Adaptively Identifying *Good* Patient Populations 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 ATI)* MV472@CAM.AC.UK

Abstract

We study the problem of adaptively identifying good patient subpopulations for 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 *good* subpopulations (and not necessarily only the *single best* subgroup) 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, and empirically investigate their (dis)advantages.

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. 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).

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]. 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. Building on these insights, we propose two meta-algorithms – AdaGGI, which constructs a subpopulation by successively discovering individual good subgroups, 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 – and investigate their (dis)advantages empirically.

2. Problem Setup

Objective. We aim to run a trial to establish efficacy of a novel drug (T) relative to an established control (C) in patient population Ω_0, which consists of K disjoint and prespecified subgroups $\Omega_1, \dots, \Omega_K$ where $\Omega_0 = \cup_{j \leq K} \Omega_j$, across which efficacy is expected to differ, e.g. due to known biological pathways. Let θ_j denote the treatment's effect in subgroup j, and let π_j denote the prevalence of Ω_j in Ω_0. To ensure success of the trial, we aim to construct a *composite subpopulation* composed of $\mathcal{S} \subseteq \mathcal{K} = \{1, \dots, K\}$, in which the treatment is *effective*; i.e. find a subpopulation $\mathcal{S}$ for which $\theta_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \frac{\pi_i}{\sum_{j \in \mathcal{S}} \pi_j} \theta_i > 0$ (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. First, we consider constructing subpopulations by identifying individual good *subgroups*, i.e. find subgroups for which we can reject the null hypothesis $H_{0j} : \theta_j = 0$ for the alternative $H_{aj} : \theta_j > 0$. We will refer to this problem as the *Good subGroup Identification* (GGI) problem. Often, however, trials are not powered to detect effects in subgroups separately; instead, the focus is set on demonstrating *average* effectiveness across a sub*population* as in [3]. Second, we therefore consider identifying a *composite* subpopulation $\mathcal{S}$ for which we can only prove that the treatment is effective *on average*, i.e. reject $H_{0\mathcal{S}} : \theta_{\mathcal{S}} = 0$ for the alternative $H_{a\mathcal{S}} : \theta_{\mathcal{S}} > 0$. 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 $H_{0\mathcal{S}}$ does not require rejecting H_{0j} for every $j \in \mathcal{S}$.

Error control. Regulators usually require control of the probability of Type 1 errors [20] as captured by the familywise error rate (FWER), which is defined for an algorithm $\mathcal{A}$ across problem instances $\mathcal{P}$ as $\text{FWER}(\mathcal{A}; \mathcal{P}) = \sup_{\rho \in \mathcal{P}} \mathbb{P}_\rho(\mathcal{A} \text{ rejects a true null hypothesis})$. FWER-control at level $\alpha \in (0, 1)$ requires that $\text{FWER}(\mathcal{A}; \mathcal{P}) \leq \alpha$. Further, 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 θ_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 [21]. That is, we aim to ensure that $\mathbb{P}(H_{0\mathcal{S}} \text{ is not rejected } | \theta_{\mathcal{S}} = \theta_{min}) \leq \beta$ for some $\beta \in (0,1)$, where $1-\beta$ is usually referred to as the power of the trial.

Environment, data structure and estimators. We assume the stylized setting of an unlimited stream of patients available for recruitment from each subgroup, where outcomes are revealed 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, a subgroup $J_t \in \mathcal{K}$ is selected 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 $Y_t^C, Y_t^T \in \mathcal{Y}$, 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. Due to randomization and under standard assumptions such as *no interference* between patients, we have that $\theta_j = \mathbb{E}[Y_t^T - Y_t^C | S_t = j]$, so that we can estimate treatment effects simply as $\hat{\theta}_{j,N_j(t)} = N_j(t)^{-1} \sum_{t'=1}^{t} \mathbb{1}\{J_{t'} = j\}(Y_{t'}^T - Y_{t'}^C)$. Whenever all subgroups i in $\mathcal{S}$ were drawn according to their relative proportion $\pi_i / \sum_{j \in \mathcal{S}} \pi_j$, we can also estimate $\hat{\theta}_{\mathcal{S},N_{\mathcal{S}}(t)} = N_{\mathcal{S}}(t)^{-1} \sum_{t'=1}^{t} \mathbb{1}\{J_{t'} \in \mathcal{S}\}(Y_{t'}^T - Y_{t'}^C)$. Finally, as [22] we assume access to *always valid confidence intervals* $\phi(t, \delta)$ which satisfy for any $\delta \in (0,1)$ that $\mathbb{P}(\cap_{t=1}^{\infty} \{|\hat{\theta}_{\mathcal{S},t} - \theta_{\mathcal{S}}| \leq \phi(t,\delta)\}) \geq 1 - \delta$, and instantiate them using Thm. 8 of [23].

