Adaptively Identifying Patient Populations With Treatment Benefit in Clinical Trials

Alicia Curth
University of Cambridge
amc253@cam.ac.uk

Alihan Hüyük
University of Cambridge
ah2075@cam.ac.uk

Mihaela van der Schaar
University of Cambridge
UCLA
The Alan Turing Institute
mv472@cam.ac.uk

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, often referred to as adaptive enrichment design, has been thoroughly studied in biostatistics with a focus on a limited number of subgroups (typically two) which make up (sub)populations, and a small number of interim analysis points. In this paper, 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, which focus on identifying good subgroups and good composite subpopulations, respectively. 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 ubiquitous in practice and include differences in treatment responses in cancer patients with specific mutations [4], psychiatric 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] 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).

arXiv:2208.05844v1 [stat.ML] 11 Aug 2022
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 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 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 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 adaptively identifying patient subpopulations that benefit from a given treatment during a clinical trial using tools from ML and make three main contributions: (1) Problem formalization and understanding: A large part of this paper is dedicated to formalising, contextualizing and understanding the population identification problem as a machine learning problem. Through this, we find that there are 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 meta-algorithm for each scenario (see Fig. 1): AdaGGI, which constructs a subpopulation by successively discovering individual subgroups with treatment benefit, and and AdaGCPI, which proceeds by successively eliminating subgroups from the full population until the average treatment effect across the leftover composite subpopulation is satisfactory. (3) Empirical Insight: We empirically investigate and provide insight into the (dis)advantages of either formulation 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 $\Omega_0$. We assume further that $\Omega_0$ is made up of $K$ disjoint and prespecified subgroups $\Omega_1, \ldots, \Omega_K$ 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 $\theta_j$ denote the
treatment’s effect (relative to control) within subgroup $j$, and let $\pi_j$ denote the prevalence of subgroup $j$ in the population.

**Goal.** To ensure success of the clinical trial, we aim to adaptively construct a composite subpopulation composed of a subset $\mathcal{S} \subseteq \mathcal{K} = \{1, \ldots, K\}$ of the full population with $\Omega_S = \bigcup_{j \in \mathcal{S}} \Omega_j$, in which the treatment is effective; that is we aim to find a subpopulation $\mathcal{S}$ with $\theta_S = \sum_{j \in \mathcal{S}} \pi_j \theta_j > 0$ (if any exists); we will refer to such subpopulations as good. 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 consider identifying individual good subgroups, i.e. find subgroups for which we can reject the null hypothesis

$$H_{0j} : \theta_j = 0 \quad (1)$$

for the 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. $\theta_{S^a} > 0$. We will 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 subpopulation 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

$$H_{0S} : \theta_S = 0 \quad (2)$$

for the alternative $H_{aS} : \theta_S > 0$. We will 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 $H_{0S}$ does not require rejecting $H_{0j}$ for every $j \in \mathcal{S}$.

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

$$\text{FWER}(A; \mathcal{P}) = \sup_{\rho \in \mathcal{P}} \Pr_{\rho}(A \text{ rejects a true null hypothesis}) \quad (3)$$

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

$$\text{FWER}_{GGI}(A; \mathcal{P}) = \sup_{\rho \in \mathcal{P}} \sum_{j=1}^{\mathcal{K}} \Pr_{\rho}(A \text{ rejects true } H_{0j}) \quad (4)$$

$$\text{FWER}_{GCPI}(A; \mathcal{P}) = \sup_{\rho \in \mathcal{P}} \sum_{\mathcal{S} \subseteq \mathcal{K}} \Pr_{\rho}(A \text{ selects subpopulation } \mathcal{S} \text{ and rejects true } H_{0S}) \quad (5)$$

**Power and minimum relevant effect.** Clinical trial designs are usually optimized for power; i.e. the ability to avoid Type 2 error (the failure to detect an effect when it does exist). Because the sample size needed to differentiate $\theta_0 = 0$ from $\theta_j > 0$ scales as $\theta_j^{-2}$, clinical trials often introduce an additional parameter, the minimum clinically relevant difference $\theta_{\min} > \theta_0 = 0$ which a trial should be powered to detect [22]. That is, we aim to ensure that $\Pr(H_{0S} \text{ is not rejected } | \theta_S = \theta_{\min}) \leq \beta$ for some $\beta \in (0,1)$, where $1 - \beta$ is usually referred to as the power of the trial.

**Mode of environment interaction, data structure and estimators.** 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 $t \in \{1, \ldots, B\}$, where $2B$ is the total patient budget of the trial, the algorithm selects a subgroup $J_t \in \mathcal{K}$ to enroll two patients from, which are then randomly assigned to one of each
treatment and control arm. This gives rise to control and treated outcome $Y^C, Y^T \in \mathcal{Y}$, which could be continuous ($\mathcal{Y} = \mathbb{R}$) or binary ($\mathcal{Y} = \{0, 1\}$), and produces a dataset of tuples $D_t = \{(J_t, Y^C_t, Y^T_t)\}_{t \leq t}$. We denote by $N_i(t) = \sum_{t < \tau} 1\{J_\tau = i\}$ and $N_S(t) = \sum_{t < \tau} 1\{J_\tau \in S\}$ the number of patient pairs enrolled from a subgroup or a subpopulation by time $t$, respectively.

