Adaptive Identification of Populations with Treatment Benefit in Clinical Trials: Machine Learning Challenges and Solutions

Alicia Curth [1] **Alihan Hüyük** [1] **Mihaela van der Schaar** [1 2]

Abstract

We study the problem of adaptively identifying patient subpopulations that benefit from a given treatment during a confirmatory clinical trial. This type of adaptive clinical trial has been thoroughly studied in biostatistics, but has been allowed only limited adaptivity so far. Here, we aim to relax classical restrictions on such designs and investigate how to incorporate ideas from the recent machine learning literature on adaptive and online experimentation to make trials more flexible and efficient. We find that the unique characteristics of the subpopulation selection problem – most importantly that (i) one is usually interested in finding subpopulations with *any* treatment benefit (and not necessarily the single subgroup with largest effect) given a limited budget and that (ii) effectiveness only has to be demonstrated across the subpopulation *on average* – give rise to interesting challenges and new desiderata when designing algorithmic solutions. Building on these findings, we propose AdaGGI and AdaGCPI, two meta-algorithms for subpopulation construction. We empirically investigate their performance across a range of simulation scenarios and derive insights into their (dis)advantages across different settings.

1. Introduction

The existence of treatment effect heterogeneity across subgroups of patients poses a challenge to both the success of clinical trials testing the effectiveness of treatments *and* the quality of treatment decisions in clinical practice when prescribing a drug that has been proven to be effective only for the average population [1–3]. Examples for such heterogeneity are ubiquituous in practice and include differences in treatment responses in cancer patients with specific mutations [4], pyschiatric patients with different forms of depression [5] and stroke patients [6]. Motivated by this, the problem of discovering treatment effect heterogeneity using *logged* experimental or observational data has received much attention in the recent machine learning (ML) literature [7], resulting in the adaptation of many supervised ML methods for post-hoc effect estimation [8–12]. The *active* counterpart to this problem, i.e. designing experiments (clinical trials) to actively discover subpopulations that respond well to a treatment, has received only limited attention in the ML literature thus far but is the focus of this paper.

The biostatistics literature on adaptive clinical trials, on the other hand, has proposed and extensively studied the use of so-called *adaptive enrichment designs*, which allow to change both enrolment criteria and the null hypothesis to be tested in a clinical trial based on interim data (see e.g. [1, 2] and Appendix A.1 for an overview). In such designs, the degree of adaptivity and flexibility is usually quite limited as the ability to adapt features is commonly restricted to a few pre-specified interim analysis points and the number of subgroups is often very small (most often set to exactly two).

In this paper, we consider a new approach to designing such adaptive enrichment trials and investigate whether and how it is possible to make them more flexible and efficient by adapting tools that were originally developed to solve pure exploration[1] multi-armed bandits [13] and other adaptive experiments problems in the recent ML literature. We find that the problem of constructing subpopulations from subgroups in which a treatment has *any positive* effect most closely resembles the *good* arm identification (or thresholding bandit) problem studied in e.g. [14–19] as there is no need to limit treatment prescription to the single subgroup with largest effect [20]. Nonetheless, we argue that there are additional unique characteristics of our problem that may change how algorithmic solutions should be designed: (i) clinical trials operate under constraints on *both* budget

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK [2]The Alan Turing Institute. Correspondence to: Alicia Curth <amc253@cam.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

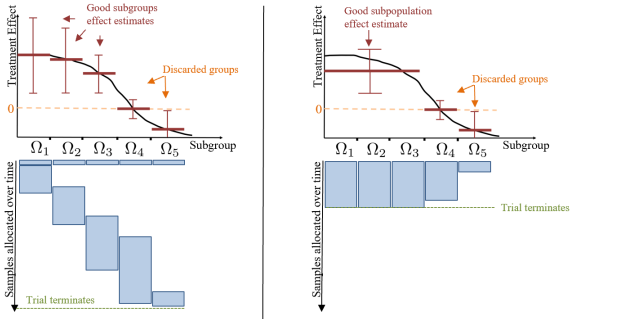
[1]Note that, as further discussed in Appendix A.2, we focus on adapting ideas from the literature on *purely explorative* bandits, which *do not* trade off exploration with exploitation and are thus very different from prototypical explore-exploit bandit problems.

and confidence, (ii) budget is *very limited* compared to e.g. online advertising settings, (iii) effectiveness only has to be demonstrated across a subpopulation *on average* and (iv) required control of false discovery and power is stricter and more nuanced. Note that solutions for problems with some of these characteristics could be of independent interest in applications beyond the clinical trial context: e.g. (i) and (ii) may appear whenever one is looking to find *any (single) good* candidate, solution or arm with high confidence as fast as possible, while (iii) appears when one only needs to identify *a collection of arms* that works well on average.

Contributions. We study the problem of adaptive identification of patient subpopulations that benefit from a treatment during a clinical trial through a ML lens. Note that our focus in this paper lies not primarily in developing novel ML methodology, but rather *in formalizing and understanding our clinical trial problem and its inherent challenges as a novel ML problem*, then allowing us to explore how to best adapt existing solutions to our setting. In doing so, we hope to introduce relevant ML communities to a new application, through showcasing that this area is full of new ML problems, demanding constraints and interesting methodological challenges. We make three main contributions:
(1) Problem formalization and understanding: We focus on *formalising*, *contextualizing* and *understanding* the population identification problem and its inherent challenges *as a ML problem*. We discover two possible formulations of the problem which differ in terms of their characteristics and investigate how these give rise to different desiderata when designing algorithmic solutions. **(2) Two new meta-algorithms:** Building on these insights and ideas from the ML literature on adaptive experiments, we then propose a solution in form of a meta-algorithm for each scenario (see Fig. 1).**(3) Empirical Insight:** We empirically investigate and provide insight into the (dis)advantages of either formulation and their solution through a range of simulation studies. Albeit not our primary objective, we believe that some of these empirical insights could be of independent interest to researchers studying the problem of good *arm* identification in a small sample regime.

2. Problem Setup

Throughout, we adopt problem setting and notation similar to [3]. Thus, we wish to run a clinical trial to establish efficacy of a novel drug (T) relative to an established control (C) in patient population Ω_0. We assume further that Ω_0 is made up of K disjoint and prespecified subgroups $\Omega_1, \dots, \Omega_K$ where $\Omega_0 = \cup_{j \leq K} \Omega_j$, across which efficacy may be expected to differ, e.g. due to known biological pathways or evidence from earlier trials. Let θ_j denote the treatment's effect (relative to control) within subgroup j, and let π_j denote the prevalence of subgroup j in the population.



(A) Adaptive Good subGroup Identification (B) Adaptive Good Composite subPopulation Identification

Figure 1. Overview of the two problem formulations and proposed solutions. (A) The adaptive good subgroup identification (AdaGGI) algorithm finds individual subgroups with treatment benefit through successive discovery. (B) The adaptive good composite subpopulation (AdaGCPI) algorithm finds a composite subpopulation by successively removing the subgroup with smallest effect until a positive average treatment effect is discovered.

Goal. To ensure success of the clinical trial, we aim to adaptively construct a *composite subpopulation* composed of a subset $\mathcal{S} \subseteq \mathcal{K} = \{1, \dots, K\}$ of the full population with $\Omega_{\mathcal{S}} = \cup_{i \in \mathcal{S}} \Omega_i$, in which the treatment is *effective* on average (if any exists); we refer to such subpopulations as *good*:

Definition 2.1. (Good Subpopulation) A subpopulation $\mathcal{S} \subseteq \mathcal{K}$ is a *good subpopulation* iff $\theta_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \frac{\pi_i}{\sum_{j \in \mathcal{S}} \pi_j} \theta_i > 0$.

Generally, to maximise patient benefit, we would like to identify the *largest* subpopulations in which the treatment is effective – i.e. if $\theta_i > \theta_j > 0$, we prefer $\mathcal{S}^{ij} = \{i, j\}$ over $\mathcal{S}^i = \{i\}$ even though $\theta_{S^{ij}} < \theta_{S^i}$.

Null hypotheses and problem types. We consider a null scenario of no treatment effect, i.e. $\theta_0 = 0$, giving rise to two types of problems and associated null hypotheses. In Sec. 3, we first identify individual good *subgroups*, i.e. find subgroups for which we can reject the null hypothesis

$$H_{0j} : \theta_j = 0 \tag{1}$$

for the one-sided alternative $H_{aj} : \theta_j > 0$. Clearly, when composing a subpopulation by including only subgroups in which the individual null hypotheses have been rejected, i.e. $\mathcal{S}^a = \{j : \theta_j > 0\}$, the subpopulation as a whole will have positive effect too, i.e. $\theta_{\mathcal{S}^a} > 0$. We refer to this problem as the *Good subGroup Identification* (GGI) problem.

Often, clinical trials are not powered to detect effects in subgroups seperately; instead (when more than two subgroups are considered), the focus is set on demonstrating *average* effectiveness across a sub*population* as in [3]. We therefore consider a second setting in Sec. 4: here, we wish to identify a *composite* subpopulation $\mathcal{S}$ for which we can prove that the treatment is effective *on average*, i.e. reject

$$H_{0\mathcal{S}} : \theta_{\mathcal{S}} = 0 \tag{2}$$

for the one-sided alternative $H_{a\mathcal{S}} : \theta_{\mathcal{S}} > 0$. We refer to this problem as the *Good Composite subPopulation Identification* (GCPI) problem. Note that the underlying requirement is strictly weaker than in the GGI problem as rejecting $H_{0\mathcal{S}}$ does not require rejecting H_{0j} for every $j \in \mathcal{S}$.

Familywise error rate control. Regulatory agencies usually require the familywise error rate (FWER), i.e. the probability of a Type 1 error, to be controlled in clinical trials [21]. Formally, the FWER of an algorithm $\mathcal{A}$ for the set of problem instances $\mathcal{P}$ under consideration is defined as

$$\text{FWER}(\mathcal{A};\mathcal{P}) = \sup_{\rho\in\mathcal{P}} \mathbb{P}_\rho(\mathcal{A} \text{ rejects a true null hypothesis})$$

and FWER-control at the level of $\alpha \in (0,1)$ requires that $\text{FWER}(\mathcal{A};\mathcal{P}) \leq \alpha$. Further, we can write

$$\text{FWER}_{GGI}(\mathcal{A};\mathcal{P}) \leq \sup_{\rho\in\mathcal{P}} \sum_{j=1}^{K} \mathbb{P}_\rho(\mathcal{A} \text{ rejects true } H_{0j})$$
$$\text{FWER}_{GCPI}(\mathcal{A};\mathcal{P}) \leq \sup_{\rho\in\mathcal{P}} \sum_{\mathcal{S}\subseteq\mathcal{K}} \mathbb{P}_\rho(\mathcal{A} \text{ selects } \mathcal{S} \text{ and rejects true } H_{0\mathcal{S}})$$

Minimum relevant effect. Clinical trials also aim to avoid Type 2 error (failure to detect a true positive effect). As the sample size needed to differentiate $\theta_0 = 0$ from $\theta_j > 0$ scales as θ_j^{-2}, trials often introduce a *minimum clinically relevant difference* $\theta_{min} > \theta_0 = 0$ which a trial should be powered to detect [22]. Thus, while not a hard requirement like FWER control, we aim to ensure that $\mathbb{P}(H_{0\mathcal{S}'} \text{ is not rejected } | \theta_{\mathcal{S}} = \theta_{min}) \leq \beta$ for at least some $\mathcal{S}' \subset \mathcal{S}$.

Mode of environment interaction and data structure. Throughout, we assume the stylized setting of an unlimited stream of patients available for recruitment from each subgroup, where outcomes are revealed to the algorithm immediately; we discuss possible extensions to more realistic scenarios in Appendix B. That is, at every time step $t \in \{1,\dots,B\}$, where $2B$ is the total patient budget of the trial, the algorithm selects a subgroup $J_t \in \mathcal{K}$ to enrol two patients from, which are then *randomly* assigned to one of each treatment and control arm. This gives rise to control and treated outcome $Y_t^C, Y_t^T \in \mathcal{Y}$, which could be continuous ($\mathcal{Y} = \mathbb{R}$) or binary ($\mathcal{Y} = \{0,1\}$), and produces a dataset of tuples $\mathcal{D}_t = \{(J_{t'}, Y_{t'}^C, Y_{t'}^T\}_{t'\leq t}$. We denote by $N_i(t) = \sum_{t'\leq t} \mathbb{1}\{J_{t'} = i\}$ and $N_{\mathcal{S}}(t) = \sum_{t'\leq t} \mathbb{1}\{J_{t'} \in \mathcal{S}\}$ the number of patient pairs enrolled from a subgroup or a subpopulation by time t, respectively.