3. The good subgroup identification problem

Related work. We begin by studying the GGI problem as it appears more closely related to problems studied in the recent ML literature: 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 [22, 24].[1] In addition to the difference in target of interest, a major difference between existing formulations and our problem are the underlying constraints: 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 δ (similarly in [18, 19, 22, 24]), while the fixed budget setting aims for the highest confidence in all classifications given a certain budget. Finally, [29] 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.

Unique characteristics & design considerations. Discovery in a clinical trial is subject to *both* budget *and* FWER constraints – combining the fixed confidence and

1. More typical exploration problems, such as best/top-k arm identification (e.g. [25–27]) are less relevant as our primary interest lies no in finding the group with *the best* response to a drug [28]

fixed budget setting that are usually considered separately. Further, the available budget is usually *very limited* relative to e.g. online advertising applications: confirmatory Phase 3 Trials usually enrol between 300-3000 patients [30]. Because it is *not* necessary to make a judgement about *all* subgroups immediately to meet our objective, it is thus advisable to *focus on promising groups*[2]. In particular, 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. 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.

3.1 Good subgroup identification using AdaGGI

We propose AdaGGI, an *Ada*ptive *G*ood sub*G*roup *I*dentification meta-algorithm. Its structure is inspired by fixed confidence GAI algorithms [18, 19, 22], but incorporates budget restrictions and other modifications: Until budget depletion, each iteration (i) uses sampling (exploration) rule $\mathcal{E}$ to choose a subgroup J_t from $\mathcal{A}_t$, the active set of unclassified subgroups, to enrol a patient pair from, (ii) screens for new good subgroups using α-dependent identification rule $\mathcal{I}$ and (iii) removes any groups demonstrating no clinical benefit using (β, θ_{min})-dependent removal rule $\mathcal{R}$. We discuss details below, pseudocode is in Appendix A.

Identification rule: Ensuring FWER control. Our identification rule needs to ensure that $FWER_{GGI} \leq \alpha$, adjusting for *multiple* hypothesis testing. We rely on a simple Bonferroni correction here[3] 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 $\sum_{j\in\mathcal{K}:\theta_j=0} \mathbb{P}(\cap_{t=1}^{\infty}\{\hat{\theta}_{j,t} - \theta_j > \phi(t, \frac{\alpha}{K})\} \leq K\frac{\alpha}{K}$.

Sampling rule $\mathcal{E}$: Finding good arms fast. The established sampling rule in the GAI literature [18, 19, 22] appears to be to use an upper-confidence bound (UCB) – which will not necessarily sample a subgroup whose null is closest to being rejected[4]. Instead, we thus propose to sample according to the best lower confidence bound (LCB), which corresponds to selecting groups that appear most promising for early identification: $\mathcal{E}_{LCB}(D_{t-1}, \mathcal{A}_{t-1}) = \arg\max_{j\in\mathcal{A}_{t-1}} \hat{\theta}_{j,N_j(t-1)} - \phi(N_j(t-1), \alpha)$.

Removal criterion: Focusing on clinically relevant effects. Finally, 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 allows requiring less evidence for discarding a "bad" subgroup than for accepting a good one. This differs from the recent GAI literature, where arms are either discarded and accepted using the same rule [18] or not discarded at all [19, 22].

4. The good composite subpopulation identification problem

Related work. Most work on adaptive enrichment designs considers a simplified version of the GCPI problem, where $\mathcal{K} = \{1, 2\}$. Initially patients from both subgroups are enrolled; at a single [2, 32, 33] or multiple [6, 34] prespecified interim analysis points it is then possible

2. This is contrary to thresholding bandits [14–16] which focus explicitly on the *hardest* to find arms.
3. To be less conservative in settings where *many* null hypotheses are false, one could use more sophisticated strategies e.g. [22]'s adapted Benjamini-Hochberg procedure, or α-investing approaches [31].
4. As confidence intervals shrink in t, we suspect that UCB-sampling encourages switching between groups when multiple good groups are similar, possibly leading to no null rejections before budget depletes.

to discontinue either subgroup. [3]'s GSDS does not restrict K but fixes subgroups included in subpopulation $\mathcal{S}$ at the first interim analysis; subsequently only early termination of the *entire* subpopulation based on normal error-spending boundaries is allowed. From a bandit perspective, the GCPI problem can be interpreted as a generic *combinatorial bandit* problem [35, 36], where each subpopulation is 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. [35, 36] would therefore resemble our GGI solution.