Due to randomization and under standard assumptions such as no interference between patients, we have that $\theta_j = E[Y^T_t - Y^C_t | J_t = j]$, so that we can estimate treatment effects simply as

$$\hat{\theta}_{j, N_j(t)} = \frac{\sum_{t < \tau} 1\{J_\tau = j\}(Y^T_t - Y^C_t)}{N_j(t)}$$

Whenever all subgroups $i$ in a subpopulation $S$ were drawn according to their relative proportion $\frac{\sum_j \pi_j}{\sum_j \pi_j}$, we can also estimate $\hat{\theta}_{S, N_S(t)} = \frac{\sum_{t < \tau} 1\{J_\tau \in S\}(Y^T_t - Y^C_t)}{N_S(t)}$. Note that the $\hat{\theta}_{j, N_j(t)}$ will generally not be unbiased for $\theta_j$ as the $J_t$ were selected in a data-adaptive manner (see e.g. [23, 24]).

Finally, we therefore also assume that we have access to always-valid confidence intervals [25]; that is, similar to [26] we rely on existence of a function $\phi(t, \delta)$ which satisfies for any $\delta \in (0, 1)$ that $\mathbb{P}(\tau \leq \infty \mid \hat{\theta}_{S, t} - \theta_S \leq \phi(t, \delta)) \geq 1 - \delta$ and instantiate it using Thm. 8 of [27] which shows that for mean-zero $\sigma^2$-(sub)gaussian variables $X_t$, $\mathbb{P}(\exists t \in \mathbb{N} : \sum_{i=1}^t X_i > \sqrt{2 \sigma^2 \phi(t, \delta) \frac{t}{t}}) \leq \delta$ for $\phi(t, \delta) = \log(1/\delta) + 3 \log \log(1/\delta) + (3/2) \log \log(at/2)$ and $\delta \leq 0.1$. As discussed in more detail in Appendix C, we can use this as $\phi(\cdot, \cdot)$ in our experiments due to the fact that (i) the difference between two $\sigma^2$-(sub)gaussian variables is $2\sigma^2$-(sub)gaussian and (ii) Bernoulli variables are $1$-subgaussian.

### 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 focuses on finding members of the set $\mathcal{H}_a = \{j : \theta_j > 0\}$, subject to FWER-a-control and budget $2B$.

**Related work.** If $\theta_j$ 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]. 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 (both above and below the threshold) to be correct with fixed confidence $\delta$, 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.

---

1 More typical exploration problems, such as best/top-k arm identification (e.g. [29, 31]) are less relevant as our primary interest lies no in finding the group with the best response to a drug [20]; see also Appendix A.
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, which is orders of magnitude smaller than sample sizes considered in the ML literature. Due to both ethical and financial considerations, clinical trials are usually considered separately. Additionally, the available budget is usually very limited.

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 unlike a best arm identification problem where relative quality of an arm matters which needs substantially more exploration to identify). Third, we may want to focus on null hypotheses closest to rejection, recognizing that for a successful trial, rejecting one null hypothesis at level $\alpha$ is better than having two hypotheses only close to rejection upon termination.

Algorithm 1 AdaGGI

**Require:** $\alpha, \beta \in (0, 1), \theta_{\min} > 0$, budget $B$, initial samples $n_0$. Sampling rule $E$, ID. rule $I$, removal rule $R$.

1. Initialise: $A_{K_0} = K$; $\forall j \in K$, sample $n_0$ times & set $D_{K_0} = \{(S_{t_i}, Y_{t_i}^C, Y_{t_i}^T)\}_{t_i < K_0}$
2. for $t \in \{K_0 + 1, B\}$ do
3. Choose subgroup $J_t = E(D_{t-1}, A_{t-1})$ to enrol, set $D_t = D_{t-1} \cup (S_t, Y_t^C, Y_t^T)$
4. Identify good subgroups $S_t = S_{t-1} \cup I(D_t, \alpha)$, set $A_t = K \setminus S_t$
5. Remove bad groups: $A_t = K \setminus R(D_t, \theta_{\min}, \beta)$
6. if $A_t = \emptyset$, Output: True if $|S_B| > 0$, $S_B$
7. Output: True if $|S_B| > 0$, $S_B$

Algorithm 2 AdaGCPI

**Require:** $\alpha, \beta \in (0, 1), \theta_{\min} > 0$, budget $B$. ID. rule $I$, removal rule $R$.

1. Initialise: $A_1 = K$, set $D_0 = \emptyset$, $t = 0$
2. while $t < B$ do
3. Sample each $j \in A_t$, obtain $D_t = \{(j, Y_{t,j}^C, Y_{t,j}^T)\}_{j \in A_t}$, set $t' = |A_t|$, update $D_t$
4. Test for positive effect in current population $I(D_t, \alpha)$: if detected, Output: True, $A_t$
5. Remove bad groups: $A_t = K \setminus R(D_t, \theta_{\min}, \beta)$
6. if $A_t = \emptyset$, Output: False, $\emptyset$
7. Output: False, $\emptyset$

The backbones of the fixed confidence algorithms for identifying good bandit arms with mean above a threshold proposed in [18] [19] [20] [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 Adaptive Good subGroup Identification meta-algorithm, which is outlined in algorithm 1. Each iteration consists of (i) choosing a subgroup $J_t$ to enrol using an exploration strategy $E$, (ii) subsequently screening for new good subgroups using an $\alpha$-dependent identification criterion $I$ and (iii) removal of any groups demonstrating no minimum clinical benefit using a $(\beta, \theta_{\min})$-dependent removal criterion $R$. We describe all in more detail below.