Estimators & Inference. Given randomization and assuming *no interference* between patients, we have that $\theta_j = \mathbb{E}[Y_t^T - Y_t^C | J_t = j]$, so that we can estimate treatment effects simply as

$$\hat{\theta}_{j,N_j(t)} = \frac{\sum_{t'=1}^{t} \mathbb{1}\{J_{t'}=j\}(Y_{t'}^T - Y_{t'}^C)}{N_j(t)} \tag{3}$$

Whenever all subgroups i in a subpopulation $\mathcal{S}$ were drawn according to their relative prevalence $\frac{\pi_i}{\sum_{j\in\mathcal{S}}\pi_j}$, we can also estimate $\hat{\theta}_{\mathcal{S},N_{\mathcal{S}}(t)} = \frac{\sum_{t'=1}^{t} \mathbb{1}\{S_{t'}\in\mathcal{S}\}(Y_{t'}^T - Y_{t'}^C)}{N_{\mathcal{S}}(t)}$. Note that the $\hat{\theta}_{j,N_j(t)}$ will generally not be unbiased for θ_j as the J_t were selected in a data-adaptive manner (see e.g. [23, 24]).

Finally, standard approaches to statistical inference will generally not be valid when experiments are stopped adaptively, and we need to account for possible bias due to continuous monitoring of experiments. To retain the ability to perform valid inference, we therefore also assume that we have access to some always-valid confidence intervals [25]; that is, similar to [26] we rely on existence of some function $\phi(t,\delta)$ which satisfies for any $\delta \in (0,1)$ that $\mathbb{P}(\cap_{t=1}^{\infty}\{|\hat{\theta}_{\mathcal{S},t} - \theta_{\mathcal{S}}| \leq \phi(t,\delta)\}) \geq 1-\delta$. Our proposed meta-algorithms allow the use of any user-specified function $\phi(t,\delta)$. As discussed further in Appendix C, we follow [26] in our experiments and instantiate it using Thm. 8 of [27] which shows that for mean-zero σ_p^2-(sub)gaussian variables X_s, $\mathbb{P}(\exists t \in \mathbb{N} : \frac{\sum_{s=1}^{t} X_s}{t} > \sqrt{\frac{2\sigma_p^2\zeta(t,\delta)}{t}}) \leq \delta$ for $\zeta(t,\delta) = \log(1/\delta) + 3\log\log(1/\delta) + (3/2)\log\log(et/2)$ and $\delta \leq 0.1$. We can use $\sqrt{\frac{2\sigma_p^2\zeta(t,\delta)}{t}}$ as $\phi(\cdot,\cdot)$ in our experiments due to the fact that (i) the difference between two σ^2-(sub)gaussian variables is $2\sigma^2$-(sub)gaussian and (ii) Bernoulli variables are $\frac{1}{4}$-subgaussian. Note that generally subgaussianity is satisfied by e.g. *bounded* (and centered) outcomes Y. Given that many medical outcomes and lab tests have bounded credible ranges, we therefore consider subgaussianity of outcomes to be a very reasonable assumption in this context.

3. Good Subgroup Identification

We begin by studying the good subgroup identification (GGI) problem as it appears more closely related to problems studied in the recent ML literature. Recall that the GGI problem focusses on finding members of the set $\mathcal{H}_a = \{j : \theta_j > 0\}$, subject to FWER-$\alpha$-control and budget $2B$.

Related work. If θ_j was the *mean of a bandit arm* (instead of a subgroup treatment effect), GGI resembles problems that have been studied in the pure exploration literature as *thresholding* bandit [14–17], *good arm* identification (GAI) [18, 19] and hypothesis testing using bandits [26, 28].[2] In addition to the difference in target of interest, a major difference between existing formulations and our problem are the constraints placed on an ideal solution. Unlike our problem, classical pure exploration problems usually operate *either* under a fixed budget *or* a fixed confidence constraint: For example, in [14]'s thresholding bandit, which aims to classify *all* arms as above or below a threshold, the fixed confidence setting requires *all* classifications

[2] More typical exploration problems, e.g. best arm identification (e.g. [29–31]) are less relevant as our interest lies no in finding the group with *the best* response to a drug [20]; see also Appendix A.

(both above and below the threshold) to be correct with fixed confidence δ, while the fixed budget setting aims for the highest confidence in all classifications given a certain budget. All of [18, 19, 26, 28] study a similar fixed confidence setting. Finally, [32] is the only ML work we are aware of that studies good subgroup discovery in a clinical trial context – they propose a Bayesian MDP-based design optimizing patient recruitment given a fixed budget but do not control Type I error rate of discoveries, which conceptually resembles a fixed-budget-only GAI setup.

3.1. Problem Characteristics and Design Considerations in the GGI Problem

Unique characteristics of the GGI objective. Discovery in a clinical trial is usually subject to *both* a budget *and* FWER constraint (i.e. a fixed confidence constraint on *each* discovery). Thus, instead of identifying *all* good arms either under a fixed budget while maximising confidence as in [14] or with fixed confidence while minimizing budget as in [18, 26], we aim to maximise the number of arms that can be discovered with fixed confidence given a budget – which is a combination of the fixed confidence and fixed budget setting that are usually considered separately. Additionally, the available budget is usually *very limited* in clinical trials relative to e.g. online advertising applications commonly considered in the bandit literature. Due to both ethical and financial considerations, clinical trials usually operate in small sample regimes – confirmatory phase 3 sample sizes usually lie between 300-3000 patients [33], which is orders of magnitude smaller than sample sizes considered in the ML literature. Finally, the distinction (or asymetry) between both confidence α and power $1-\beta$, and null threshold θ_0 and minimum relevant effect θ_{min} is usually not found in e.g. GAI problems.

Design considerations. The unique characteristics of the GGI objective give rise to a number of desiderata while designing algorithms: First, there is a *need to focus on promising groups*, as budget is limited and to meet our objective it is *not* necessary to make a judgement about *all* subgroups immediately. Thus we should focus our attention on subgroups that look promising and leave subgroups with effects that are hard to distinguish from the null for last (this is the opposite strategy to thresholding bandit solutions [14–16] that focus explicitly on the arms that are hardest to identify[3]). Second, we may wish to *limit the degree of exploration* and explore only so long until a promising good subgroup has been identified (this is unlike a *best* arm identification problem where *relative* quality of an arm matters which needs substantially more exploration to identify).

[3]If identifying *all* good groups is desired, it matters how fast the last (most difficult) group is found; while our goal to identify *many good groups quickly* necessitates early focus on 'easier' ones.

Algorithm 1 AdaGGI

Require: $\alpha, \beta \in (0,1)$, $\theta_{min} > 0$, budget B, initial samples n_0, sampling rule $\mathcal{E}$, identification rule $\mathcal{I}$, removal rule $\mathcal{R}$
1: Initialise: $\mathcal{A}_{Kn_0} = \mathcal{K}$; $\forall j \in \mathcal{K}$, sample n_0 times, set $D_{Kn_0} = \{(S_{t'}, Y_{t'}^C, Y_{t'}^T\}_{t' \leq Kn_0}$
2: **for** $t \in \{Kn_0 + 1, B\}$ **do**
3: Choose subgroup $J_t = \mathcal{E}(D_{t-1}, \mathcal{A}_{t-1})$ to enrol, set $\mathcal{D}_t = \mathcal{D}_{t-1} \cup (S_t, Y_t^C, Y_t^T)$
4: Identify good subgroups $\mathcal{S}_t = \mathcal{S}_{t-1} \cup \mathcal{I}(\mathcal{D}_t, \alpha)$, set $\mathcal{A}_t = \mathcal{K} \setminus \mathcal{S}_t$
5: Remove bad groups: $\mathcal{A}_t = \mathcal{K} \setminus \mathcal{R}(\mathcal{D}_t, \theta_{min}, \beta)$
6: If $\mathcal{A}_t = \emptyset$, **Output:** `True` if $|\mathcal{S}_B|>0$, $\mathcal{S}_B$
7: **end for**
8: **Output:** `True` if $|\mathcal{S}_B|>0$, $\mathcal{S}_B$

Third, we may want to *focus on null hypotheses closest to rejection*, recognizing that for a successful trial, rejecting one null hypothesis at level α is better than having two hypotheses only close to rejection upon termination.

The backbones of the fixed confidence algorithms for identifying good bandit arms with mean above a threshold proposed in [18, 19, 26, 28] *do* lend themselves to be adapted to our combined fixed confidence - fixed budget setting: these algorithms sequentially move arms from the *active set* under exploration to a passive (output) set containing all good arms identified with fixed confidence thus far, and could in principle solve our fixed budget setting by simply stopping testing additional arms once the budget is reached. Below, we discuss this approach and our modifications in more detail.

3.2. AdaGGI: A Meta-algorithm for Good Subgroup Identification

We propose AdaGGI, an *Ada*ptive *G*ood sub*G*roup *I*dentification meta-algorithm, presented in Alg. 1. As described in detail below, each iteration consists of (i) choosing a subgroup J_t to enrol using an exploration (sampling) strategy $\mathcal{E}$, (ii) subsequently screening for new good subgroups using an α-dependent identification criterion $\mathcal{I}$ and (iii) removal of any groups demonstrating no minimum benefit using a (β, θ_{min})-dependent removal criterion $\mathcal{R}$.

Sampling strategies $\mathcal{E}$: Finding good arms fast. The established choice for sampling (exploration) strategy $\mathcal{E}$ in the GAI literature [18, 19, 26] appears to be to use an optimistic upper-confidence bound (UCB) approach, i.e.

$$\mathcal{E}_{UCB}(D_{t-1}, \mathcal{A}_{t-1}) = \arg \max_{j \in \mathcal{A}_{t-1}} \hat{\theta}_{j,N_j(t-1)} + \phi(N_j(t-1), \alpha)$$

However, this strategy does not necessarily *exploit* accumulated knowledge by repeatedly sampling a subgroup whose null is close to being rejected; in fact, as $\phi(t, \delta)$ shrinks with increasing t, we suspect that $\mathcal{E}_{UCB}$ may encourage frequent switching between subgroups when the effects in multiple good subgroups are similar which may

lead to no null being rejected when budget is very limited.

Therefore, we explore the use of two new sampling strategies for this problem. As we discuss below, identification using $\mathcal{I}(\cdot)$ will rely on the criterion $\mathbb{1}\{\hat{\theta}_{j,N_j(t)} - \phi(N_j(t), \epsilon) > 0\}$ for some $\epsilon \in (0, 1)$; therefore, sampling according to the best lower confidence bound (LCB) would correspond to selecting arms that appear most promising for early identification, i.e. be more *exploitative*. Thus, we also consider using

$$\mathcal{E}_{LCB}(D_{t-1}, \mathcal{A}_{t-1}) = \arg \max_{j \in \mathcal{A}_{t-1}} \hat{\theta}_{j,N_j(t-1)} - \phi(N_j(t-1), \alpha)$$

Because this strategy conversely may risk *getting stuck* on a subgroup which only *appeared* good early on, we consider a final strategy $\mathcal{E}_{LUCB}(D_{t-1}, \mathcal{A}_{t-1}) = \mathcal{E}_{LCB}(D_{t-1}, \mathcal{A}_{t-1}) \cup \mathcal{E}_{UCB}(D_{t-1}, \mathcal{A}_{t-1})$, allowing enrolment from two subgroups whenever sampling according to UCB and LCB disagree (thus t increases by 2).

Identification criterion: Ensuring FWER control. Our identification criterion needs to ensure that $FWER_{GGI} \leq \alpha$ by adjusting for the fact that we perform *multiple* hypothesis tests. As we consider only a moderate number of subgroups K, we rely on a simple Bonferroni correction here and use

$$\mathcal{I}^K_{BF}(\mathcal{D}_t, \alpha) = \{j \in \mathcal{K} : \hat{\theta}_{j,N_j(t)} - \phi(N_j(t), \frac{\alpha}{K}) > 0\}$$

which controls FWER as

$$\sum_{j \in \mathcal{K}: \theta_j = 0} \mathbb{P}(\cap_{t=1}^{\infty} \{\hat{\theta}_{j,t} - \theta_j > \phi(t, \frac{\alpha}{K})\} \leq K \frac{\alpha}{K}$$

To create tighter confidence bounds in settings where *many* null hypotheses are false and recycling α from previously rejected hypotheses is thus possible, one could implement more sophisticated strategies based on the adapted Benjamini-Hochberg procedure from [26], or other α-investing approaches such as those discussed in [34].

