \dots, X_i)^T$ for $i \ge j$, and where N_b is the b-dimensional normal distribution, $\mu_0 = (0, \dots, 0)^T \in \mathbb{R}^b$ and $\Sigma = (\sigma_{ij})_{i,j=1,\dots,b} \in \mathbb{R}^{b\times b}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho \in (0, 1)$ for all $i \in \{1, \dots, b\}$ and $j \ne i$. For each $i \in \{1, \dots, n\}$, we test the null hypothesis $H_i : \mathbb{E}[Z_i] \le 0$, where $Z_i = X_i + \mu_A$, $\mu_A > 0$, with probability $\pi_A \in (0, 1)$ and $Z_i = X_i + \mu_N$, $\mu_N \le 0$, otherwise. The p-values are calculated by $p_i = \Phi(-Z_i)$, where Φ is the cumulative distribution function (CDF) of a standard normal distribution. Thus, for a p-value of a true hypothesis $p_i = \Phi(-X_i - \mu_N)$, $i \in I_0^{\mathbb{P}}$ and $x \in [0, 1]$:

$$\mathbb{P}(p_i \le x) = \mathbb{P}\big(X_i \le \Phi^{-1}(x) + \mu_N\big) = \Phi\big(\Phi^{-1}(x) + \mu_N\big).$$

If $\mu_N = 0$, we have uniformly distributed null p-values, which means that $\mathbb{P}(p_i \le x) = x$ for all $i \in I_0^{\mathbb{P}}$ and $x \in [0, 1]$ while a null p-value p_i, $i \in I_0^{\mathbb{P}}$, is said to be conservative, if $\mathbb{P}(p_i \le x) < x$ for some $x \in [0, 1]$ [39]. Since $\Phi(x)$ is increasing in x, the null p-values are

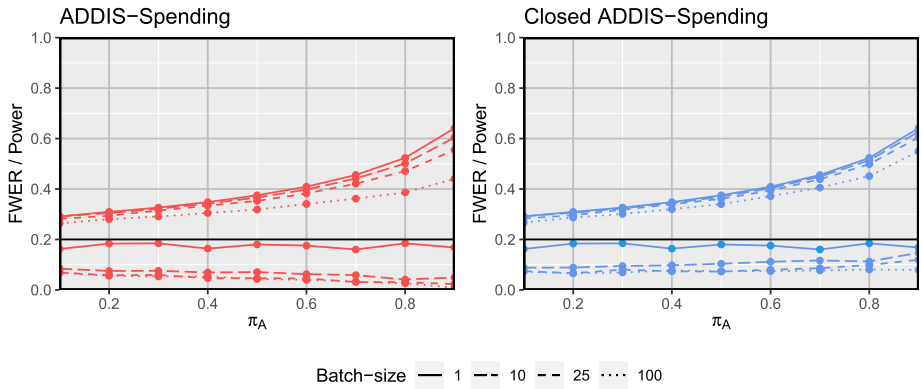

FIG. 4. *Comparison of ADDIS-Spending and closed ADDIS-Spending in terms of power and FWER* ($\alpha = 0.2$) *based on locally dependent p-values for different batch sizes and proportions of false null hypotheses* (π_A); $n = 1000$, $\mu_N = 0$, $\mu_A = 4$ *and* $\rho = 0.8$ *in both plots.*

conservative if and only if $\mu_N < 0$ and the conservativeness grows with decreasing μ_N. The parameters π_A and μ_A can be interpreted as proportion of false hypotheses and strength of the alternative, respectively.

For each considered scenario, we simulated 2000 independent trials and estimated the power and FWER using closed ADDIS-Spending and ADDIS-Spending.

5.2. *Comparison of closed ADDIS-spending and ADDIS-spending through simulations.* We compared the results when using closed ADDIS-Spending and ADDIS-Spending with respect to FWER and power for different batch-sizes and proportions of false null hypotheses. The FWER is represented by the lines below the global significance level $\alpha = 0.2$ and the power by the lines above it. Thereby, the power is defined as the expected number of rejected hypotheses among all false hypotheses. Both procedures are applied with the parameters $\gamma_i = \frac{6}{\pi^2 i^2}$, $\lambda_i = 0.3$ and $\tau_i = 0.8$ for all $i \in \mathbb{N}$. We do not claim that this choice leads to the highest possible power, but it works well to show the differences between the presented procedures.

The results with uniformly distributed null p-values and $\mu_A = 4$ can be found in Figure 4. As discussed before, closed ADDIS-Spending and ADDIS-Spending coincide under independence of the p-values ($b = 1$), and by this the solid lines are identical. However, when some of the p-values become dependent ($b > 1$), the power and FWER decrease drastically using ADDIS-Spending, while closed ADDIS-Spending decelerates this decrease such that a higher power is obtained. In addition, we see that closed ADDIS-Spending exhausts the FWER more. Simulations for other parameter choices of n, μ_N and μ_A can be found in Appendix B. In all cases, no differences in the behavior of closed ADDIS-Spending and ADDIS-Spending are observed compared to those shown in Figure 4.