---

2 When the goal is to identify all good groups, it matters how fast the last (most difficult to identify) group is found, while our goal of identifying many good groups quickly necessitates an initial focus on “easier” groups.
Sampling (exploration) strategies $\mathcal{E}$: Finding good arms fast. The established choice for sampling strategy $\mathcal{E}$ in the GAI literature [18, 19, 26] appears to be to use an optimistic upper-confidence bound (UCB) approach, i.e. $\mathcal{E}_{UCB}(D_{t-1}, \mathcal{A}_{t-1}) = \arg \max_{j \in \mathcal{A}_{t-1}} \hat{\theta}_{j,N,(t-1)} + \phi(N_j(t-1), \alpha)$. 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 $\mathbb{I} \{ \hat{\theta}_{j,N,(t)} - \phi(N_j(t), \epsilon) > 0 \}$ 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. Thus, we also consider using $\mathcal{E}_{LCB}(D_{t-1}, \mathcal{A}_{t-1}) = \arg \min_{j \in \mathcal{A}_{t-1}} \hat{\theta}_{j,N,(t-1)} - \phi(N_j(t-1), \alpha)$. 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.

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 $\mathcal{K}$, we rely on a simple Bonferroni correction here and use $\mathcal{I} \{ \cdot \} = \{ j \in \mathcal{K} : \hat{\theta}_{j,N,(t)} - \phi(N_j(t), \frac{\alpha}{|\mathcal{K}|}) > 0 \}$, which controls FWER as $\sum_{j \in \mathcal{K}, \theta_j = 0} \mathbb{P}(\cap_{t=1}^{\infty} \{ \hat{\theta}_{j,t} - \theta_j > \phi(t, \frac{\alpha}{|\mathcal{K}|}) \}) \leq \frac{\alpha}{|\mathcal{K}|}$. To create tighter confidence bounds in settings where many null hypotheses are false and recycling $\alpha$ from previously rejected hypotheses is thus possible, one could implement more sophisticated strategies based on the adapted Benjamini-Hochberg procedure from [26], or other $\alpha$-investing approaches similar to those discussed in [34].

Removal criterion: Focusing on clinically relevant effects. Finally, we employ removal criterion $\mathcal{R}_{fut}(D_t, \theta_{min}, \beta) = \{ j \in \mathcal{K} : \hat{\theta}_{j,N,(t)} + \phi(N_j(t), \beta) < \theta_{min} \}$. 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 in the GGI problem, we now move to the Good Composite subPopulation Identification (GCPI) problem which tackles the problem of 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 problem should be easier to solve — i.e. we would expect a smaller sample size to be required for a trial to be successful: given a $\mathcal{S}$, rejection of $H_{\mathcal{S}}$ 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 $\mathcal{K}$: 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 \textit{combinatorial bandit} problem \cite{29, 30}, where each subpopulation could be seen as a \textit{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. \cite{29, 30} would therefore resemble our GGI solution.

4.1 Unique problem characteristics and design considerations in the GCPI problem

\textbf{Unique characteristics of the GCPI objective.} Relative to the GGI problem, we consider two additional features key to the GCPI problem: On the one hand, the weaker requirement of identification of a positive \textit{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 \textit{combinatorial} and there are \(2^K\) possible subpopulations and null hypotheses, possibly making the problem harder.

\textbf{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 \textit{successive elimination} algorithms \cite{29, 30}, 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, uniformly allocating samples across all groups would lead to rejection of the \textit{full population} composite null hypothesis using the same expected number of samples that the GGI problem would need to identify a \textit{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 \textit{best arm} identification problem, where UCB-style algorithms empirically dominate successive elimination algorithms which are too wasteful in that context (see e.g. \cite{31}).

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 \textit{Adaptive Good Composite subPopulation Identification} 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 enrolling two patients from each. For ease of presentation we assume equal sized subgroups \((\pi_i = \frac{1}{K})\) 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 \textit{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.