Removal criterion: Focusing on significant effects. We employ removal criterion

$$\mathcal{R}_{fut}(\mathcal{D}_t, \theta_{min}, \beta) = \{j \in \mathcal{K} : \hat{\theta}_{j,N_j(t)} + \phi(N_j(t), \beta) < \theta_{min}\}$$

This ensures that subgroups can be removed early for *futility* while power to detect a clinically relevant effect is preserved. Note that this ensures that the burden of proof to discard a bad subgroup can be much lower than what is needed to identify it as good. This differs from the recent GAI literature, where arms are either discarded and accepted using the same threshold/confidence [18] or not discarded at all [19, 26].

4. Good Composite Subpopulation Identification

Instead of finding good subgroups *separately* as before, we now move to the Good Composite subPopulation Identification (GCPI) problem which considers finding a good *composite* subpopulation directly, i.e. finding $\mathcal{S} \subseteq \mathcal{K}$ such that $\theta_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \frac{\pi_i}{\sum_{j \in \mathcal{S}} \pi_j} \theta_i > 0$. Intuitively, this should be *easier* to solve – i.e. we would expect a smaller sample size to be required for a trial to be successful: given $\mathcal{S}$, rejection of $H_{0\mathcal{S}}$ is a strictly weaker requirement than rejecting all constituent elementary null hypotheses separately and it should be possible to share statistical strength (i.e. exploit larger sample size) *across* subgroups contained in $\mathcal{S}$.

Related work. Most work from the adaptive enrichment clinical trial literature appears to solve a simplified version of the GCPI problem, where $\mathcal{K} = \{1, 2\}$ and initially patients from both subgroups are enrolled. At either a single (e.g. [2, 35, 36]) or multiple (e.g. [6, 37]) prespecified interim analysis points it is then possible to discontinue either subgroup, where decisions are usually based on precalculated (normal) stopping boundaries. The setting considered in [3] is most similar to our setup as no restrictions are placed on K: here, the choice of subgroups to include in the selected subpopulation $\mathcal{S}$ is *fixed* at the first interim analysis and all subsequent analyses allow only early termination of the *entire* subpopulation based on efficacy/futility error-spending boundaries which are calculated based on the assumption that all $\theta_j \geq 0$ (i.e. negative effects are not allowed). From a bandit perspective, the GCPI problem can be interpreted as a generic *combinatorial bandit* problem [38, 39], where each subpopulation could be seen as a *super-arm*; however, to the best of our knowledge no existing solutions exploit the idea of sharing statistical strength across arms by pooling samples and solutions derived from e.g. [38, 39] would therefore resemble our GGI solution.

4.1. Unique Problem Characteristics and Design Considerations in the GCPI Problem

Unique characteristics of the GCPI objective. Relative to GGI, we consider two additional features key to the GCPI problem: On the one hand, the weaker requirement of identification of a positive *average* effect should make it possible to share statistical strength *across* subgroups, which may make the problem *easier*. On the other hand, while the GGI problem has only K subgroups with associated hypotheses to consider, the subpopulation construction problem is *combinatorial* and there are 2^K possible subpopulations and null hypotheses, possibly making the problem *harder*.

Design considerations. While the need to identify single groups fast in the GGI problem led us to consider highly non-uniform sampling schemes, the possibility to share statistical strength across subgroups in the GCPI problem makes *successive elimination* algorithms [29, 40], which uniformly sample all subgroups that have not yet been eliminated for futility, a more attractive alternative: intuitively speaking, if all subgroups had exactly the same (positive) effect, uni-

Algorithm 2 AdaGCPI

Require: $\alpha, \beta \in (0,1)$, $\theta_{min} > 0$, budget B, identification rule $\mathcal{I}$, removal rule $\mathcal{R}$
Initialise: $\mathcal{A}_1 = \mathcal{K}$, set $\mathcal{D}_0 = \emptyset$, $t = 0$
while $t < B$ **do**
  Sample each $j \in \mathcal{A}_t$, obtain $\mathcal{D}' = \{j, Y^C_{t+j}, Y^T_{t+j}\}_{j \in \mathcal{A}_t}$, set $t \leftarrow t + |\mathcal{A}_t|$, update $\mathcal{D}_t$ with $\mathcal{D}'$.
  Test for positive effect in current population $\mathcal{I}(\mathcal{D}_t, \alpha)$: if detected, **Output:** `True`, $\mathcal{A}_t$
  Remove bad groups: $\mathcal{A}_t = \mathcal{K} \backslash \mathcal{R}(\mathcal{D}_t, \theta_{min}, \beta)$ and remove their samples from $\mathcal{D}_t$
  If $\mathcal{A}_t = \emptyset$, **Output:** `False`, $\emptyset$
end while
Output: `False`, $\emptyset$

formly allocating samples across all groups would lead to rejection of the *full population* composite null hypothesis using the same expected number of samples that the GGI problem would need to identify a *a single* group. Note that such potential efficiency of successive elimination in the GCPI problem stands in stark contrast to what has been observed for the *best arm* identification problem, where UCB-style algorithms empirically dominate successive elimination algorithms which are too wasteful in that context (see e.g. [41]). Further, successive elimination has the inherent advantage that it substantially limits the number of subpopulations (and associated null hypotheses) the algorithm will consider: if subgroups are irreversibly eliminated one-by-one, an algorithm will consider at most K (nested) subpopulations.

4.2. AdaGCPI: A Meta-algorithm for Good Composite Subpopulation Identification

To solve the GCPI problem, we propose AdaGCPI, an *Ada*ptive *G*ood *C*omposite sub*P*opulation *I*dentification meta-algorithm, as formalized in Algorithm 2. At each time step t, the algorithm proceeds by uniformly sampling all subgroups in the active set $\mathcal{A}_t$ by enroling two patients from each. For ease of presentation we assume equal sized subgroups ($\pi_j = \frac{1}{K}$) here but note that this could easily be avoided by sampling (with replacement) K indices from the active set according to the subgroup prevalence $\pi_j / \sum_{i \in \mathcal{A}_t} \pi_i$. We then apply an identification criterion $\mathcal{I}$ that tests for evidence of an *average* positive subpopulation effect across the active set. Upon success, the algorithm terminates; when evidence is not statistically significant, removal criterion $\mathcal{R}$ checks whether groups should be eliminated before enrolment continues. We discuss identification and removal criterion in turn below.

Identification criterion: Ensuring (approximate) FWER control. A full Bonferroni-style adjustment would require the significance level to be adjusted by 2^K, the number of hypotheses that could *potentially* be tested. As we only select *at most* K hypotheses for testing in practice, this adjustment is clearly overly conservative. If the K hypothesis tests were independent[4], we could use

$$\mathcal{I}^K_{BF}(\mathcal{D}_t, \alpha) = \mathbb{1}\{\hat{\theta}_{\mathcal{A}_t, N_{\mathcal{A}_t}(t)} - \phi(N_{\mathcal{A}_t}(t), \frac{\alpha}{K}) > 0\}$$

Clearly, they are not independent as datasets used for testing overlap, so identification using $\mathcal{I}^K_{BF}$ will not lead to exact FWER control. However, between selection and testing of a new hypothesis, at least $|\mathcal{A}_t|$ new samples accrue (and often many more), so any dependence decreases due to the online data collection. In experiments (Appendix D), we observe that FWER-α seems to hold empirically when using $\mathcal{I}^K_{BF}$, so we rely on it in our implementations.

Removal criterion: Exploiting subgroup *and* subpopulation signals. Using criterion $\mathcal{R}_{fut}(\mathcal{D}_t, \theta_{min}, \beta)$ as in AdaGGI, we remove *individual* subgroups for futility if their individual effects are insufficient. In addition, we exploit full subpopulation information by realising that the event $\mathcal{F}_t = \mathbb{1}\{\hat{\theta}_{\mathcal{A}_t, N_{\mathcal{A}_t}(t)} + \phi(N_{\mathcal{A}_t}(t), \beta) < \theta_{min}\}$ provides evidence that *at least* one subgroup has no sufficient treatment effect. Thus, if $\mathcal{F}_t$ is true, we remove the empirically worst subgroup through the rule $\mathcal{R}_{pop\text{-}fut}(\mathcal{D}_t, \mathcal{A}_t, \theta_{min}, \beta) = \arg\min_{j \in \mathcal{A}_t} \hat{\theta}_{j, N_j(t-1)} - \phi(N_j(t-1), \alpha)$ if $\mathcal{F}_t$ else $\emptyset$.

5. Experiments

5.1. Stylized Simulations: Understanding the (Dis)advantages of Different Strategies

Setup: In this section, we consider a stylized simulation setup to gain insight into the (dis)advantages of different sampling strategies and algorithms. Only here we assume that we observe a treatment effect signal $Y^\theta_t \sim \mathcal{N}(\theta_{J_t}, 1)$ *directly*; this also ensures that all our observations immediately generalize to the good *arm* identification problem. We consider $K = 10$ groups, $\pi_j = \frac{1}{K}, \forall j \in \mathcal{K}$ and let $\theta_{min} = 0.5$, $\alpha = 0.05$, $\beta = 0.1$. In the main results presented in Fig. 2, we let $\theta_k \in \{\theta_b, \theta_g\}$, where $\theta_b = 0$ and $\theta_g = 0.5$ unless stated otherwise, and vary $n_g = |\{j : \theta_j \geq 0.5\}|$. Throughout, we do not restrict budget and report t_{stop}, the stopping time of the algorithm (i.e. the time when *all* subgroups are classified as good or not), as well as $t^{id,j}_{good}$ and $t^{id,j}_{bad}$, the time taken to identify the j^{th} good group and to discard the j^{th} bad group, respectively;

[4] To gain further intuition, let $T_\mathcal{S}$ denote whether hypothesis $H_\mathcal{S}$ is selected for testing at any time, and $R_\mathcal{S}$ whether it is rejected. Using an argument adapted from the discussion of discard-spending in [34], we note that $FWER \leq \mathbb{E}[\sum_{\mathcal{S}:\theta_\mathcal{S} \leq 0} T_\mathcal{S} R_\mathcal{S}]$ by Markov's inequality. Further, $\mathbb{E}[\sum_{\mathcal{S}:\theta_\mathcal{S} \leq 0} T_\mathcal{S} R_\mathcal{S}] = \sum_{\mathcal{S}:\theta_\mathcal{S} \leq 0} \mathbb{E}[R_\mathcal{S} | T_\mathcal{S} = 1] P(T_\mathcal{S} = 1)$. If the data used to determine hypothesis selection $T_\mathcal{S}$ was independent of that used to determine rejection $R_\mathcal{S}$, we would have that $\mathbb{E}[R_\mathcal{S} | T_\mathcal{S} = 1] = \mathbb{E}[R_\mathcal{S}] = \mathbb{P}(\cap_{t=1}^\infty \{\hat{\theta}_\mathcal{S} - \theta_\mathcal{S} \geq \phi(t, \frac{\alpha}{K})\}) \leq \frac{\alpha}{K}$ so that $\mathbb{E}[\sum_{\mathcal{S}:\theta_\mathcal{S} \leq 0} T_\mathcal{S} R_\mathcal{S}] \leq \frac{\alpha}{K} \mathbb{E}[\sum_{\mathcal{S}:\theta_\mathcal{S} \leq 0} T_\mathcal{S}] \leq \frac{\alpha}{K} K$ as at most K hypotheses will be tested.