6. Application on real data. In this section, we apply the presented procedures to real data. First, we consider the IMPC data set, which aims to identify the influence on the phenotype of each protein-coding mouse gene [24] and which is a standard data set used in the online multiple testing literature [29, 37]. Second, we apply the procedures on the RECOVERY platform trial [33]. Here, several treatments for severe COVID-19 diseases are tested against a standard of care. The two data sets have significant differences. The IMPC data includes several thousand experiments, while the RECOVERY trial has only tested 12 treatments to date. It is important to note that we are not attempting to draw any conclusions from these data sets; they are merely being used for illustrative purposes.

TABLE 1
Number of rejections obtained by different procedures applied on IMPC data

Procedure	Number of rejections		
	$h = 1.3$	$h = 1.1$	$h = 1.05$
Alpha-Spending	1348	1404	1404
Closed Alpha-Spending	1394	1426	1427
Online-Graph	1424	1426	1423
ADDIS-Spending	1427	1459	1454
Closed ADDIS-Spending	1434	1465	1460

6.1. *IMPC data.* The International Mouse Phenotyping Consortium (IMPC) is coordinating a large-scale study to determine the function of all protein-coding mouse genes. To this end, each gene is systematically knocked out and the effect on the phenotype is explored. As the data set grows over time due to the testing of additional genes, the use of online multiple testing procedures is appropriate [29, 37]. Our evaluation is based on the data set in the Zenodo repository https://zenodo.org/record/2396572 [30]. The contained p-values resulted from the analysis in [19] and follow a batch dependence structure since the same group of mice was used for testing several consecutive hypotheses [37]. In our evaluation, we restrict to 5000 of these p-values that are arranged in 84 batches.

We compare the number of rejections obtained by Alpha-Spending, closed Alpha-Spending, Online-Graph, ADDIS-Spending and closed ADDIS-Spending at the FWER level $\alpha = 0.2$. We choose $(\gamma_i)_{i\in\mathbb{N}}$ such that $\gamma_i \propto 1/i^h$ for all $i \in \mathbb{N}$ and $\sum_{i\in\mathbb{N}} \gamma_i = 1$, where $h \in \{1.05, 1.1, 1.3\}$. Note that the larger the h, the faster $(\gamma_i)_{i\in\mathbb{N}}$ converges to 0. As done in [37], we set $\tau_i = 0.8$ and $\lambda_i = 0.16$ for the ADDIS procedures. In addition, we choose $g_{j,i} = \gamma_{i-j}$ for the Online-Graph.

The results are summarized in Table 1. As expected, the Alpha-Spending leads to the least rejections. Closed Alpha-Spending and Online-Graph performed similarly. However, Online-Graph led to more rejections when $(\gamma_i)_{i\in\mathbb{N}}$ decreased faster ($h = 1.3$), while closed Alpha-Spending was superior in case of a slowly decreasing $(\gamma_i)_{i\in\mathbb{N}}$ ($h = 1.05$). Both procedures were outperformed by ADDIS-Spending, which was further improved by closed ADDIS-Spending.

6.2. *RECOVERY platform trial.* In a platform trial, several treatment arms $T_1, T_2, \ldots$ are compared to the same control group. In contrast to multiarm trials, the treatment arms do not enter or leave the trial at the same time and the total number of treatments under evaluation is not predefined, leading to an online testing problem [28]. Usually, concurrent control data is used, meaning that a treatment arm is only compared to control patients that were randomized while the treatment arm was in the platform. This leads to a local dependence structure of the p-values, since treatment arms that overlap share some control data for testing and those that do not overlap can be considered as independent.

In this section, we compare the rejections achieved by the considered methods when applied to a real ongoing platform trial. The Randomized Evaluation of COVID-19 Therapy (RECOVERY) trial has already tested twelve treatments for severe COVID-19 diseases against a standard of care, while another one is currently recruiting [33]. All p-values are available at the website https://www.recoverytrial.net/. The overlapping structure is illustrated in a publication by the data monitoring committee [33].

We apply the same procedures as in Section 6.1. However, since the required evidence in such clinical trials is usually higher, we set $\alpha = 0.05$. In addition, we choose $\gamma_i = i^q(1-q)/q$, $i \in \mathbb{N}$, for $q \in \{0.6, 0.7, 0.8\}$. This change is because harmonic sequences tend to decrease

TABLE 2
Number of rejections and current significance level α_{13} obtained by different procedures applied on the RECOVERY trial

Procedure	Number of rejections			α_{13}		
	$q = 0.6$	$q = 0.7$	$q = 0.8$	$q = 0.6$	$q = 0.7$	$q = 0.8$
Alpha-Spending	1	2	2	0.00004	0.00021	0.00069
Closed Alpha-Spending	2	2	3	0.00012	0.00042	0.00134
Online-Graph	2	2	3	0.00012	0.00042	0.00134
ADDIS-Spending	2	3	3	0.00060	0.00112	0.00168
Closed ADDIS-Spending	2	3	3	0.00060	0.00161	0.00210

very fast at the beginning of the sequence. This is negligible if the h is rather low, as in Section 6.1. However, in this example, we would choose larger h as the number of hypotheses is much lower. This is why we expect a geometric sequence to perform better in low-scale settings. Note that in this case, $(\gamma_i)_{i\in\mathbb{N}}$ decreases faster for lower q.