\textbf{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 \textit{potentially} be tested. As we only select \textit{at most} \(K\) hypotheses for testing in practice, this adjustment is clearly overly conservative. To gain further intuition, let \(T_S\) denote whether hypothesis \(H_S\) is selected for testing at any time, and \(R_S\) whether it is independent of that used to determine rejection \(R_S\), we would have that 
\[
\text{FWER} \leq \mathbb{E}[\sum_{S,t} \mathbb{I}_{T_S=R_S}] 
\]
by Markov’s inequality. Further, 
\[
\mathbb{E}[\sum_{S,t} \mathbb{I}_{T_S=R_S}] = \sum_{S,t} \mathbb{E}[\mathbb{I}_{T_S=R_S} | T_S=1] \mathbb{P}(T_S=1) = \sum_{S,t} \mathbb{E}[\mathbb{I}_{T_S=R_S} | T_S=1] \mathbb{P}(T_S=1) 
\]
If the data used to determine hypothesis selection \(T_S\) was independent of that used to determine rejection \(R_S\), we would have that 
\[
\mathbb{E}[\sum_{S,t} \mathbb{I}_{T_S=R_S}] = \sum_{S,t} \mathbb{E}[\mathbb{I}_{T_S=R_S} | T_S=1] \mathbb{P}(T_S=1) = \sum_{S,t} \mathbb{E}[\mathbb{I}_{T_S=R_S} | T_S=1] \mathbb{P}(T_S=1) 
\]
so that 
\[
\frac{\mathbb{E}[\sum_{S} \mathbb{I}_{T_S=R_S}]}{\sum_{S,t} \mathbb{I}_{T_S=R_S}} \leq \frac{1}{\sum_{S,t} \mathbb{I}_{T_S=R_S}} \leq \frac{1}{K} 
\]
As at most \(K\) hypotheses will be tested. Clearly, \(T_S\) and \(R_S\) are not independent in our setting, so identification using \(T_S^K\) will not lead to exact FWER control. However, note that 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 our experiments (Appendix D), we observe that FWER-$\alpha$ seems to hold empirically when using $\mathcal{I}^{KF}_t$ (generally even on the conservative side, most likely due to the conservative nature of the anytime confidence intervals), so we rely on it in our implementations.

**Removal criterion: exploiting both subgroup and subpopulation signals.** Using criterion $R_{\text{fut}}(\mathcal{D}_t, \theta_{\text{min}}, \beta)$ as in AdaGGI, we remove individual subgroups for futility if their individual effects are insufficient. In addition, we exploit full subpopulation information by realizing that the event $F_t = \mathbb{I}\{\theta_{\mathcal{A}_t}, N_{\mathcal{A}_t}(t) + \phi(N_{\mathcal{A}_t}(t), \beta) < \theta_{\text{min}}\}$ provides evidence that at least one subgroup has no sufficient treatment effect. Thus, if $F_t$ is true, we remove the empirically worst subgroup through the rule $R_{\text{pop-fut}}(\mathcal{D}_t, \mathcal{A}_t, \theta_{\text{min}}, \beta) = \arg \min_{j \in \mathcal{A}_t} \theta_j, N_j(t) - \phi(N_j(t-1), \alpha)$ if $F_t$ 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 increases as Time to identify the we remove the empirically worst subgroup through the rule $R_{\text{pop-fut}}(\mathcal{D}_t, \mathcal{A}_t, \theta_{\text{min}}, \beta) = \arg \min_{j \in \mathcal{A}_t} \theta_j, N_j(t-1) - \phi(N_j(t-1), \alpha)$ if $F_t$ 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 $Y_j^q \sim N(\theta_j, \alpha)$ 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 $\theta_{\text{min}} = 0.5, \alpha = 0.05, \beta = 0.1$. 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 $\pi_g = \{|j : \theta_j \geq 0.5|\}$. Throughout, we do not restrict budget and report $t_{\text{stop}}$, the stopping time of the algorithm (i.e. the time when all subgroups are classified as good or not), as well as $i_{\text{d,j}}^{\text{good}}$ and $i_{\text{d,j}}^{\text{bad}}$, the time taken to identify the $j^{th}$ good group and to discard the $j^{th}$ bad group, respectively; doing so allows us to understand what the algorithm would have identified given any budget. We compare AdaGGI with different sampling strategies – $\mathcal{E}_{\text{UCB}}, \mathcal{E}_{\text{LCB}}$ and $\mathcal{E}_{\text{LUCB}}$ as discussed in Sec. 3.2 as well as two baselines (discussed further in Appendix A.2): $\mathcal{E}_{\text{unif}}$, which uniformly samples groups that have not yet been identified, and $\mathcal{E}_{\text{APT}}$, which corresponds to [13]'s thresholding bandit solution – to AdaGCPI with different removal strategies ($R_{\text{fut}}$ and $R_{\text{fut}} + R_{\text{pop-fut}}$). 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 $t_{\text{stop}}$. 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($R_{\text{fut}}$), which is identical to AdaGGI($\mathcal{E}_{\text{unif}}$) 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($R_{\text{fut}} + R_{\text{pop-fut}}$) 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 $\mathcal{E}_{\text{APT}}$, focussing on the groups hardest to distinguish from the threshold, performs even worse. Within the other adaptive strategies, $\mathcal{E}_{\text{LUCB}}$ indeed makes the first discoveries faster than $\mathcal{E}_{\text{UCB}}$ in this setting, as the latter will unnecessarily switch between good groups as upper bounds cross (because the underlying means are identical); as expected, $\mathcal{E}_{\text{LUCB}}$ lies inbetween.
Figure 2: Results describing time until (A) termination, (B&C) identification of good groups and (D& E) removal of bad groups, avg. across 1000 replications. (A): Time to termination $t_{stop}$ by number of good groups $n_g$. (B&C): Number of good group identifications over time, for $n_g = 4$ (B) and $n_g = 8$ (C). (D&E): Number of removals of bad groups over time, for $n_g = 2$ (D) and $n_g = 6$ (E).