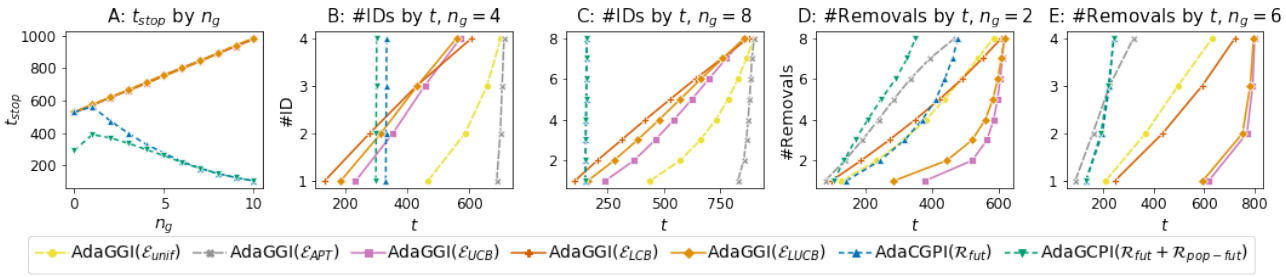


Figure 2. Results describing time until (A) termination, (B&C) identification of good groups and (D&E) removal of bad groups (1000 replications). (A): Time to termination t_{stop} by # of good groups n_g. (B&C): # of good group identifications over time, for $n_g = 4$ (B) and $n_g = 8$ (C). (D&E): # of removals of bad groups over time, for $n_g = 2$ (D) and $n_g = 6$ (E).

doing so allows us to understand what the algorithm would have identified given *any* budget. We compare AdaGGI with different sampling strategies – $\mathcal{E}_{UCB}$, $\mathcal{E}_{LCB}$ and $\mathcal{E}_{LUCB}$ as discussed in Sec. 3.2, as well as two baselines (discussed further in Appendix A.2): $\mathcal{E}_{unif}$, which uniformly samples groups that have not yet been identified, and $\mathcal{E}_{APT}$, which corresponds to [14]'s thresholding bandit solution – to AdaGCPI with different removal strategies ($\mathcal{R}_{fut}$ and $\mathcal{R}_{fut} + \mathcal{R}_{pop\text{-}fut}$). Some existing bandit algorithms arise as special cases of AdaGGI for the various sampling strategies we consider (see Appendix A.2.1). We discuss insights in turn below and present additional results in Appendix D.

Natural stopping times. In Fig. 2A, we investigate *how long* it would take the different algorithms to select/discard *all* subgroups (arms) for different n_g. First, we observe that the sampling strategy of AdaGGI has no impact on the stopping time; this is expected as identification of the final/worst group determines t_{stop}. Second, the total time to termination *increases* as n_g increases for AdaGGI because the identification criterion is *stricter* than the removal criterion. Third, AdaGCPI($\mathcal{R}_{fut}$), which is identical to AdaGGI($\mathcal{E}_{unif}$) except for the subpopulation-based identification criterion, performs identically to AdaGGI when $n_g \leq 1$ but begins to terminate earlier when n_g increases as sample size can be shared across $n_g \geq 2$ good subgroups. Finally, AdaGCPI($\mathcal{R}_{fut} + \mathcal{R}_{pop\text{-}fut}$) terminates fastest throughout, as it shares statistical strength across subgroups *both* when discarding and accepting subgroups; thus, the more homogeneous the population (n_g close to 0 or 10) the faster it terminates.

Time to identify the j^{th} good group. In Fig. 2B&C, we investigate *when* the different algorithms make *good* group discoveries, for $n_g = 4, 8$. When comparing algorithms, we find that AdaGGI generally makes the *first* discovery before AdaGCPI, as AdaGCPI makes *all* discoveries at the same time (yet this often happens before AdaGGI even makes its second discovery). When comparing sampling strategies within AdaGGI, major differences become visible. (Non-adaptive) uniform sampling now clearly appears suboptimal; as expected, the thresholding approach $\mathcal{E}_{APT}$, focussing on the groups hardest to distinguish from the threshold, performs even worse. Within the other adaptive strategies, $\mathcal{E}_{LCB}$ indeed makes the first discoveries faster than $\mathcal{E}_{UCB}$ in this setting, as the latter will unnecessarily switch between good groups as upper bounds cross (because the underlying means are identical); as expected, $\mathcal{E}_{LUCB}$ lies inbetween.

If the good groups were to exhibit quantitatively very different effects, the group with the largest θ_j should need least samples to be discovered – thus we would expect UCB-type strategies that haven proven successful in *best arm* identification [31] to be advantageous in this context. In Fig 3, we therefore further investigate the relative performance of sampling strategies when altering the underlying simulation: when the means in good groups are very different (Scen. 1: $\theta_1 = 0.5, \theta_2 = 1$; $\theta_j = 0, j > 2$) the relative performance indeed reverses. With more good arms and less spacing between means (Scen. 2: $\theta_j = 0.5 + \frac{0.5}{7}(j-1), j \leq 8$; $\theta_j = 0, j > 8$), this difference becomes less pronounced. In Appendix D, we additionally investigate how sampling strategies compare when outcome variance is known to differ across groups, and find that $\mathcal{E}_{LCB}$ can dominate as it intrinsically makes use of the fact that arms with *lower* variance need less samples to be identified, while $\mathcal{E}_{UCB}$ may erroneously focus on groups with high variance.

Time to discard the j^{th} bad group. In Fig. 2D&E, we investigate when the different algorithms *discard* groups that do not appear good. First, we observe that, unsurprisingly, AdaGCPI – an algorithm operating by suc-

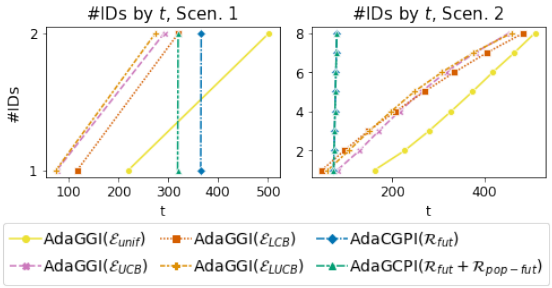


Figure 3. Good group identifications over time for two additional scenarios.

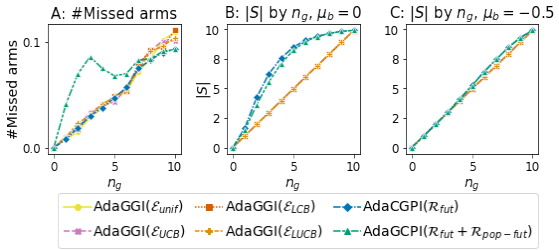


Figure 4. (A): Avg. number of missed groups by n_g. (B& C): Avg. |S| by n_g, for $\theta_b = 0, -0.5$.

cessive elimination – discards groups much faster than AdaGGI (with the exception of AdaGGI($\mathcal{E}_{APT}$), which essentially acts like a more aggressive elimination algorithm due to its focus on the threshold). Second, we observe that AdaGCPI($\mathcal{R}_{fut} + \mathcal{R}_{pop\text{-}fut}$) indeed benefits from the population-based elimination criterion as groups are discarded faster esp. when n_g small, which is when the population-based removal criterion will be met earlier. Third, we note that uniform sampling leads to faster elimination than (L)UCB-based sampling, which is expected as the latter actively avoid sampling groups that appear bad. Perhaps more surprisingly, LCB sampling leads to similarly fast discarding of the first bad groups, which we attribute to LCB being more likely to continue sampling from a group that has already been sampled often.

Incorrectly classified groups. Finally, we consider whether subgroups are (in)correctly classified as good. First, we note that, as we show in Appendix D, Type I error is not only controlled at level α but essentially 0 throughout (even when we remove the Bonferroni correction); we attribute this to the used anytime confidence intervals being unnecessarily conservative as $t \ll \infty$ here. Second, in Fig. 4A, we observe that good groups are seldomly missed by either algorithm (again, likely due to conservativeness of the bounds, the rate lies far below $\beta * n_g$); only AdaGCPI occasionally removes a good group with the aggressive removal criterion $\mathcal{R}_{pop\text{-}fut}$. Third, in Fig. 4B, we observe interesting differences in groups without effect that are included in the selected subpopulation $\mathcal{S}$ (note: for AdaGCPI, this does *not* necessarily constitute a Type I error as long as $\theta_S > 0$). As AdaGGI identifies groups individually, $|\mathcal{S}| \approx n_g$ throughout, while AdaGCPI allows *free-riding* of groups without effect on the larger effects of other groups, i.e. $|\mathcal{S}| > n_g$, especially when n_g is large, which leads to dilution of the effect on the full subpopulation but retains the average positive effect estimate. In Fig. 4C we set $\theta_b = -0.5$ instead of 0, and observe that this behavior ceases when groups contribute sufficiently large negative effects.

5.2. Application: Simulating a Clinical Trial

Finally, we apply our methods to a clinical trial setup. Because ground truth treatment effects are never observed in real data, papers on adaptive clinical trials (and the literature on treatment effect estimation more generally [42]) usually have to resort to simulation studies to evaluate the quality of their algorithms (e.g [3, 32]), where simulations are often semi-synthetic in that they are designed to reflect some qualities of real data. Here, we do so by building off the simulation setting presented in Section 6 of [3], which is in turn motivated by the I-SPY 2 breast cancer trial for neoadjuvant therapies [43]. We consider 3 equal sized subgroups with unknown treatment effect vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$ and as [3] let $\theta_{min} = 0.2, \alpha = 0.025$ and $\beta = 0.1$. Their setup considers binary outcomes $(Y_j^C \sim \mathcal{B}(\mu_{0,j}), Y_j^T \sim \mathcal{B}(\mu_{0,j} + \theta_j))$; in Appendix D we also consider normal outcomes. Using their budget calculations we set a budget of $B = 800$ pairs of patients. We compare AdaGCPI and AdaGGI to [3]'s proposed GSDS procedure as a baseline, which is structured similarly to AdaGCPI but (i) allows only n_a (preprespecified) interim analyses (in their study and here $n_a = 2$, allowing a single interim analysis halfway), (ii) selects and fixes subpopulation $\mathcal{S}$ at the first interim analysis and (iii) relies on explicitly calculated normal error-spending boundaries. GSDS and the simulation are further described in Appendix C.

Table 1. Results of 1000 simulated trials: Prop. of successful trials, avg. size of discovered subpopulation and, as prop. of budget: Avg. time to termination and avg. time to identification of the first good and bad group.

Scenario: $\boldsymbol{\theta}$	Method	%Succ.	$\lvert\mathcal{S}\rvert$	$\frac{t_{stop}}{B}$	$\frac{t_{1g}}{B}$	$\frac{t_{1b}}{B}$
A:$[0, 0, 0]$	GSDS	2.6	0.04	0.74		0.5
	AdaGGI	**0**	0	0.64		0.24
	AdaGCPI	**0**	0	**0.49**		**0.23**
B:$[-0.2, 0, 0.2]$	GSDS	**99.3**	1.19	0.64	0.64	0.5
	AdaGGI	97.9	0.98	0.63	**0.46**	0.38
	AdaGCPI	95	1.04	**0.61**	0.61	**0.15**
C:$[0, 0.1, 0.3]$	GSDS	**100**	2.03	**0.50**	0.50	0.50
	AdaGGI	99	1.00	0.55	**0.29**	0.59
	AdaGCPI	89	2.28	0.89	0.55	**0.44**
D:$[0.2, 0.2, 0.2]$	GSDS	**100**	2.98	0.50	0.5	
	AdaGGI	99.8	2.27	0.94	**0.36**	
	AdaGCPI	99.8	2.99	**0.37**	0.37	
E:$[0.3, 0.3, 0.3]$	GSDS	100	3	0.5	0.5	
	AdaGGI	100	3	0.49	**0.16**	
	AdaGCPI	100	3	**0.17**	0.17	

The original experiment in [3] has $\boldsymbol{\theta} \approx [0, 0.05, 0.1]$, i.e. all $\theta_j < \theta_{min}$, so that none of the designs are powered to detect any effect; indeed we find that across 1000 replications GSDS declares the trial successful 67% of the time, while AdaGGI and AdaGCPI[5] declare success only in 13% and 7% – a direct consequence of our designs discarding effects below the minumum clinically relevant θ_{min}. To gain more interesting insights into relative performance, we therefore consider five scenarios with varying $\boldsymbol{\theta}$ in Table 1.

[5]We focus on comparison with GSDS and use Sec. 5.1's overall best versions, AdaGGI($\mathcal{E}_{LCB}$) and AdaGCPI($\mathcal{R}_{fut}+\mathcal{R}_{pop\text{-}fut}$), as AdaGGI and AdaGCPI; full results are in Appendix D.