We compare the number of rejections and the current individual significance level α_{13} that would be used to test the next treatment, which is already in the trial but has not yet finished recruitment (see Table 2). The behavior of the procedures looks similar as in Section 6.1. However, while the number of rejections does not differ much, closed ADDIS-Spending tests the current hypothesis at the highest level such that the differences between the number of rejections will possibly be larger when further hypotheses are tested. As noted in Section 4.2, closed Alpha-Spending and Online-Graph coincide when $(\gamma_i)_{i\in\mathbb{N}}$ is proportional to a geometric sequence and $g_{j,i} = \gamma_{i-j}$.

7. Discussion. Contemporary problems, for example, platform trials and genetics research studies, require control of the FWER in unbounded and sequential multiple testing settings [28, 29]. Since the closure principle is fundamental for the construction of multiple testing procedures with FWER control, an extension of the theory is essential. We introduced a novel online closure principle, including a predictability condition for the family of online intersection tests, which ensures that the resulting closed testing procedure can be applied in the online setting. Important properties that hold in the classical multiple testing case were transferred to the class of online closed procedures. It was shown that all online procedures with FWER control are also online closed procedures. With this, one can focus on the construction of online closed procedures when aiming for FWER control. Moreover, we proved that one can restrict to consonant families of intersection tests and provided a sufficient condition under which the event of rejecting any hypothesis cannot be enlarged. In addition, we showed how shortcuts of online closed procedures can be obtained under consonance. These have a simpler form than in classical multiple testing as the rejection of an intersection hypothesis uniquely determines which individual hypothesis is to be rejected. We used this to derive individual significance levels for shortcuts that are based on α-adjustments. With that, new online closed procedures can be derived easily that often improve existing ones, which we have demonstrated with the examples of Alpha-Spending and ADDIS-Spending.

In this paper, we focused on the construction of FWER controlling procedures. In online multiple testing, however, many applications aim for a less conservative error rate, for example, FDR. The reason for this can also be seen when we look at the online procedures that were considered in this paper (e.g., Definitions 4.5 and 4.11). Except for unrealistic extreme cases, the individual significance levels $(\alpha_i)_{i\in\mathbb{N}}$ of all these procedures will tend to 0 for i to infinity. That is because every additional test increases the probability of committing at least

one type I error, and thus increases the FWER, whereas rejections may lead to a decrease of the FDR. One could say that both types of procedures have an overall level α available at the beginning of the testing process. But while FWER controlling procedures can only spend the level on testing, FDR controlling procedures also allow to gain additional significance level from rejections [9]. Nevertheless, in practice an online multiple testing problem does not necessarily mean that thousands or even millions of hypotheses will be tested, but rather that the number and concrete structure of hypotheses that will be tested is unknown. For example, in platform trials, there may be only a low number of hypotheses to be tested. But since new treatments will be tested over time, online control of the FWER might be required [28]. In another approach, that was especially constructed for the modification of machine learning algorithms, the considered online error rate is only controlled over some time window [6]. In this way, significance level is gained back when hypotheses leave the window, which also makes it reasonable to consider conservative error measures, such as the FWER, in the online setting. A similar approach was considered in [25], where the past is ignored in a smooth manner.

Furthermore, there are approaches to use the closure principle to control other error rates than FWER, such as false discovery proportion (FDP) tail probabilities [14]. Various types of error rates fall under FDP, such as k-FWER and false discovery exceedance (FDX), and it can even be shown that any admissible FDP procedure must also be a closed procedure [12]. In Appendix A, we show that our approach can trivially be extended to obtain an online closure principle for FDP control. However, it is unclear whether the admissibility results derived in [12] and the shortcuts derived in [13] still apply in this case. This could be addressed in future work.

There also exist connections between FDR and closed testing. In [27], they introduced an approach to use graphical procedures for FDR control, and in [13], a connection between Simes-based closed testing and the Benjamini–Hochberg procedure [1] was shown. In addition, every FDR controlling procedure provides weak FWER control and thereby defines an α-level intersection test. Hence, all procedures that were constructed for FDR control can be used to derive new closed testing procedures with FWER or FDP control. This is especially interesting in the online case, as the literature is more advanced for FDR control than for FWER and FDP control.