If the good groups were to exhibit quantitatively very different effects, the group with the largest $\theta_j$ should need least samples to be discovered – thus we would expect UCB-type strategies that have 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: $\theta_j = 0.5 + 0.5(j-1), j \leq 8; \theta_j = 0, j > 8$), 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 $E_{LCB}$ can dominate as it intrinsically makes use of the fact that arms with lower variance need less samples to be identified, while $E_{UCB}$ 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 successive elimination – discards groups much faster than AdaGGI (with the exception of AdaGGI($E_{APT}$), which essentially acts like a more aggressive elimination algorithm due to its focus on the threshold). Second, we observe that AdaGCPI($R_{fut} + R_{pop-fut}$) 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 discarding 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 $\alpha$ but essentially 0 throughout (even when we remove the Bonferroni correction); we attribute this to the used anytime confidence intervals being unnecessarily conservative at $t \approx \infty$ here. Figure 4: (A): Avg. number of missed groups by $n_g$. (B & C): Avg. $|S|$ by $n_g$, for $\theta_0 = 0, -0.5$. 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 $\beta \cdot n_g$); only AdaGCPI occasionally removes a good group with the aggressive removal criterion $R_{pop-fut}$. Third, in Fig. 4B, we observe interesting differences in groups without effect that are included in the selected subpopulation $S$ (note: for AdaGCPi, this does not necessarily constitute a Type I error as long as $\theta_0 > 0$). As AdaGGI identifies groups individually, $|S| \approx n_g$ throughout, while AdaGCPI allows free-riding of groups without effect on the larger effects of other groups, i.e. $|S| > n_g$, 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. 1C 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 with three subgroups

Finally, we study a clinical trial setup inspired by the one presented in Section 6 of [3]. We compare AdaGCPI and AdaGGI to [3]’s proposed GSDS procedure as a baseline, which is structured similarly to AdaGCPI but (i) allows only \( n_a \) (prespecified) interim analyses (in their study and here \( n_a = 2 \)), (ii) selects and fixes subpopulation \( S \) at the first interim analysis and (iii) relies on explicitly calculated normal error-spending boundaries. We consider 3 equal sized subgroups with unknown treatment effect vector \( \theta = [\theta_1, \theta_2, \theta_3] \) and as [3] let \( \theta_{\text{min}} = 0.2, \alpha = 0.025 \) and \( \beta = 0.1 \). Their setup considers binary outcomes \( (Y^C \sim B(\mu_{0,j}), Y_j^f \sim B(\mu_{0,j} + \theta_j)) \); in Appendix D we also consider normal outcomes. Using their budget calculations we set a budget of \( B = 800 \) pairs of patients, of which GSDS uses half in each stage. GSDS and the simulation are further described in Appendix C.

The original experiment in [3] has \( \theta \approx [0, 0.05, 0.1] \), i.e. all \( \theta_j < \theta_{\text{min}} \), 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 declare success only in 13\% and 7\% -- a direct consequence of our designs discarding effects below the minimum clinically relevant \( \theta_{\text{min}} \). To gain more interesting insights into relative performance, we therefore consider five scenarios with varying \( \theta \) in Table 1.



\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|}
\hline
Scenario: \( \theta \) & Method & \%Succ. & \( t_{stop} \) & \( t_{FUT} \) & \( R_\beta \) \\
\hline
A: [0, 0.0] & GSDS & 2.6 & 0.04 & 0.74 & 0.37 \\
 & AdaGGI & 0 & 0 & 0.64 & 0.24 \\
 & AdaGCPI & 0 & 0 & 0.49 & 0.23 \\
\hline
B: [0, 0.2, 0.2] & GSDS & 99.3 & 1.19 & 0.64 & 0.64 \\
 & AdaGGI & 97.9 & 0.98 & 0.63 & 0.46 \\
 & AdaGCPI & 95 & 1.04 & 0.61 & 0.15 \\
\hline
C: [0.1, 0.3] & GSDS & 100 & 2.03 & 0.50 & 0.56 \\
 & AdaGGI & 99 & 1.00 & 0.55 & 0.29 \\
 & AdaGCPI & 89 & 2.28 & 0.89 & 0.55 \\
\hline
D: [0.2, 0.2, 0.2] & GSDS & 100 & 2.98 & 0.50 & 0.54 \\
 & AdaGGI & 99.8 & 2.27 & 0.94 & 0.36 \\
 & AdaGCPI & 99.8 & 2.99 & 0.37 & 0.37 \\
\hline
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 \\
\hline
\end{tabular}
\caption{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.}
\end{table}

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 \( \theta_{\text{min}} \) 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 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 \( \theta_j > \theta_{\text{min}}, \) 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.