We observe that GSDS generally has more power to detect smaller effects. This is not surprising because (i) GSDS does not automatically discard groups below θ_{min} and (ii) the used anytime confidence intervals in both our algorithms are, as discussed above, overly conservative – especially when compared to the exact normal confidence bounds used in GSDS. Nonetheless, compared to our fully adaptive approaches, GSDS suffers from its rigidity (i.e. being restricted to pre-specified interim analysis points). In Scenarios B-D, it is apparent that both AdaGGI and AdaGCPI can make judgements about a single subgroup much before GSDS' first scheduled interim analysis (as before, AdaGGI generally finds the first good group faster, while AdaGCPI discards the first bad subgroup faster). In Scenarios A&E, where outcomes are extreme (all $\theta_j = 0$ and $\theta_j > \theta_{min}$, respectively), the advantage of the flexibility of AdaGCPI relative to GSDS is most obvious, as, due to the lack of restriction on analysis points, AdaGCPI can terminate *much* earlier than the first scheduled interim analysis of GSDS.

In summary, we thus find that our algorithms can outperform GSDS in some scenarios because they are much less constrained in terms of *when* they can terminate. At the same time, their performance in other scenarios is limited by the used anytime confidence intervals $\phi(\cdot, \cdot)$, which are less tight than the exact normal intervals used in GSDS. Investigating the use of other confidence intervals to instantiate our meta-algorithms would thus be an interesting avenue for future work.

6. Conclusion

We investigated how to adaptively identify patient subpopulations with treatment benefit during a clinical trial using ideas from ML, and proposed two problem formulations and associated meta-algorithms with different characteristics. We highlighted that the elimination-based AdaGCPI algorithm generally terminates using fewer samples, but may include subgroups that have no true benefit from treatment in the selected subpopulation if other groups have a sufficiently positive effect. Using AdaGGI, which discovers individual subgroups, this can generally be avoided – if one is willing to use substantially more samples. As we discuss further in Appendix B, we believe that the formalization of the population identification problem presented in this paper opens up many interesting avenues for future ML research in this context. In particular, we believe there is great potential for extending our setting to incorporate further practical requirements – e.g. allowing for delayed feedback or discovery of (not pre-specified) subgroups – and theoretical analysis of the considered algorithms and sampling strategies.

Societal Impact. There is a clear (ethical) tradeoff when deciding between algorithms to use in practice: AdaGCPI has the advantage that it may allow to bring a novel treatments to larger audiences faster and, due to uniform enrolment, does not give (arbitrary) preference to a single subgroup – but it may lead to prescription recommendations that include subgroups without effect. Conversely, AdaGGI has the advantage that it will recommend treatment only in truly good subgroups, yet highly non-uniform enrolment may lead to fairness concerns (e.g. due to the randomness in deciding which equally good group to recruit first) and trials may require much larger sample sizes and hence delay the release of a potentially life-saving treatment. [44] discusses similar issues for enrichment designs more generally.

Further, we note that the choice of using of adaptive trial designs instead of conventional non-adaptive trials is always highly situational [45]. There are definitely cases where one may want to continue to rely on conventional trial designs – e.g. applications where there are very large delays between patient enrolment and realisation of outcomes. There are, however, also cases in which adaptivity can be expected to be beneficial (it is sometimes even argued that adaptive designs are the only ethically permissible experimental designs, see e.g. [46] for a discussion). We therefore believe that adaptive enrichment designs like ours would be of most value in applications where high between-subgroup variability of effectiveness is expected.

Acknowledgements

We would like to thank anonymous reviewers for insightful comments and discussions on earlier drafts of this paper. AC gratefully acknowledges funding from AstraZeneca. AH is funded by the US Office of Naval Research (ONR).

References

[1] Peter F Thall. Adaptive enrichment designs in clinical trials. *Annual Review of Statistics and its Application*, 8:393–411, 2021.

[2] Nigel Stallard, Thomas Hamborg, Nicholas Parsons, and Tim Friede. Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of biopharmaceutical statistics*, 24(1):168–187, 2014.

[3] Baldur P Magnusson and Bruce W Turnbull. Group sequential enrichment design incorporating subgroup selection. *Statistics in medicine*, 32(16):2695–2714, 2013.

[4] Rita Nahta and Francisco J Esteva. Her-2-targeted therapy: lessons learned and future directions. *Clinical Cancer Research*, 9(14):5078–5084, 2003.

[5] Jay C Fournier, Robert J DeRubeis, Steven D Hollon, Sona Dimidjian, Jay D Amsterdam, Richard C Shelton, and Jan Fawcett. Antidepressant drug effects

and depression severity: a patient-level meta-analysis. *Jama*, 303(1):47–53, 2010.

[6] Michael Rosenblum, Brandon Luber, Richard E Thompson, and Daniel Hanley. Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in Medicine*, 35(21):3776–3791, 2016.

[7] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela Van Der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100, 2021.

[8] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[9] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[10] Ahmed Alaa and Mihaela van der Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138, 2018.

[11] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[12] Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

[14] Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.

[15] Jie Zhong, Yijun Huang, and Ji Liu. Asynchronous parallel empirical variance guided algorithms for the thresholding bandit problem. *arXiv preprint arXiv:1704.04567*, 2017.

[16] Chao Tao, Saúl Blanco, Jian Peng, and Yuan Zhou. Thresholding bandit with optimal aggregate regret. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Subhojyoti Mukherjee, Kolar Purushothama Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Thresholding bandits with augmented ucb. *arXiv preprint arXiv:1704.02281*, 2017.

[18] Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Machine Learning*, 108(5):721–745, 2019.

[19] Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pages 1781–1791. PMLR, 2020.

[20] Christopher Jennison and Bruce W Turnbull. Adaptive seamless designs: selection and prospective testing of hypotheses. *Journal of biopharmaceutical statistics*, 17(6):1135–1161, 2007.

[21] US Food and Drug Administration. Enrichment strategies for clinical trials to support determination of effectiveness of human drugs and biological products: Guidance for industry. 2019.

[22] Anne G Copay, Brian R Subach, Steven D Glassman, David W Polly Jr, and Thomas C Schuler. Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal*, 7(5):541–546, 2007.

[23] Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR, 2018.

[24] Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? *Advances in Neural Information Processing Systems*, 32, 2019.

[25] Ramesh Johari, Leo Pekelis, and David J Walsh. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*, 2015.

[26] Kevin G Jamieson and Lalit Jain. A bandit approach to sequential experimental design with false discovery control. *Advances in Neural Information Processing Systems*, 31, 2018.

[27] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

[28] Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. A unified framework for bandit multiple testing. *Advances in Neural Information Processing Systems*, 34, 2021.

[29] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53. Citeseer, 2010.

[30] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 25, 2012.

[31] Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

[32] Onur Atan, William R Zame, and Mihaela Schaar. Sequential patient recruitment and allocation for adaptive clinical trials. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1891–1900. PMLR, 2019.

[33] US Food and Drug Administration. The drug development process: Step 3, clinical research. 2018.

[34] Jinjin Tian and Aaditya Ramdas. Online control of the familywise error rate. *Statistical Methods in Medical Research*, 30(4):976–993, 2021.

[35] Martin Jenkins, Andrew Stone, and Christopher Jennison. An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics*, 10(4):347–356, 2011.

[36] Tim Friede, N Parsons, and Nigel Stallard. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in medicine*, 31(30):4309–4320, 2012.

[37] Michael Rosenblum, Tianchen Qian, Yu Du, Huitong Qiu, and Aaron Fisher. Multiple testing procedures for adaptive enrichment designs: combining group sequential and reallocation approaches. *Biostatistics*, 17(4):650–662, 2016.

[38] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27, 2014.

[39] Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh, Ronald Ortner, and Peter Bartlett. Improved learning complexity in combinatorial pure exploration bandits. In *Artificial Intelligence and Statistics*, pages 1004–1012. PMLR, 2016.

[40] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

[41] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[42] Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[43] AD Barker, CC Sigman, GJ Kelloff, NM Hylton, DA Berry, and LJs Esserman. I-spy 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100, 2009.

[44] Boris Freidlin, Zhuoxin Sun, Robert Gray, and Edward L Korn. Phase iii clinical trials that integrate treatment and biomarker evaluation. *Journal of Clinical Oncology*, 31(25):3158, 2013.

[45] Christopher R Palmer and William F Rosenberger. Ethics and practice: alternative designs for phase iii randomized clinical trials. *Controlled clinical trials*, 20(2):172–186, 1999.

[46] Nicolas Fillion. Clinical equipoise and adaptive clinical trials. *Topoi*, 38(2):457–467, 2019.

[47] Boris Freidlin and Richard Simon. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical cancer research*, 11(21):7872–7878, 2005.

[48] Boris Freidlin, Wenyu Jiang, and Richard Simon. The cross-validated adaptive signature design. *Clinical cancer research*, 16(2):691–698, 2010.

[49] Zhiwei Zhang, Meijuan Li, Min Lin, Guoxing Soon, Tom Greene, and Changyu Shen. Subgroup selection in adaptive signature designs of confirmatory clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(2):345–361, 2017.

[50] Noah Simon and Richard Simon. Adaptive enrichment designs for clinical trials. *Biostatistics*, 14(4):613–625, 2013.

[51] Kevin SS Henning and Peter H Westfall. Closed testing in pharmaceutical research: Historical and recent developments. *Statistics in biopharmaceutical research*, 7(2):126–147, 2015.

[52] Sue-Jane Wang, HM James Hung, and Robert T O'Neill. Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(2):358–374, 2009.

[53] Werner Brannath, Emmanuel Zuber, Michael Branson, Frank Bretz, Paul Gallo, Martin Posch, and Amy Racine-Poon. Confirmatory adaptive designs with

bayesian decision tools for a targeted therapy in oncology. *Statistics in medicine*, 28(10):1445–1463, 2009.

[54] Yi-Da Chiu, Franz Koenig, Martin Posch, and Thomas Jaki. Design and estimation in clinical trials with subpopulation selection. *Statistics in medicine*, 37(29):4335–4352, 2018.

[55] Alihan Hüyük, Zhaozhi Qian, and Mihaela van der Schaar. When to make and break commitments? In *The Eleventh International Conference on Learning Representations*, 2023.

[56] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[57] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

[58] Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

[59] Séebastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265. PMLR, 2013.

[60] Alexandra Carpentier and Andrea Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.

[61] Oded Maron and Andrew Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. *Advances in neural information processing systems*, 6, 1993.

[62] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

[63] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pages 672–679, 2008.

[64] Shivaram Kalyanakrishnan and Peter Stone. Efficient selection of multiple bandit arms: Theory and practice. In *ICML*, 2010.

[65] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[66] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

[67] Julian Katz-Samuels and Clay Scott. Feasible arm identification. In *International Conference on Machine Learning*, pages 2535–2543. PMLR, 2018.

[68] Yoan Russac, Christina Katsimerou, Dennis Bohle, Olivier Cappé, Aurélien Garivier, and Wouter M Koolen. A/b/n testing with control in the presence of subpopulations. *Advances in Neural Information Processing Systems*, 34, 2021.

[69] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

[70] Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[71] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[72] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[73] Xuelin Huang, Jing Ning, Yisheng Li, Elihu Estey, Jean-Pierre Issa, and Donald A Berry. Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Statistics in medicine*, 28(12):1680–1689, 2009.

[74] Aditya Grover, Todor Markov, Peter Attia, Norman Jin, Nicolas Perkins, Bryan Cheong, Michael Chen, Zi Yang, Stephen Harris, William Chueh, et al. Best arm identification in multi-armed bandits with delayed feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 833–842. PMLR, 2018.

[75] Roland Gerard Gera and Tim Friede. Blinded sample size re-calculation in multiple composite population designs with normal data and baseline adjustments. *arXiv preprint arXiv:2011.14735*, 2020.

A. Appendix A: Additional Literature Review

A.1. Extended review of adaptive clinical trial literature on enrichment designs

Below, we discuss in some more detail the clinical trials literature on adaptive enrichment trials which allow discontinuation of subgroups and changes of the population (and hence hypothesis) under consideration in a clinical trial. We focus here on designs where the subgroups under investigation are prespecified; subgroup discovery in the presence of single or multiple biomarkers is covered in e.g. the literature on so-called *adaptive signature designs*. [47–50]. For a broader review of adaptive enrichment designs, refer to [1].

[20] (Section 6) describe a generic two-stage enrichment design with K subpopulations that are *not* necessarily disjoint (as in our case) or nested, which allows for selection of an arbitrary population j^* after the first stage, after which recruitment is focussed on j^* and the hypothesis to be tested is H_{0j*}, where error is controled through application of closed testing procedures [51]. [52] also consider two-stage trials under different types of restrictions with multiple nested subpopulations determined by biomarker interactions.