APPENDIX A: ONLINE CLOSURE PRINCIPLE FOR FDP CONTROL

The *false discovery proportion* (FDP) for a some $S \in 2^{\mathbb{N}_f}$, where we denote by $2^{\mathbb{N}_f}$ the set of all finite subsets of $\mathbb{N}$, is defined as

$$\mathrm{FDP}_{\mathbb{P}}(S) = \frac{|S \cap I_0^{\mathbb{P}}|}{|S| \vee 1} \quad (\mathbb{P} \in \mathcal{P}).$$

Note that we focus on finite $S \subseteq \mathbb{N}$. On the one hand, $\mathrm{FDP}_{\mathbb{P}}(S)$ is not well-defined for infinite $S \subseteq \mathbb{N}$. On the other hand, from a practical point of view, in most applications we will not have an infinite number of hypotheses at hand, as the infinite testing process is only assumed because we do not know how many hypotheses are to be tested in the future. Therefore, we will only be interested in $\mathrm{FDP}_{\mathbb{P}}(S)$ for finite S. Following the notation in [12], we are searching for some random function $\boldsymbol{q} : 2^{\mathbb{N}_f} \to [0, 1]$ such that for all $\mathbb{P} \in \mathcal{P}$:

$$\mathbb{P}\big(\boldsymbol{q}(S) \geq \mathrm{FDP}_{\mathbb{P}}(S) \text{ for all } S \in 2^{\mathbb{N}_f}\big) \geq 1 - \alpha.$$

Providing an upper bound for $\mathrm{FDP}(S)$ is equivalent to providing a lower bound for the number of true discoveries $|S \cap I_1|$ [12]. Since the number of true discoveries is easier to handle, we

are focusing on it in the following. A procedure with *true discovery guarantee* is a random function $\boldsymbol{d} : 2^{\mathbb{N}_f} \to \mathbb{R}$ such that for all $\mathbb{P} \in \mathcal{P}$:

$$\mathbb{P}\big(\boldsymbol{d}(S) \leq |S \cap I_1^{\mathbb{P}}| \text{ for all } S \in 2^{\mathbb{N}_f}\big) \geq 1 - \alpha.$$

Furthermore, we call $\boldsymbol{d}$ *online true discovery procedure*, if $\boldsymbol{d}(S)$ is measurable with respect to $\mathcal{F}_{\max(S)}$ for all $S \in 2^{\mathbb{N}_f}$. Thus, the idea is that $\boldsymbol{d}(S)$ must be fixed as soon as we can decide on all the individual hypotheses H_i with $i \in S$. The following theorem is based on the results in [11, 12, 14].

THEOREM A.1 (Online closure principle for true discovery control). *Let a family of α-level intersection tests $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$ be given. Then the closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$, where $\boldsymbol{d}^{\boldsymbol{\phi}}(S) := \min\{|S \setminus I| : I \subseteq \mathbb{N}, \phi_I = 0\}$ for all $S \in 2^{\mathbb{N}_f}$, has true discovery guarantee at level α. In addition, if each ϕ_I is an online intersection test and the family of online intersection tests $\boldsymbol{\phi}$ is predictable (Definition 3.1), then $\boldsymbol{d}^{\boldsymbol{\phi}}$ is an online procedure.*

PROOF. Let $\mathbb{P} \in \mathcal{P}$ be arbitrary. For showing true discovery control of $\boldsymbol{d}^{\boldsymbol{\phi}}$, note that $\boldsymbol{d}^{\boldsymbol{\phi}}(S) > |S \cap I_1^{\mathbb{P}}|$ implies that $\phi_I = 1$ for all $I \subseteq \mathbb{N}$ with $|S \setminus I| \leq |S \cap I_1^{\mathbb{P}}|$. Since $|S \setminus I_0^{\mathbb{P}}| = |S \cap I_1^{\mathbb{P}}|$, we especially have $\phi_{I_0^{\mathbb{P}}} = 1$. However, this happens with probability at most α. Thus, $\mathbb{P}(\boldsymbol{d}^{\boldsymbol{\phi}}(S) \leq |S \cap I_1^{\mathbb{P}}| \text{ for all } S \subseteq I) \geq 1 - \alpha$. Furthermore, analogously to Lemma 3.3, we can show that due to the predictability of $\boldsymbol{\phi}$, it holds $\boldsymbol{d}^{\boldsymbol{\phi}}(S) = \min\{|S \setminus I| : I \subseteq \{1, \ldots, \max(S)\}, \phi_I = 0\}$. Since all ϕ_I are online intersection tests, $\boldsymbol{d}^{\boldsymbol{\phi}}(S)$ is measurable with respect to $\mathcal{F}_{\max(S)}$. □

APPENDIX B: ADDITIONAL SIMULATION RESULTS

In this section, we provide additional simulation results based on the design described in Section 5.1. We applied closed ADDIS-Spending and ADDIS-Spending with the same parameters as in Section 5.2. Figure 5 shows the results for conservative p-values ($\mu_N = -2$). In Figure 6 and Figure 7, we reduced the number of hypotheses to $n = 100$ and the strength of the alternative to $\mu_A = 3$, respectively, and in Figure 8 we reduced these parameters simultaneously. When we considered the lower number of $n = 100$ hypotheses, we also reduced the batch sizes to $b \in \{1, 5, 10, 25\}$. All plots show the similar behavior that closed ADDIS-Spending is less sensitive than ADDIS-Spending to an increase of the batch size, and thus to locally dependent p-values.