6 Conclusion

We investigated how to adaptively identify patient subpopulations with treatment benefit during a clinical trial, 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. We believe that this paper opens up many interesting avenues for future research, which we discuss in Appendix B.

\[ \text{Here, we focus on comparison with GSDS and use Sec. 5.1’s all-around best versions, AdaGGI(\text{E}_{LCD}) and AdaGCPI(\text{R}_{fut} + \text{R}_{pop-fut}), as AdaGGI and AdaGCPI. Full results are presented in Appendix D.} \]
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 treatment 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. [12] discusses similar issues for enrichment designs more generally.

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 Naita 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, Satul 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] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

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

[51] Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.
[54] Séebastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In International Conference on Machine Learning, pages 258–265. PMLR, 2013.

[55] 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.

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

[57] 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.

[58] 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.

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

[60] 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.

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

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

[63] 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.

[64] 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.

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

[66] 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.

[67] 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.

[68] 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.

[69] Aditya Grover, Todor Markov, Peter Attia, Norman Jin, Nicolas Perkins, Bryan Cheung, Michael Chen, Zi Yang, Stephen Harris, William Chuah, 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.

[70] 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.

[71] 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.
Appendix

This appendix is structured as follows: Section A contains an additional literature review, which consists of (i) an extended review of the adaptive clinical trial literature on enrichment designs and (ii) an extended contextualization of the GGI and GCPI problems within the bandit literature. Section B discusses possible extensions and future work. Section C contains details on the experiments performed, and Section D presents additional results.

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.\[43\]\[46\]. 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 \[47\]. \[48\] also consider two-stage trials under different types of restrictions with multiple nested subpopulations determined by biomarker interactions.

\[35\]\[36\]\[49\] all consider a setting where one can either consider the full population or a single pre-specified subgroup 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 \[50\] 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 only 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 most similar to good arm identification under a fixed budget (only) setting.

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. \[51\]). 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.

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, 54, 55], 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, 56, 61]. 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, 62] 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 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 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

| Problem            | Ref. | Type of arms identified | Number of arms identified | Budget | Confidence | Formulation |
|--------------------|------|-------------------------|---------------------------|--------|------------|-------------|
| BAI w/ FB          | [29] | Best arms               | Variable                  | Variable | Minimized  | Fixed (T)   |
| BAI w/ FC          | [30] | Top-K arms              | Fixed (T)                 | Fixed (1 − δ) |            |             |
| GAI w/ FB          | [14] | Good arms               | Fixed (T)                 | Maximized | Fixed (1 − δ) | Minimized |
| GAI w/ FC          | [19] | Good composite arms     | Fixed (T)                 | Maximized | Fixed (1 − δ) | Minimized |
| GGI (Ours)         |      | All good arms           | Fixed (T)                 | Fixed (1 − δ) | Minimized |             |
| GCPI (Ours)        |      | Good composite arms     | Fixed (T)                 | Maximized | Fixed (1 − δ) | Minimized |

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.
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, 20] 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 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 \(\mathcal{E}_{APT}(D_{t-1}, A_{t-1}) = \arg\min_{j\in A_{t-1}} \sqrt{N_j(t-1)}(\hat{\theta}_{j,N_j(t-1)} - \theta_0)\) 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 \(\mathcal{E}_{UCB}\) tries to accomplish and cannot be expected to perform well in our context. Because the original paper [14] focusses on a fixed 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 [64]'s Combinatorial Upper Confidence Bound (CUCB) algorithm in more detail as it permits straightforward adaptation to our setting. The general setting considered in [64] allows to play a super-arm \(S\) at each time \(t\), and their algorithm assumes existence of an oracle that outputs the optimal \(S\) 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 \(\hat{\theta}_{j,t} = \theta_{j,N_j(t-1)} + \phi(N_j(t-1), \beta)\) on the means of all arms, and then applies the oracle to the \(\hat{\theta}_{j,t}\), outputting a super-arm \(S_t\) to sample. In our context, this would sample all arms for which it holds that \(\theta_{j,N_j(t-1)} + \phi(N_j(t-1), \beta) > \theta_0\). Note that this essentially corresponds to AdaGGI with removal criterion \(R_{\text{unif}}(D_t, \theta_0, \beta)\) instead of \(R_{\text{mut}}(D_t, \theta_{\text{min}}, \beta)\), and uniform sampling of the active set. As discussed above, setting \(\theta_{\text{min}} \neq \theta_0\) 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**

**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 [65]. AdaGCPI could directly accommodate this: instead of recruiting \(|A_t|\) 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 [66, 67] – 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 [68], while the bandit literature has developed approaches for decision making under delayed feedback [69]; 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. [70]). 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 [43–46]; 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 \( B < \infty \) and to therefore construct alternatives that instead of allowing for infinitely many peaks at the data, allow only \( L \leq B \) decision points which may lead to less necessity to be conservative.