[35, 36, 53] all consider a setting where one can either consider the full population or a *single* pre-specified subpopulation of heightened interest; at an interim analysis it is to be decided whether to continue with the full population or within the subpopulation only (or not at all). These designs differ in both the rules for population selection and the hypothesis tests used, but all rely on closed testing principles. [2] empirically compares some of these and other approaches for population selection and hypothesis testing in the adaptive enrichment problem with two subpopulation and a single interim analyses.

[6] also considers a setting where either the full population or a single promising subgroup is of interest, however, instead of only one interim analysis the trial has multiple analyses where the trial can be stopped for efficacy/futility in either the full population or the subgroups based on normal stopping boundaries. Finally, [3] propose a design ([54] discuss a multistage design analogous to theirs) that is most closely related to our AdaGCPI approach, where the main differences lie in that (i) [3] *fix* the selected subpopulation after the first stage and (ii) exact probability boundaries are calculated for termination. We describe [3]'s proposed GSDS procedure in more detail in Appendix C.

The most relevant related work from the ML literature that we are aware of is [32]; they also consider adaptive recruitment to discover all good subgroups and do so using a Bayesian MDP-based design that learns by optimizing an objective function that *trades off* Type I and II error given a limited budget. As such, type I error is neither controlled nor is multiplicity considered, making this approach (objective) conceptually, i.e. abstracting away specific implementation choices, most similar to good arm identification and thresholding bandits under a *fixed budget* (only) setting. Finally, [55] recently considered subpopulation selection in adaptive clinical trials but for *portfolio-level management* of trials rather than the sample-by-sample decisions we consider here.

A.2. Extended contextualization of the GGI and GCPI problems within bandit literature

GGI and GCPI are closely related to multi-armed bandit problems as one can interpret each considered subgroup as an *arm* and their (unknown) treatment effect as the *mean reward* of that arm. Typically, the goal in a bandit problem is to maximize the rewards of all arms that are "played" (e.g. [56]). Since the mean rewards are unknown initially, this requires striking a balance between *exploring* arms to gain information about their rewards and *exploiting* arms that appear to have high rewards. In our setting, this conventional objective would have corresponded to maximizing the benefit received by all patients recruited into the trial. Instead, we focus on what is known as *pure exploration* in the bandit literature, where the rewards of played arms do not matter except for that of a singular arm identified at the end [13, 38, 57, 58].

Different purely-exploratory objectives have been considered in the multi-armed bandit literature. Best arm identification (BAI) problems aim to identify the arm (or the top-K arms) with the largest mean reward (e.g. [29]). Here, the success can be measured via the reward gap between the identified arm and the true best arm. In the *fixed budget* setting, the goal is to maximize the probability of the identified arm indeed being the best given a fixed budget of samples [30, 59, 60], while in the *fixed confidence* setting, the goal is to minimize the number of samples necessary to guarantee a fixed level of confidence [30, 31, 61–66]. Good arm identification (GAI) problems (sometimes called pure exploration in thresholding bandits) aim to identify arms with mean rewards that are higher than a pre-specified threshold. These problems too can be considered either in fixed budget [14, 17, 67] or fixed confidence [18, 19] settings.

GGI is essentially a type of GAI problem but it requires both the budget as well as the confidence in each identified arm being good to be fixed, and given those constraints, aims to identify as many good arms as possible. In existing formulations

Table 2. Comparison of pure exploration problems. GGI and GCPI uniquely require both the budget as well as the confidence to be fixed, and aim to identify as many suitable arms as possible within those constraints. In contrast, other problems aim to identify all suitable arms, which is only possible with the more relaxed constraint of either just the budget or just the confidence being fixed. FB and FC stand for fixed budget and fixed confidence respectively.

Problem	Ref.	Type of arms identified	Number of arms identified	Budget	Confidence	Formulation
BAI	[29]	Best arms $i^* = \text{argmax}_i \theta_i$	Top-K arms	Variable	Variable	minimize $\theta_{i^*} - \theta_{\hat{i}^*}$
BAI w/ FB	[60]			Fixed (T)	Maximized	maximize $\mathbb{P}(\hat{i}^*(T) = i^*)$
BAI w/ FC	[66]			Minimized	Fixed ($1-\delta$)	minimize T s.t. $\mathbb{P}(\hat{i}^*(T) \neq i^*) \leq \delta$
GAI w/ FB	[14]	Good arms $\mathcal{I} = \{i : \theta_i > \xi\}$	All good arms	Fixed (T)	Maximized	maximize $\mathbb{P}(\hat{\mathcal{I}}(T) = \mathcal{I})$
GAI w/ FC	[19]			Minimized	Fixed ($1-\delta$)	minimize T s.t. $\mathbb{P}(\hat{\mathcal{I}}(T) \neq \mathcal{I}) \leq \delta$
GGI	**(Ours)**	Good arms $\mathcal{I} = \{i : \theta_i > \xi\}$	Maximized	Fixed (T)	Fixed ($1-\delta$) w.r.t. type I error	maximize $\lvert\hat{\mathcal{I}}(T)\rvert$ s.t. $\mathbb{P}(\hat{\mathcal{I}}(T) \setminus \mathcal{I} \neq \emptyset) \leq \delta$
GCPI	**(Ours)**	Good composite arms $\mathcal{I} : \frac{1}{\lvert\mathcal{I}\rvert}\sum_{i\in\mathcal{I}} \theta_i > \xi$	Maximized	Fixed (T)	Fixed ($1-\delta$)	maximize $\lvert\hat{\mathcal{I}}(T)\rvert$ s.t. $\mathbb{P}\Big(\frac{1}{\lvert\hat{\mathcal{I}}(T)\rvert}\sum_{i\in\hat{\mathcal{I}}(T)} \theta_i \not> \xi\Big) \leq \delta$

of GAI, the aim is usually to identify *all* good arms, which is only possible with the more relaxed constraint of either just the budget or the confidence being fixed (but not both at the same time). GCPI is similar to GGI in that it too requires both the budget and the confidence to be fixed but it only aims to identify a collection of arms that are good *on average*[6] rather than arms that are all individually good. Table 2 formally compares GGI and GCPI with the existing pure exploration problems.

A.2.1. How existing pure exploration solutions arise as special cases of AdaGGI

One of the main goals of this paper is to formalize, contextualize and understand the trial population identification problem as a pure exploration bandit problem. Because our paper considers a new problem formulation, there – to the best of our knowledge – are no off-the-shelf solutions from the bandit literature that have already solved this exact problem. Therefore, this paper studies how to apply and adapt solutions proposed for related problems and empirically investigates how different approaches work in our context.

To do so, we study very generic meta-algorithms, which give rise to adaptations of some existing combinatorial bandit solutions as special cases, allowing for fair comparison of different approaches. Note that both the thresholding bandit and good arm identification (GAI) are combinatorial bandit instances and their specific problem formulations are closer to our problem setting than generic combinatorial bandits, making their solutions more likely to perform well in our context. Below, we discuss in detail how GAI algorithms, thresholding bandits and a generic combinatorial bandit solution arise as variations of AdaGGI and can thus be seen as 'bandit baselines' in our experiments.

GAI algorithms – AdaGGI($\mathcal{E}_{UCB}$) As outlined in Section 3, the GAI algorithms proposed in [18, 26] proved most suitable to adapt to our setting and thus share a very similar backbone to AdaGGI. The main conceptual differences to existing implementations lies in that (i) they exclusively rely on UCB-sampling and (ii) have no [26] or a stricter [18] removal criterion. The special case $\mathcal{E}_{UCB}$ could thus be seen as a GAI-bandit baseline with adapted removal criterion. Adaptation of the removal criterion to allow discarding of groups without clinically relevant effect greatly improves those algorithms with respect to stopping time; the original criteria lead to infinite running times when 'bad' group effects are exactly zero (as in our experiments).

Thresholding bandit – AdaGGI($\mathcal{E}_{APT}$) Another approach that could be adapted to our setting is [14]'s thresholding bandit solution. Because the thresholding bandit problem aims at correctly classifying *all* arms as either good or bad using a fixed budget, [14]'s Anytime Parameter-free Thresholding (APT) algorithm tries to equalize the confidence in the classification of all arms by ensuring that $N_j(t)(\hat{\theta}_{j,N_j(t)} - \theta_0)^2$ is constant across arms. This corresponds to a sampling strategy $\mathcal{E}_{APT}(\mathcal{D}_{t-1}, \mathcal{A}_{t-1}) = \arg\min_{j\in\mathcal{A}_{t-1}} \sqrt{N_j(t-1)}(\hat{\theta}_{j,N_j(t-1)} - \theta_0)$ with $\theta_0 = 0$ in our setup. Conceptually, this will lead to sampling the groups that are *furthest* from being identified – this is the opposite strategy to what $\mathcal{E}_{LCB}$ tries to accomplish and cannot be expected to perform well in our context. Because the original paper [14] focusses on a fixed

[6] [68] also consider a GAI problem involving weighted averages over collections of subgroups in a population, but there, the viable collections and weights are fixed and not part of the optimization problem, which is very different from the GCPI problem where the collections are optimized over.

budget only setting, it is lacking some form of identification and removal criterion. We therefore instantiate it using the AdaGGI backbone and simply use $\mathcal{E}_{APT}$ as the sampling strategy.

Generic combinatorial bandit baseline – AdaGGI($\mathcal{E}_{unif}$) Finally, we consider adapting more generic combinatorial bandit solutions, which generally aim to optimize some objective over collections of arms. Here, we consider [69]'s Combinatorial Upper Confidence Bound (CUCB) algorithm in more detail as it permits straightforward adaptation to our setting. The general setting considered in [69] allows to play a super-arm $\mathcal{S}$ at each time t, and their algorithm assumes existence of an oracle that outputs the optimal $\mathcal{S}$ whenever provided with the underlying distributions of all arms; in the GAI setting this simply picks all arms whose means exceed the threshold. The algorithm proceeds by constructing upper confidence bounds $\tilde{\theta}_{j,t} = \hat{\theta}_{j,N_j(t-1)} + \phi(N_j(t-1), \beta)$ on the means of all arms, and then applies the oracle to the $\tilde{\theta}_{j,t}$, outputting a super-arm $\mathcal{S}_t$ to sample. In our context, this would sample all arms for which it holds that $\hat{\theta}_{j,N_j(t-1)} + \phi(N_j(t-1), \beta) > \theta_0$. Note that this essentially corresponds to AdaGGI with removal criterion $\mathcal{R}_{fut}(\mathcal{D}_t, \theta_0, \beta)$ instead of $\mathcal{R}_{fut}(\mathcal{D}_t, \theta_{min}, \beta)$, and uniform sampling of the active set. As discussed above, setting $\theta_{min} \neq \theta_0$ can only improve the algorithm's performance; thus AdaGGI($\mathcal{E}_{unif}$) – i.e. simple uniform sampling of the active set – corresponds to a straightforward adaptation of the CUCB algorithm to our setting.

B. Appendix B: Possible Extensions and Future Work

We believe that this paper opens up many interesting avenues for future research; natural next steps lie in (i) extending the setting under consideration to incorporate more realistic problem features and (ii) further studying and improving components of the algorithms.

Extending the setting. Multiple modifications to the data generating process might lead to a more realistic setting and interesting research problems at the same time:

- **Considering batched (grouped) observations:** In practice, it might be operationally difficult to collect and reveal *individual* patient responses as they come in; instead it might be more easily feasible to release patient responses in *batches* or *groups* as is commonly done in *group sequential designs* [70]. AdaGCPI could directly accommodate this: instead of recruiting $|\mathcal{A}_t|$ patient pairs uniformly and evaluating the subpopulation immediately, a larger batch of patients could be recruited (uniformly from the active set) before using the updated dataset for testing the hypothesis. Doing the same for AdaGGI may not be optimal, as – because the original sampling strategies are *deterministic* – one would then have to recruit an entire batch of patients from the same subgroup, which may explore insufficiently. Instead, sampling strategies that resemble Thompson sampling [71, 72] – i.e. strategies that are *random* and recruit patients proportionally to *the probability of their subgroup being good* – may be more suited to this scenario.