Unique characteristics & design considerations. Relative to the GGI problem, GCPI has two interesting characteristics: First, the weaker requirement of establishing a positive *average* effect should make it possible to share statistical strength *across* subgroups. While the need to identify individual groups fast led us to consider non-uniform sampling schemes for GGI, this possibility thus makes *successive elimination* algorithms [25, 37], 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 would lead to rejection of the *full population* null hypothesis using the same expected number of samples that the GGI problem would need to identify a *a single* group[5]. Second, while GGI considers K separate subgroups/hypotheses, the GCPI problem is *combinatorial* as there are 2^K possible subpopulations (and null hypotheses). Also here, successive elimination lends itself as a solution as it substantially limits the number of subpopulations (and associated null hypotheses) to be considered – if subgroups are irreversibly eliminated, at most K (nested) subpopulations will be considered.

4.1 Good composite subpopulation identification using AdaGCPI

We propose AdaGCPI, an *Ada*ptive *G*ood *C*omposite sub*P*opulation *I*dentification meta-algorithm. Until budget is depleted, the algorithm proceeds by uniformly sampling all subgroups in the active set $\mathcal{A}_t$ by enroling two patients from each[6]. We then apply an identification criterion $\mathcal{I}$ that tests for evidence of an *average* positive subpopulation effect across $\mathcal{A}_t$. 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 rule below, pseudo-code is in Appendix A.

Identification rule: 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[7], 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,

5. Note that such potential efficiency of successive elimination in the GCPI problem stands in 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. [27]).
6. 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 $\mathcal{A}_t$ according to prevalence $\pi_j / \sum_{i \in \mathcal{A}_t} \pi_i$.
7. To gain 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 [31], 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.

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 rule: Using subgroup *and* subpopulation signals. As in AdaGGI, we remove subgroups for futility if their individual effects are insufficient using $\mathcal{R}_{fut}(\mathcal{D}_t, \theta_{min}, \beta)$. 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 $\mathcal{R}_{pop\text{-}fut}(\mathcal{D}_t, \mathcal{A}_t, \theta_{min}, \beta) = \arg\min_{j \in \mathcal{A}_t} \hat{\theta}_{j,N_j(t)} - \phi(N_j(t), \alpha)$ if $\mathcal{F}_t$ else $\emptyset$.

5. Experiments: Clinical Trial Simulation adapted from [3]

We compare AdaGCPI and AdaGGI to [3]'s GSDS with one interim analysis (as in [3]). We consider 3 equal sized subgroups with $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$ and as [3] let $\theta_{min} = 0.2, \alpha = 0.025$, $\beta = 0.1$ and $B = 800$. [3]'s 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. GSDS and the simulation setup are described further in Appendix C. The original experiment in [3] has $\boldsymbol{\theta} \approx [0, .05, .1]$, i.e. all $\theta_j < \theta_{min}$, so that no design is powered to detect any effect[8]. To gain more interesting insights into relative performance, we instead vary $\boldsymbol{\theta}$ in Table 1. In addition to trial success and $|\mathcal{S}|$, we examine stopping time of the algorithm (i.e. $t_{stop} \stackrel{\text{def}}{=} t : \mathcal{A}_t = \emptyset \vee t = B$), as well as $t_g^{id,1}(t_b^{id,1})$, the time taken to identify the 1^{st} good group (discard the 1^{st} bad group).

Table 1: Results of 1000 simulated trials.

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	

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 overly conservative (see Appendix D) – especially when compared to the exact normal boundaries used in GSDS. Nonetheless, compared to our fully adaptive approaches, GSDS suffers from its rigidity (i.e. being restricted to pre-specified analysis times). 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. Comparing the two, AdaGGI generally finds the first good group faster, while AdaGCPI discards the first bad subgroup faster and is able to stop much faster as it can exploit shared statistical strength both in accepting and discarding subgroups. In Scenarios A&E, where outcomes are extreme, the advantage of the flexibility of AdaGCPI relative to GSDS is most obvious, as, due to the lack of restriction on analysis times, AdaGCPI can terminate *much* earlier than GSDS's first scheduled interim analysis.