**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, \( \pi_j = \frac{1}{K}, \forall j \in K \), and outcomes are normally distributed according to \( N(\theta_k, 1) \), 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 set 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 \( Y^\theta \) is assumed normally distributed with known variance \( \sigma^2 = 1 \), we use \( \phi(t, \delta) = \ldots \)
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 [71]. The assumed end point of interest is the occurrence of pathologic complete response (pCR), patients, and not patients. We have adapted budget calculations accordingly and with

$$\hat{\theta} = \theta$$

denote the means in treated and control arm, we have

$$\sqrt{\frac{2\log(1/\delta)+3\log\log(1/\delta)+(3/2)\log\log(\text{et}/2)}{t}}$$

as in [26].

In their example with binary outcomes, if we let $$Y_j^T \sim \mathcal{B}(0.4)$$ 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] 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 $\sigma^2$ is normal with variance $2\sigma^2$, we use $\phi(t, \delta) = 2\sqrt{\frac{\log(1/\delta)+3\log\log(1/\delta)+(3/2)\log\log(\text{et}/2)}{t}}$ for the normal outcomes, and, as bernoulli variables are $\frac{1}{2}$ subgaussian, we use $\phi(t, \delta) = \sqrt{\frac{\log(1/\delta)+3\log\log(1/\delta)+(3/2)\log\log(\text{et}/2)}{t}}$ 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 $n_j$, a test statistic $Y_j(t)$ and associated Fisher information $I_j(t)$, a desired significance level $\alpha$ and power $1 - \beta, \delta$ is used to calculate stopping boundaries $\{(l_p, u_p)\}^n_{p=\alpha}$, for each interim analysis. $\beta$ is used to calculate a maximum information level $I_{max}$, 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: $S^* = \{j \in K : Y_j(t_1)\sqrt{I_j(t_1)} > l_1\}$. If $Y_j(t_1)\sqrt{I_j(t_1)} < u_1$, the trial terminates immediately for futility; otherwise the trial continues and at all $n_a - 1$ subsequent stages, the trial is terminated for futility if $Y_j(t_k)\sqrt{I_j(t_k)} > u_k$ and terminated for futility if $Y_j(t_k)\sqrt{I_j(t_k)} < l_k$.

**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 $I_{max} = 1495.5$.

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

$$Y = \hat{p}^T - \hat{p}^C$$

and $I = \frac{b}{2\hat{p}(1-\hat{p})}$

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

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

$$Y = \hat{\mu}^T - \hat{\mu}^C$$

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

---

5We 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.
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.](image1)

![Figure 6: Results describing removal of bad groups over time, for $n_g \in \{0, 2, 4, 6, 8\}$; avg. across 1000 replications.](image2)

D.1.2 Type I error

In Fig. 7, we plot Type I errors 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 $S$ 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 Type I errors remain below $(10 - n_g) \star \alpha$ (in fact, they even lie below $\alpha$). 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 $K$).

D.2 Additional simulation scenarios

Varying means We present additional results on the setting presented in Fig. 3 of the main text: for $n_g \in \{2, \ldots, 10\}$ 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: \(E_{\text{LCB}}\) generally performs worse than \(E_{\text{UCB}}\) here; with increasing \(n_g\) and hence decreasing spacing between the good means, this effect reduces.

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 \(E_{\text{UCB}}\) is much larger than that for \(E_{\text{LCB}}\) 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.
Finally, we present additional clinical trial simulation results, which include all 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 $\mathcal{E}_{LUCB}$ works better, and the addition of $R_{\text{pop-fut}}$ has
the largest effect on AdaGCPIs performance whenever there is no effect across all groups.