- **Allowing delayed feedback:** Another difficulty likely to be encountered in practice, particularly when considering time-to-event data or other long term outcomes, might be that not all outcomes of previously recruited patients are available when making the next recruitment decision. The biostatistics literature has investigated how one can use available *short term outcomes* that are indicative of the long term outcomes in such scenarios [73], while the bandit literature has developed approaches for decision making under delayed feedback [74]; it would be interesting to investigate how to incorporate either into our framework.

- **Incorporating covariates and discovering subgroups:** An interesting extension to the setting considered here would be to make use of any other patient information (context) that may be available, e.g. *prognostic* information that may explain some baseline variation likely to exist in practice and hence improve precision of estimators (as in e.g. [75]). When no subgroups are pre-specified, one may also make use of such information to *discover* subgroups that differ in their treatment response through so-called *adaptive signature designs* [47–50]; investigating how to better use ML tools to efficiently discover such subgroups may be a natural next step.

Analyzing problem settings and algorithms. We believe that a number of our empirical findings both motivate further theoretical analyses *and* suggest that improvements to our implementations may be possible:

- **Comparing problem complexity of GGI and GCPI theoretically:** Our experiments confirmed the intuition that the GCPI problem can be easier (faster) to solve than the GGI problem, especially when subgroups are close to

homogeneously *all* good or bad. It would be an interesting avenue for future work to confirm and analyze this theoretically.

- **Comparing sampling strategies theoretically:** Our experiments also confirmed the intuition that, depending on the underlying problem structure, different (non-uniform) sampling strategies are better at discovering (the first) good arm fast, and it would thus be interesting to formally derive scenarios in which either UCB or LCB strategies could be expected to have an advantage.
- **Improving the used confidence intervals:** We observed in our experiments that the $\phi(\cdot,\cdot)$ that we used seemed to create overly conservative confidence intervals in our settings. One possibility to improve this may be to rely on the fact that usually $B \ll \infty$ and to therefore construct alternatives that instead of allowing for infinitely many peeks at the data, allow only $L \leq B$ decision points which may lead to less necessity to be conservative.
- **Explicitly incorporating budget in sampling and elimination strategies:** Finally, we note that it may be an interesting avenue for future work to develop a removal criterion $\mathcal{R}_{Budget}$ that forces early removal of a subgroup, either permanently from AdaGCPI whenever it is expected that there is insufficient budget left to prove treatment effectiveness with confidence α in the current subpopulation (this would be the case if the average effect in the current subpopulation is likely too low to do so; more aggressively removing the subgroups that appear worst may be appropriate in this context) or temporarily from consideration for sampling in AdaGGI. Alternatively, one could also investigate new multi-stage meta-algorithms with an initial more exploratory stage and a later more exploitative stage, where the number of samples allocated to each stage or the transition between stages would depend on budget.

C. Appendix C: Experimental Details

C.1. Stylized simulations (Section 5.1)

All data was generated according to the setup described in Section 5.1: There are $K = 10$ groups, $\pi_j = \frac{1}{K}, \forall j \in \mathcal{K}$ and outcomes are normally distributed according to $\mathcal{N}(\theta_k, 1)$, where $\sigma^2 = 1$ is assumed *known*. In the main results presented in Fig. 2, we let $\theta_k = 0.5$ for $k \leq n_g$ and $\theta_k = 0$ for $k > n_g$, for $n_g \in \{0, \ldots, 10\}$. In Fig. 4, we set bad means equal to -0.5, and in Fig. 3 we let good means vary between 0.5 and 1 by setting them equal to $\theta_j = 0.5 + \frac{0.5}{n_g - 1}(j-1), j \leq n_g$.

For all algorithms we set $\theta_{min} = 0.5$, $\alpha = 0.05$, $\beta = 0.1$ and $n_0 = 1$. As Y^θ is assumed normally distributed with known variance $\sigma^2 = 1$, we use $\phi(t,\delta) = \sqrt{2\frac{\log(1/\delta) + 3\log\log(1/\delta) + (3/2)\log\log(et/2))}{t}}$ as in [26]. Note that we can use this for both AdaGGI and AdaGCPI as all outcomes in any subpopulation are distributed equally under the null hypothesis (regardless of subgroup, under the null hypothesis all outcomes are distributed according to $\mathcal{N}(0,1)$).

C.2. Simulated trials (Section 5.2)

In Section 5.2, we use a modified version of the experiment in section 6 of [3], which is in turn motivated by the I-SPY 2 breast cancer trial for neoadjuvant therapies [43]. The assumed end point of interest is the occurrence of pathologic complete response (pCR), [3] assume this to follow a Bernoulli distribution where for the controls $Y^C \sim \mathcal{B}(0.4)$ for all subgroups while the outcomes in treated individuals can differ across subgroups as $Y_j^T \sim \mathcal{B}(0.4 + \theta_j)$. As [3] we consider 3 subgroups, for simplicity we assume them to be equal sized ($\pi_k = \frac{1}{3}$) here. In addition to the Bernoulli setting from the main text, we also consider an additional setting with normally distributed outcomes in Appendix D (with known $\sigma^2 = 1$) i.e. $Y_j^C \sim \mathcal{N}(0,1), Y_j^T \sim \mathcal{N}(\theta_j, 1), \forall j \in [3]$.

As [3] we let $\theta_{min} = 0.2$, $\alpha = 0.025$ and $\beta = 0.1$. For our algorithms we additionally let $n_0 = 5$ due to the higher variance induced by considering *a difference* between random variables now. As the difference between two normal random variables with variance σ^2 is normal with variance $2\sigma^2$, we use $\phi(t,\delta) = 2\sqrt{\frac{\log(1/\delta) + 3\log\log(1/\delta) + (3/2)\log\log(et/2))}{t}}$ for the normal outcomes, and, as bernoulli variables are $\frac{1}{4}$ subgaussian, we use $\phi(t,\delta) = \sqrt{\frac{\log(1/\delta) + 3\log\log(1/\delta) + (3/2)\log\log(et/2))}{t}}$ for the difference between binary outcomes.

Description of GSDS. We now briefly formally describe the group sequential design for subgroups (GSDS) proposed in [3]. The design requires: a pre-specified number of interim analyses n_a, a test statistic $Y_j(t)$ and associated Fisher information $\mathcal{I}_j(t)$, a desired significance level α and power $1 - \beta$. $\dashv$ is used to calculate stopping boundaries $\{(l_p, u_p)\}_{p=1}^{n_a}$ for each interim analysis. β is used to calculate a *maximum information* level $\mathcal{I}_{max}$, which is in turn used to determine the

sample size. The algorithm proceeds as follows: at the first interim analysis at time t_1, a subpopulation is selected through exclusion of all bad subgroups: $\mathcal{S}^* = \{j \in \mathcal{K} : Y_j(t_1)\sqrt{\mathcal{I}_j(t_1)} > l_1\}$. If $Y_{\mathcal{S}^*}(t_1)\sqrt{\mathcal{I}_{\mathcal{S}^*}(t_1)} > u_1$, the trial terminates immediately for efficacy; otherwise the trial continues and at all $n_a - 1$ subsequent stages, the trial is terminated for efficacy if $Y_{\mathcal{S}^*}(t_k)\sqrt{\mathcal{I}_{\mathcal{S}^*}(t_k)} > u_k$ and terminated for futility if $Y_{\mathcal{S}^*}(t_k)\sqrt{\mathcal{I}_{\mathcal{S}^*}(t_k)} < l_k$.

Budget calculation. We follow the example in [3] who calculate that for a two stage trial with $\alpha = 0.025$, $\beta = 0.1$ and $\theta_{min} = 0.2$, we have $(l_1, u_1) = (0.7962, 2.7625)$ and $l_2 = u_2 = 2.5204$ and $\mathcal{I}_{max} = 1495.5$.

In their example with binary outcomes, if we let b denote the number of *pairs* of recruited patients[7], and $\hat{p}^C, \hat{p}^T$ the observed binary proportions in each group, we have that

$$Y = \hat{p}^T - \hat{p}^C \text{ and } \mathcal{I} = \frac{b}{2\tilde{p}(1-\tilde{p})} \tag{4}$$

where $\tilde{p}$ is the average response rate and is conservatively set to 0.5. Solving $\mathcal{I}_{max}$ for b yields a (rounded) budget $B = 800$ pairs of patients.

Similarly, when doing the same for normal outcomes with known variance σ^2, if we let $\hat{\mu}^C, \hat{\mu}^T$ denote the means in treated and control arm, we have

$$Y = \hat{\mu}^T - \hat{\mu}^C \text{ and } \mathcal{I} = \frac{b}{2\sigma^2} \tag{5}$$

and with $\sigma^2 = 1$ this yields a rounded budget of $B = 3000$.

D. Appendix D: Additional Results

D.1. Additional simulation results (Sec 5.1)

D.1.1. Identifications: Complete results

In Fig. 5, we present results capturing time until identification of each good group for $n_g \in \{2, 4, 6, 8, 10\}$ (only $n_g = 4, 8$ are presented in the main text). In Fig. 6, we present results capturing time until removal of each bad group for $n_g \in \{0, 2, 4, 6, 8\}$ (only $n_g = 2, 6$ are presented in the main text). These results reflect the same insights as those presented in the main text, both in terms of comparing algorithms and in terms of comparing sampling strategies.

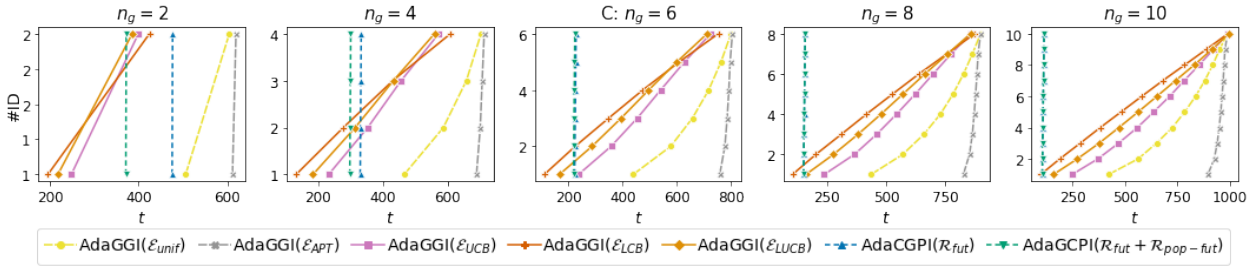


Figure 5. Results describing identification of good groups over time, for $n_g \in \{2, 4, 6, 8, 10\}$; avg. across 1000 replications.

D.1.2. Type I error

In Fig. 7 we plot Type I errrors committed over 1000 simulation runs, both with and without Bonferroni correction. (Note that a Type I error is defined as *any* null hypothesis being incorrectly rejected; for AdaGGI this includes any *single* subgroup being incorrectly declared good, while for AdaGCPI this would mean that the selected subpopulation $\mathcal{S}$ does not have a positive *average* effect.) We make multiple interesting observations: First, with Bonferroni correction, all identification criteria are clearly overly conservative – incorrect rejections only happen when *all* groups are bad, and even then this lies much below the used $\alpha = 0.05$. Second, this is not primarily due to the conservativeness of the Bonferroni correction, but due to the conservativeness of the used anytime confidence interval: even when we remove the Bonferroni correction, all

[7] We believe there is a typo in Sec. 6 of [3], so that n should denote the number of *pairs* of patients, and not patients. We have adapted budget calculations accordingly

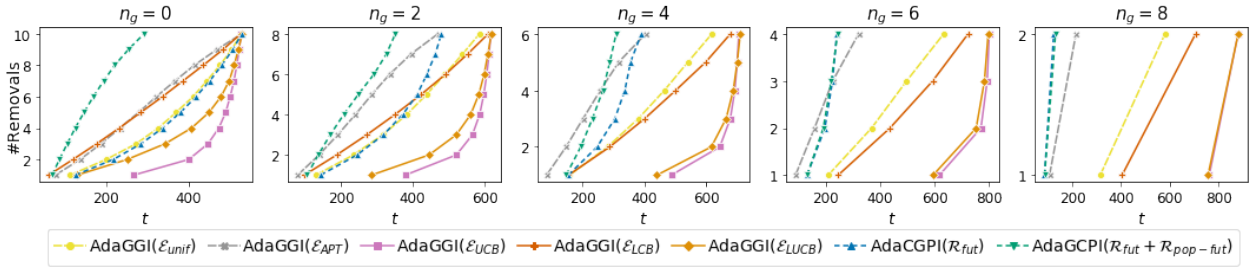


Figure 6. Results describing removal of bad groups over time, for $n_g \in \{0, 2, 4, 6, 8\}$; avg. across 1000 replications.