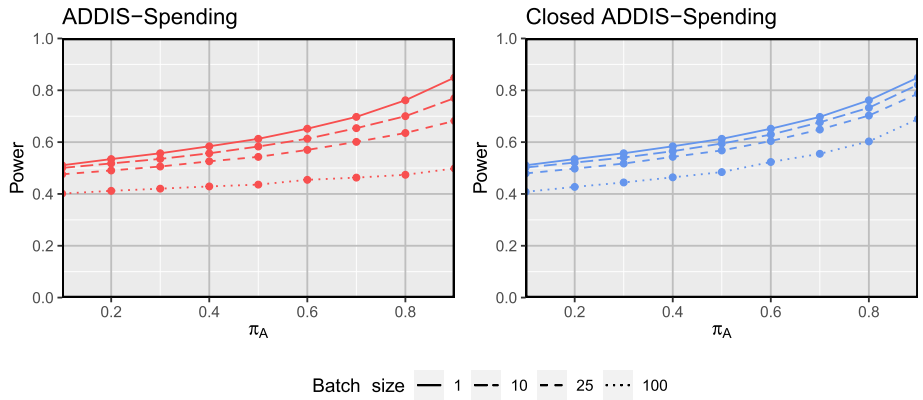

FIG. 5. *Comparison of ADDIS-Spending and closed ADDIS-Spending in terms of power based on locally dependent p-values for different batch sizes and proportions of false null hypotheses (π_A); $n = 1000$, $\mu_N = -2$, $\mu_A = 4$ and $\rho = 0.8$ in both plots.*

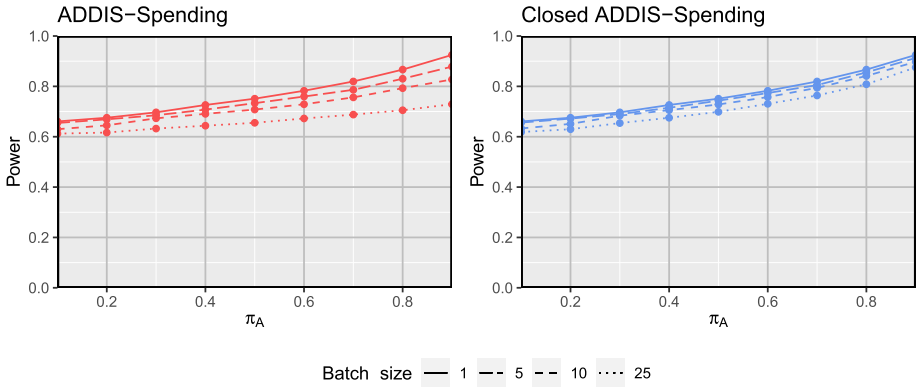

FIG. 6. *Comparison of ADDIS-Spending and closed ADDIS-Spending in terms of power based on locally dependent p-values for different batch sizes and proportions of false null hypotheses* (π_A); $n = 100$, $\mu_N = 0$, $\mu_A = 4$ *and* $\rho = 0.8$ *in both plots.*

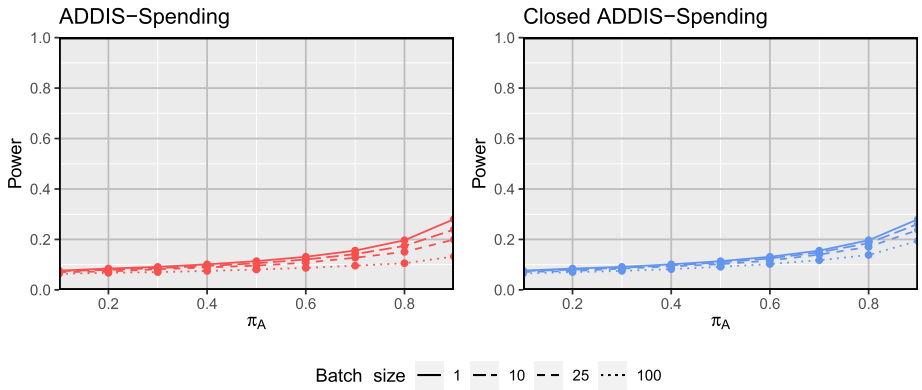

FIG. 7. *Comparison of ADDIS-Spending and closed ADDIS-Spending in terms of power based on locally dependent p-values for different batch sizes and proportions of false null hypotheses* (π_A); $n = 1000$, $\mu_N = 0$, $\mu_A = 3$ *and* $\rho = 0.8$ *in both plots.*