| $\theta$ | Method | Binary | Normal |
|---------|--------|--------|--------|
| A: $[0,0,0]$ | GSDDS | $2.6$ | $0.04$ | $0.74$ | $0.5$ | $2.4$ | $0.04$ | $0.75$ | $0.51$ |
| AdaGGI($\mathcal{E}_{LCB}$) | $0$ | $0$ | $0.64$ | $0.24$ | $0$ | $0$ | $0.69$ | $0.25$ |
| AdaGGI($\mathcal{E}_{LUCB}$) | $0$ | $0$ | $0.63$ | $0.53$ | $0$ | $0$ | $0.69$ | $0.60$ |
| AdaGGI($\mathcal{E}_{\text{unif}}$) | $0$ | $0$ | $0.64$ | $0.48$ | $0$ | $0$ | $0.69$ | $0.54$ |
| AdaGCPI($R_{\text{fut}}$) | $0$ | $0$ | $0.64$ | $0.36$ | $0$ | $0$ | $0.69$ | $0.39$ |
| AdaGCPI($R_{\text{fut}} + R_{\text{pop-fut}}$) | $0$ | $0$ | $0.49$ | $0.23$ | $0$ | $0$ | $0.54$ | $0.26$ |
| B: $[-0.2,0.2,0]$ | GSDDS | $99.3$ | $1.19$ | $0.64$ | $0.64$ | $0.5$ | $97.9$ | $1.18$ | $0.68$ | $0.68$ | $0.5$ |
| AdaGGI($\mathcal{E}_{LCB}$) | $97.9$ | $0.98$ | $0.63$ | $0.46$ | $0.38$ | $96.6$ | $1$ | $0.69$ | $0.52$ | $0.57$ |
| AdaGGI($\mathcal{E}_{LUCB}$) | $98.4$ | $0.98$ | $0.63$ | $0.48$ | $0.51$ | $96.4$ | $0.96$ | $0.69$ | $0.62$ | $0.8$ |
| AdaGGI($\mathcal{E}_{\text{unif}}$) | $96$ | $0.96$ | $0.64$ | $0.61$ | $0.15$ | $90.2$ | $0.90$ | $0.70$ | $0.64$ | $0.16$ |
| AdaGCPI($R_{\text{fut}}$) | $96$ | $1.05$ | $0.63$ | $0.62$ | $0.15$ | $91.9$ | $0.994$ | $0.68$ | $0.66$ | $0.16$ |
| AdaGCPI($R_{\text{fut}} + R_{\text{pop-fut}}$) | $95$ | $1.04$ | $0.61$ | $0.61$ | $0.15$ | $92$ | $0.97$ | $0.67$ | $0.65$ | $0.16$ |
| C: $[0,0.1,0.3]$ | GSDDS | $100$ | $2.03$ | $0.50$ | $0.50$ | $0.50$ | $100$ | $1.98$ | $0.51$ | $0.51$ | $0.5$ |
| AdaGGI($\mathcal{E}_{LCB}$) | $99$ | $1.00$ | $0.55$ | $0.29$ | $0.59$ | $79$ | $0.87$ | $0.93$ | $0.34$ | $0.57$ |
| AdaGGI($\mathcal{E}_{LUCB}$) | $100$ | $1.09$ | $0.90$ | $0.25$ | $0.81$ | $100$ | $1.08$ | $0.93$ | $0.29$ | $0.87$ |
| AdaGGI($\mathcal{E}_{\text{unif}}$) | $99.9$ | $1.08$ | $0.90$ | $0.26$ | $0.83$ | $100$ | $1.06$ | $0.93$ | $0.29$ | $0.85$ |
| AdaGCPI($R_{\text{fut}}$) | $99.3$ | $2.28$ | $0.55$ | $0.55$ | $0.53$ | $98.5$ | $1.03$ | $0.96$ | $0.65$ | $0.59$ |
| AdaGCPI($R_{\text{fut}} + R_{\text{pop-fut}}$) | $89$ | $2.28$ | $0.89$ | $0.55$ | $0.44$ | $0.98$ | $2.26$ | $0.59$ | $0.59$ | $0.46$ |
| D: $[0.2,0.2,0.2]$ | GSDDS | $100$ | $2.98$ | $0.50$ | $0.5$ | $100$ | $2.97$ | $0.5$ | $0.5$ | $0.5$ |
| AdaGGI($\mathcal{E}_{LCB}$) | $99.8$ | $2.27$ | $0.94$ | $0.36$ | $100$ | $2.06$ | $0.96$ | $0.4$ | $0.5$ |
| AdaGGI($\mathcal{E}_{LUCB}$) | $95.9$ | $2.07$ | $0.94$ | $0.51$ | $93.5$ | $1.83$ | $0.96$ | $0.54$ | $0.5$ |
| AdaGGI($\mathcal{E}_{\text{unif}}$) | $83$ | $1.76$ | $0.94$ | $0.65$ | $75.1$ | $1.48$ | $0.96$ | $0.65$ | $0.04$ |
| AdaGCPI($R_{\text{fut}}$) | $99.7$ | $2.97$ | $0.37$ | $0.37$ | $99.7$ | $2.99$ | $0.41$ | $0.41$ | $0.41$ |
| AdaGCPI($R_{\text{fut}} + R_{\text{pop-fut}}$) | $99.8$ | $2.99$ | $0.37$ | $0.37$ | $99.7$ | $2.98$ | $0.41$ | $0.4$ | $0.4$ |
| E: $[0.3,0.3,0.3]$ | GSDDS | $100$ | $3$ | $0.5$ | $0.5$ | $100$ | $3$ | $0.5$ | $0.5$ | $0.5$ |
| AdaGGI($\mathcal{E}_{LCB}$) | $100$ | $3$ | $0.49$ | $0.16$ | $100$ | $3$ | $0.53$ | $0.18$ | $0.18$ |
| AdaGGI($\mathcal{E}_{LUCB}$) | $100$ | $3$ | $0.49$ | $0.25$ | $100$ | $3$ | $0.53$ | $0.26$ | $0.26$ |
| AdaGGI($\mathcal{E}_{\text{unif}}$) | $100$ | $3$ | $0.49$ | $0.24$ | $100$ | $3$ | $0.53$ | $0.26$ | $0.26$ |
| AdaGCPI($R_{\text{fut}}$) | $100$ | $3$ | $0.49$ | $0.33$ | $100$ | $3$ | $0.53$ | $0.34$ | $0.34$ |
| AdaGCPI($R_{\text{fut}} + R_{\text{pop-fut}}$) | $100$ | $3$ | $0.17$ | $0.17$ | $100$ | $3$ | $0.18$ | $0.18$ | $0.18$ |

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

Column legend: (1) %Succ. : prop. of trials which found a significant effect in some group. (2) $|S|$: Average size of discovered subpopulation $S$. (3) $t_{\text{stop}}/B$: 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).