Type 1 errors remain below $(10 - n_g) * \alpha$ (in fact, they even lie below α). Finally, we note that the approximate Bonferroni correction we chose for AdaGCPI therefore does not appear to be problematic; in the plot *without* any correction we also observe that AdaGCPI does not seem to be more likely to commit a Type I error than AdaGGI even without any correction (despite the number of hypotheses that could potentially be tested being exponential versus linear in K).

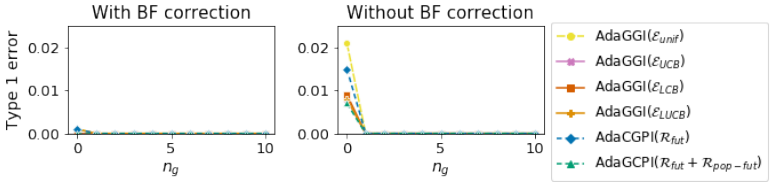


Figure 7. Type 1 error (runs with *any* incorrectly rejected null hypothesis) by n_g across 1000 replications. Identification *with* Bonferroni correction (left) and without (right).

D.2. Additional simulation scenarios

Varying means We present additional results on the setting presented in Fig. 3 of the main text: for $n_g \in \{2, \dots, 10\}$ we let $\theta_j = 0.5 + 0.5\frac{j-1}{n_g-1}$ for $j \leq n_g$ and $\theta_j = 0$ otherwise. As discussed in the main text, we observe that the relative performance of sampling strategies changes in this setting: $\mathcal{E}_{LCB}$ generally performs worse than $\mathcal{E}_{UCB}$ here; with increasing n_g and hence decreasing spacing between the good means, this effect reduces.

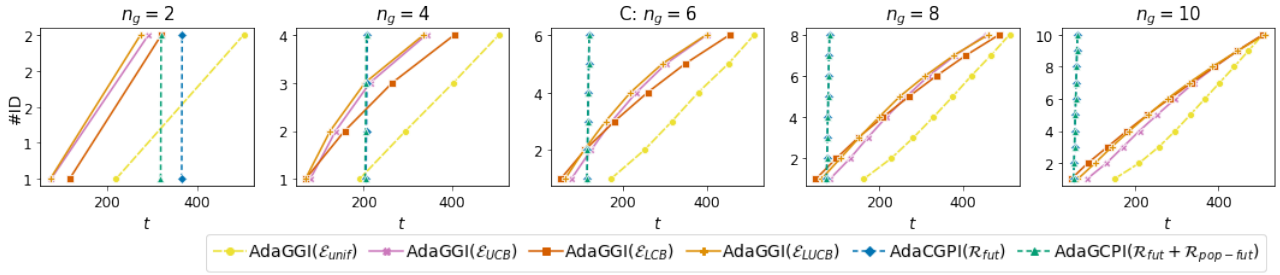
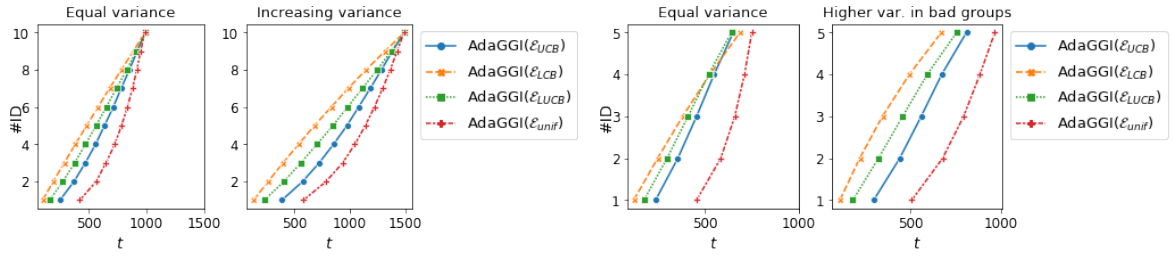


Figure 8. Results describing identification of good groups over time, for $n_g \in \{2, 4, 6, 8, 10\}$ for a setting with varying means; avg. across 1000 replications.

Different variances Next we consider how changing variance affects the performance of the different sampling algorithms. In the Fig. 9(a), we compare the original setting with $n_g = 10$ and $\sigma^2 = 1$ for all groups to one where the means are the same but $\sigma_j^2 = 1 + \frac{j-1}{n_g}$ grows across groups. In the right Fig. 9(b), we compare the original setting with $n_g = 5$ and $\sigma^2 = 1$ for all groups to one where $\sigma^2 = 2$ for the bad groups, while $\sigma^2 = 1$ for the good groups. We observe that identification times worsen across the board in both settings, but that the time increase for the first identifications for $\mathcal{E}_{UCB}$ is much larger than that for $\mathcal{E}_{LCB}$ in absolute terms – most likely because UCB-style algorithms may erroneously enrol groups with larger variance as the UCB will generally be higher for these groups.



(a) $n_g = 10$ (b) $n_g = 5$

Figure 9. Results describing identification of good groups over time, for $n_g = 10$ with variances possibly increasing across groups (left) and $n_g = 5$ with possibility for higher variance in bad groups (right)

D.3. Additional clinical trial simulation results (Sec 5.2)

Finally, we present additional clinical trial simulation results, which include more versions of the two algorithms and an additional setting with normal outcomes, in Table 3. We observe that the results with normal outcomes are largely in line with the results with binary outcomes. The relative performance of AdaGGI using different sampling strategies and AdaGCPI using different removal rules is also in line with what has been observed in Sec. 5.1 in the main text; in particular, $\mathcal{E}_{LCB}$ continues to dominate unless there is one group with a much better effect than others in which case $\mathcal{E}_{(L)UCB}$ works better, and the addition of $\mathcal{R}_{pop\text{-}fut}$ has the largest effect on AdaGCPIs performance whenever there is no effect across all groups.

$\boldsymbol{\theta}$	Method	Binary %Succ.	Binary $\lvert\mathcal{S}\rvert$	Binary $\frac{t_{stop}}{B}$	Binary $\frac{t_{1g}}{B}$	Binary $\frac{t_{1b}}{B}$	Normal %Succ.	Normal $\lvert\mathcal{S}\rvert$	Normal $\frac{t_{stop}}{B}$	Normal $\frac{t_{1g}}{B}$	Normal $\frac{t_{1b}}{B}$
A:$[0,0,0]$	GSDS	2.6	0.04	0.74		0.5	2.4	0.04	0.75		0.51
	AdaGGI($\mathcal{E}_{LCB}$)	0	0	0.64		0.24	0	0	0.69		**0.25**
	AdaGGI($\mathcal{E}_{UCB}$)	0	0	0.63		0.53	0	0	0.69		0.60
	AdaGGI($\mathcal{E}_{LUCB}$)	0	0	0.64		0.48	0	0	0.69		0.54
	AdaGGI($\mathcal{E}_{unif}$)	0	0	0.63		0.35	0	0	0.70		0.40
	AdaGCPI($\mathcal{R}_{fut}$)	0	0	0.64		0.36	0	0	0.69		0.39
	AdaGCPI ($\mathcal{R}_{fut}+\mathcal{R}_{pop\text{-}fut}$)	0	0	**0.49**		**0.23**	0	0	**0.54**		0.26
B:$[-0.2,0,0.2]$	GSDS	**99.3**	1.19	0.64	0.64	0.5	**97.9**	1.18	0.68	0.68	0.5
	AdaGGI($\mathcal{E}_{LCB}$)	97.9	0.98	0.63	**0.46**	0.38	96.6	1	0.69	0.52	0.57
	AdaGGI($\mathcal{E}_{UCB}$)	98	0.98	0.63	0.48	0.51	96.4	0.96	0.69	0.52	0.8
	AdaGGI($\mathcal{E}_{LUCB}$)	98.4	0.98	0.63	0.48	0.52	96.7	0.97	0.68	**0.51**	0.57
	AdaGGI($\mathcal{E}_{unif}$)	96	0.96	0.64	0.61	**0.15**	90.2	0.90	0.70	0.64	**0.16**
	AdaGCPI($\mathcal{R}_{fut}$)	96	1.05	0.63	0.62	**0.15**	91.9	0.994	0.68	0.66	**0.16**
	AdaGCPI ($\mathcal{R}_{fut}+\mathcal{R}_{pop\text{-}fut}$)	95	1.04	**0.61**	0.61	**0.15**	92	0.97	**0.67**	0.65	**0.16**
C:$[0,0.1,0.3]$	GSDS	**100**	2.03	**0.50**	0.50	0.50	**100**	1.98	**0.51**	0.51	0.5
	AdaGGI($\mathcal{E}_{LCB}$)	99	1.00	0.55	0.29	0.59	79	0.87	0.93	0.34	0.57
	AdaGGI($\mathcal{E}_{UCB}$)	**100**	1.09	0.90	**0.25**	0.81	**100**	1.08	0.93	**0.29**	0.87
	AdaGGI($\mathcal{E}_{LUCB}$)	99.9	1.08	0.90	0.26	0.83	**100**	1.06	0.93	**0.29**	0.85
	AdaGGI($\mathcal{E}_{unif}$)	99.6	1.06	0.91	0.45	0.53	98.5	1.03	0.96	0.65	0.59
	AdaGCPI($\mathcal{R}_{fut}$)	99.3	2.28	0.55	0.55	0.53	97.7	2.25	0.6	0.59	0.47
	AdaGCPI ($\mathcal{R}_{fut}+\mathcal{R}_{pop\text{-}fut}$)	89	2.28	0.55	0.55	**0.44**	0.98	2.26	0.59	0.59	**0.46**
D:$[0.2,0.2,0.2]$	GSDS	**100**	2.98	0.50	0.5		**100**	2.97	0.5	0.5	
	AdaGGI($\mathcal{E}_{LCB}$)	99.8	2.27	0.94	**0.36**		99.7	2.06	0.96	**0.4**	
	AdaGGI($\mathcal{E}_{UCB}$)	93.8	2.02	0.94	0.53		91.4	1.81	0.96	0.53	
	AdaGGI($\mathcal{E}_{LUCB}$)	95.9	2.07	0.94	0.51		93.5	1.83	0.96	0.54	
	AdaGGI($\mathcal{E}_{unif}$)	83	1.76	0.94	0.65		75.1	1.48	0.96	0.65	
	AdaGCPI($\mathcal{R}_{fut}$)	99.7	2.97	0.37	0.37		99.7	2.99	0.41	0.41	
	AdaGCPI ($\mathcal{R}_{fut}+\mathcal{R}_{pop\text{-}fut}$)	99.8	2.99	**0.37**	0.37		99.7	2.98	**0.41**	**0.4**	
E:$[0.3,0.3,0.3]$	GSDS	100	3	0.5	0.5		100	3	0.5	0.5	
	AdaGGI($\mathcal{E}_{LCB}$)	100	3	0.49	**0.16**		100	3	0.53	**0.18**	
	AdaGGI($\mathcal{E}_{UCB}$)	100	3	0.49	0.25		100	3	0.53	0.26	
	AdaGGI($\mathcal{E}_{LUCB}$)	100	3	0.49	0.24		100	3	0.53	0.26	
	AdaGGI($\mathcal{E}_{unif}$)	100	3	0.49	0.33		100	3	0.53	0.34	
	AdaGCPI($\mathcal{R}_{fut}$)	100	3	0.17	0.17		100	3	0.18	**0.18**	
	AdaGCPI ($\mathcal{R}_{fut}+\mathcal{R}_{pop\text{-}fut}$)	100	3	**0.17**	0.17		100	3	**0.18**	**0.18**	

Column legend: (1) %Succ. : prop. of trials which found a significant effect in *some* group. (2) $|\mathcal{S}|$: Average size of discovered subpopulation $\mathcal{S}$. (3) t_{stop}/B: Average algorithm termination time (as prop. of budget). (4) t_{1g}/B: Average time it took to identify the *first* good arm (as prop. of budget). (5) t_{1b}/B: Average time it took to discard the *first* bad arm (as prop. of budget).

Table 3. Results for simulated clinical trials with binary outcomes (left) and normal outcomes (right) using different treatment effect vectors θ; averaged across 1000 replications.