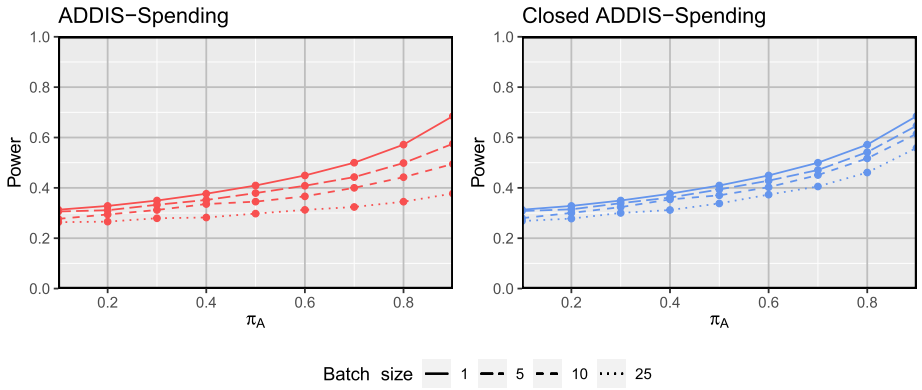

FIG. 8. *Comparison of ADDIS-Spending and closed ADDIS-Spending in terms of power based on locally dependent p-values for different batch sizes and proportions of false null hypotheses* (π_A); $n = 100$, $\mu_N = 0$, $\mu_A = 3$ *and* $\rho = 0.8$ *in both plots.*

APPENDIX C: OMITTED PROOFS

PROOF OF THEOREM 3.9. We show by induction that $d_i^{\boldsymbol{\phi}} = d_i^{\boldsymbol{\phi},s}$ for all $i \in \mathbb{N}$.

Initial case ($i = 1$): The predictability of $\boldsymbol{\phi}$ and Lemma 3.3 immediately implies that $d_1^{\boldsymbol{\phi}} = 1$ if and only if $d_1^{\boldsymbol{\phi},s} = \phi_{I_1} = 1$.

Induction Hypothesis (IH): We assume that $d_j^{\phi} = d_j^{\phi,s} = \phi_{I_j}$ for all $j \leq i-1$, where $i \geq 2$ is arbitrary but fixed.

Induction step ($i-1 \to i$): "$\leq$" Since $i \in I_i$, $d_i^{\boldsymbol{\phi}} = 1$ immediately implies $d_i^{\boldsymbol{\phi},s} = \phi_{I_i} = 1$. "$\geq$" Assume $d_i^{\boldsymbol{\phi},s} = \phi_{I_i} = 1$ and consider an arbitrary subset $J \subseteq \{1,\dots,i\}$ with $i \in J$. The consonance property implies that there exists a $k \in I_i$ such that $\phi_K = 1$ for all $K \subseteq I_i$ with $k \in K$. Since $I_k \subseteq I_i$ for all $k \in I_i$ and $\phi_{I_k} = 0$ for all $k \in I_i \setminus \{i\}$, the index satisfying the consonance property has to be i. Thus, H_J can be rejected by ϕ_J if $J \subseteq I_i$. If $J \nsubseteq I_i$, the definition of I_i ensures that there exists a $j \in J$ with $j < i$ such that $\phi_{I_j} = 1$. The induction hypothesis then implies that H_j is rejected by $\boldsymbol{d}^{\boldsymbol{\phi}}$, and hence H_J is rejected by ϕ_J. Since $J \subseteq \{1,\dots,i\}$ was arbitrary, all H_J with $J \subseteq \{1,\dots,i\}$ and $i \in J$ can be rejected. Moreover, Lemma 3.3 implies that H_i is rejected by $\boldsymbol{d}^{\boldsymbol{\phi}}$. □

PROOF OF THEOREM 4.2. The theorem follows immediately by Theorem 3.9 and Lemma C.1 below. □

LEMMA C.1. *Assume $(\boldsymbol{\alpha}_I)_{I\subseteq\mathbb{N}}$, where $\boldsymbol{\alpha}_I = (\alpha_i^I)_{i\in I}$, is a predictable family of online sub α-adjustments and $(\phi_I)_{I\subseteq N}$ the online intersection tests defined by (3). Then it holds for all $i \in \mathbb{N}$ that $\phi_{I_i} = 1$, where I_i is defined in Theorem 3.9, if and only if $p_i \leq \alpha_i^{I_i}$.*

PROOF. "$\Rightarrow$" Assume $\phi_{I_i} = 1$. Then there exists a $j \in I_i$ such that $p_j \leq \alpha_j^{I_i}$. Since $I_i = I_j \cup \{k \in \mathbb{N} : j < k < i, \phi_{I_k} = 0\} \cup \{i\}$, the predictability of $(\boldsymbol{\alpha}_I)_{I\subseteq\mathbb{N}}$ ensures that $\alpha_j^{I_i} = \alpha_j^{I_j}$ and hence $p_j \leq \alpha_j^{I_j}$. Because $\phi_{I_k} = 0$ for all $k \in I_i \setminus \{i\}$, we have $j = i$, meaning $p_i \leq \alpha_i^{I_i}$. "$\Leftarrow$" $p_i \leq \alpha_i^{I_i}$ implies $\phi_{I_i} = 1$ by definition. □

Acknowledgments. The authors are grateful for the valuable comments of two anonymous referees and an Associate Editor, which have led to significant improvements of the paper. In addition, the authors would like to thank Jelle Goeman for a useful discussion.

Funding. L. Fischer acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project number 281474342/GRK2224/2.

M. Bofill Roig is a member of the EU Patient-centric clinical trial platform (EU-PEARL). EU-PEARL has received funding from the Innovative Medicines Initiative 2 Joint Undertaking 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 nonprofit organization, Spring-works 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.

SUPPLEMENTARY MATERIAL

R code (DOI: 10.1214/24-AOS2370SUPP; .zip).

REFERENCES

[1] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

[2] BITTMAN, R. M., ROMANO, J. P., VALLARINO, C. and WOLF, M. (2009). Optimal testing of multiple hypotheses with common effect direction. *Biometrika* **96** 399–410. MR2507151 https://doi.org/10.1093/biomet/asp006

[3] BRETZ, F., MAURER, W., BRANNATH, W. and POSCH, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Stat. Med.* **28** 586–604. MR2655732 https://doi.org/10.1002/sim.3495

[4] DMITRIENKO, A., OFFEN, W. W. and WESTFALL, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Stat. Med.* **22** 2387–2400.

[5] DÖHLER, S., MEAH, I. and ROQUAIN, E. (2024). Online multiple testing with super-uniformity reward. *Electron. J. Stat.* **18** 1293–1354. MR4718473 https://doi.org/10.1214/24-ejs2230

[6] FENG, J., EMERSON, S. and SIMON, N. (2021). Approval policies for modifications to machine learning-based software as a medical device: A study of bio-creep. *Biometrics* **77** 31–44. MR4229719 https://doi.org/10.1111/biom.13379

[7] FENG, J., PENNLLO, G., PETRICK, N., SAHINER, B., PIRRACCHIO, R. and GOSSMANN, A. (2022). Sequential algorithmic modification with test data reuse. In *Uncertainty in Artificial Intelligence* 674–684. PMLR.

[8] FISCHER, L., BOFILL ROIG, M. and BRANNATH, W. (2024). Supplement to "The online closure principle." https://doi.org/10.1214/24-AOS2370SUPP

[9] FOSTER, D. P. and STINE, R. A. (2008). α-investing: A procedure for sequential control of expected false discoveries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 429–444. MR2424761 https://doi.org/10.1111/j.1467-9868.2007.00643.x

[10] GABRIEL, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. *Ann. Math. Stat.* **40** 224–250. MR0240931 https://doi.org/10.1214/aoms/1177697819

[11] GENOVESE, C. R. and WASSERMAN, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.* **101** 1408–1417. MR2279468 https://doi.org/10.1198/016214506000000339

[12] GOEMAN, J. J., HEMERIK, J. and SOLARI, A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *Ann. Statist.* **49** 1218–1238. MR4255125 https://doi.org/10.1214/20-aos1999

[13] GOEMAN, J. J., MEIJER, R. J., KREBS, T. J. P. and SOLARI, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika* **106** 841–856. MR4046036 https://doi.org/10.1093/biomet/asz041

[14] GOEMAN, J. J. and SOLARI, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26** 584–597. MR2951390 https://doi.org/10.1214/11-STS356

[15] GRECHANOVSKY, E. and HOCHBERG, Y. (1999). Closed procedures are better and often admit a shortcut. *J. Statist. Plann. Inference* **76** 79–91. MR1673341 https://doi.org/10.1016/S0378-3758(98)00125-6

[16] HOMMEL, G., BRETZ, F. and MAURER, W. (2007). Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Stat. Med.* **26** 4063–4073. MR2405792 https://doi.org/10.1002/sim.2873

[17] IOANNIDIS, J. P. (2005). Why most published research findings are false. *PLoS Med.* **2** e124.

[18] JAVANMARD, A. and MONTANARI, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *Ann. Statist.* **46** 526–554. MR3782376 https://doi.org/10.1214/17-AOS1559

[19] KARP, N. A., MASON, J., BEAUDET, A. L., BENJAMINI, Y., BOWER, L., BRAUN, R. E., BROWN, S. D., CHESLER, E. J., DICKINSON, M. E. et al. (2017). Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nat. Commun.* **8** 1–12.

[20] KOHAVI, R., DENG, A., FRASCA, B., WALKER, T., XU, Y. and POHLMANN, N. (2013). Online controlled experiments at large scale. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1168–1176.

[21] LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

[22] MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056 https://doi.org/10.1093/biomet/63.3.655

[23] MAURER, W., HOTHORN, L. and LEHRMACHER, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: A-priori ordered hypotheses. In *Biometrie in der Chemisch-Pharmazeutischen Industrie* (V. Joachim, ed.) 3–18. Fischer Verlag, Stuttgart.

[24] MUÑOZ-FUENTES, V., CACHEIRO, P., MEEHAN, T. F., AGUILAR-PIMENTEL, J. A., BROWN, S. D., FLENNIKEN, A. M., FLICEK, P., GALLI, A., MASHHADI, H. H. et al. (2018). The International Mouse Phenotyping Consortium (IMPC): A functional catalogue of the mammalian genome that informs conservation. *Conserv. Genet.* **19** 995–1005.

[25] RAMDAS, A., YANG, F., WAINWRIGHT, M. J. and JORDAN, M. I. (2017). Online control of the false discovery rate with decaying memory. In *Advances in Neural Information Processing Systems* **30**. Curran Associates, Red Hook.

[26] RAMDAS, A., ZRNIC, T., WAINWRIGHT, M. and JORDAN, M. (2018). SAFFRON: An adaptive algorithm for online control of the false discovery rate. In *International Conference on Machine Learning* 4286–4294. PMLR.

[27] ROBERTSON, D. S., WASON, J. M. S. and BRETZ, F. (2020). Graphical approaches for the control of generalized error rates. *Stat. Med.* **39** 3135–3155. MR4151924 https://doi.org/10.1002/sim.8595

[28] ROBERTSON, D. S., WASON, J. M. S., KÖNIG, F., POSCH, M. and JAKI, T. (2023). Online error rate control for platform trials. *Stat. Med.* **42** 2475–2495. MR4596806 https://doi.org/10.1002/sim.9733

[29] ROBERTSON, D. S., WASON, J. M. S. and RAMDAS, A. (2023). Online multiple hypothesis testing. *Statist. Sci.* **38** 557–575. MR4665026 https://doi.org/10.1214/23-sts901

[30] ROBERTSON, D. S., WILDENHAIN, J., JAVANMARD, A. and KARP, N. A. (2019). onlineFDR: An R package to control the false discovery rate for growing data repositories. *Bioinformatics* **35** 4196–4199.

[31] ROMANO, J. P., SHAIKH, A. and WOLF, M. (2011). Consonance and the closure method in multiple testing. *Int. J. Biostat.* **7** 12. MR2775079 https://doi.org/10.2202/1557-4679.1300

[32] ROMANO, J. P. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821 https://doi.org/10.1198/016214504000000539

[33] SANDERCOCK, P. A., DARBYSHIRE, J., DEMETS, D., FOWLER, R., LALLOO, D. G., MUNAVVAR, M., STAPLIN, N., WARRIS, A., WITTES, J. et al. (2022). Experiences of the data monitoring committee for the RECOVERY trial, a large-scale adaptive platform randomised trial of treatments for patients hospitalised with COVID-19. *Trials* **23** 881.

[34] SCHWEDER, T. and SPJØTVOLL, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69** 493–502.

[35] SONNEMANN, E. and FINNER, H. (1988). Vollständigkeitssätze für multiple Testprobleme. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing* (P. Bauer, G. Hommel and E. Sonnemann, eds.) 121–135. Springer, Berlin.

[36] TIAN, J. and RAMDAS, A. (2019). ADDIS: An adaptive discarding algorithm for online FDR control with conservative nulls. In *Advances in Neural Information Processing Systems* **32**. Curran Associates, Red Hook.

[37] TIAN, J. and RAMDAS, A. (2021). Online control of the familywise error rate. *Stat. Methods Med. Res.* **30** 976–993. MR4259882 https://doi.org/10.1177/0962280220983381

[38] ZEHETMAYER, S., POSCH, M. and KOENIG, F. (2022). Online control of the False Discovery Rate in group-sequential platform trials. *Stat. Methods Med. Res.* **31** 2470–2485. MR4513312 https://doi.org/10.1177/09622802221129051

[39] ZHAO, Q., SMALL, D. S. and SU, W. (2019). Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *J. Amer. Statist. Assoc.* **114** 1291–1304. MR4011780 https://doi.org/10.1080/01621459.2018.1497499

[40] ZRNIC, T., JIANG, D., RAMDAS, A. and JORDAN, M. (2020). The power of batching in multiple hypothesis testing. In *International Conference on Artificial Intelligence and Statistics* **108** 3806–3815. PMLR.

[41] ZRNIC, T., RAMDAS, A. and JORDAN, M. I. (2021). Asynchronous online testing of multiple hypotheses. *J. Mach. Learn. Res.* **22** 33. MR4253726 https://doi.org/10.1515/ijnsns-2019-0210