8. 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 minumum clinically relevant θ_{min}.

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] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

[26] 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.

[27] 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.

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

[29] 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.

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

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[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.

Appendix A. Pseudo-code and illustrations of Algorithms

A.1 Pseudo-code

Algorithm 1 AdaGGI

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

Initialise: $\mathcal{A}_{Kn_0} = \mathcal{K}$; $\forall j \in \mathcal{K}$, sample n_0 times & set $D_{Kn_0} = \{(S_{t'}, Y_{t'}^C, Y_{t'}^T\}_{t' \leq Kn_0}$

for $t \in \{Kn_0 + 1, B\}$ **do**

  Choose subgroup $J_t = \mathcal{E}(D_{t-1}, \mathcal{A}_{t-1})$ to enrol, set $\mathcal{D}_t = \mathcal{D}_{t-1} \cup (S_t, Y_t^C, Y_t^T)$

  Identify good subgroups $\mathcal{S}_t = \mathcal{S}_{t-1} \cup \mathcal{I}(\mathcal{D}_t, \alpha)$, set $\mathcal{A}_t = \mathcal{K} \setminus \mathcal{S}_t$

  Remove bad groups: $\mathcal{A}_t = \mathcal{K} \setminus \mathcal{R}(\mathcal{D}_t, \theta_{min}, \beta)$

  If $\mathcal{A}_t = \emptyset$, **Output:** `False`, $\emptyset$

Output: `True` if $|\mathcal{S}_B| > 0$, $\mathcal{S}_B$

Algorithm 2 AdaGCPI

Require: $\alpha, \beta \in (0,1)$, $\theta_{min} > 0$, budget B. ID. 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 $\mathcal{D}' = \{(j, Y_{t+j}^C, Y_{t+j}^T\}_{j \in \mathcal{A}_t}$, set $t+ = |\mathcal{A}_t|$, update $\mathcal{D}_t$.

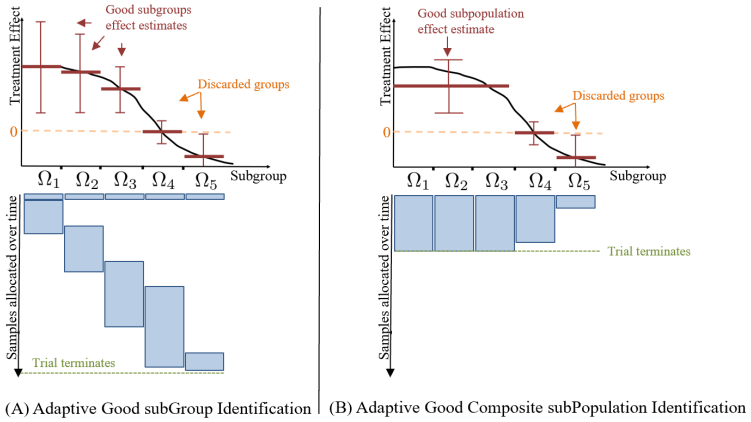
  Test for positive effect in current population $\mathcal{I}(\mathcal{D}_t, \alpha)$: if detected, **Output:** `True`, $\mathcal{A}_t$

  Remove bad groups: $\mathcal{A}_t = \mathcal{K} \setminus \mathcal{R}(\mathcal{D}_t, \theta_{min}, \beta)$

  If $\mathcal{A}_t = \emptyset$, **Output:** `False`, $\emptyset$

Output: `False`,$\emptyset$

A.2 Illustration of the two algorithms



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

Figure 1: Overview of the two considered problem formulations and proposed solutions. (A) The adaptive good subgroup identification (AdaGGI) algorithm finds individual good subgroups by successively discovering the next good group. (B) The adaptive good composite subpopulation (AdaGCPI) algorithm finds a composite subpopulation by successively removing the worst subgroup until a positive average treatment effect is discovered.

Appendix B. Possible extensions to more realistic settings

We believe that this paper opens up many interesting avenues for future research; a particularly interesting natural next steps lies in extending the setting under consideration to incorporate more realistic problem features. 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* [38]. 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 [39, 40] – 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 [41], while the bandit literature has developed approaches for decision making under delayed feedback [42]; it would be interesting to investigate how to incorporate either into our framework.

Appendix C. Experimental details

In Section 5, 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 $n_0 = 5$ (the number of initialization samples) due to the higher variance induced by considering *a difference* between random variables. To construct confidence intervals, as [22] we use Thm. 8 of [23] 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 this 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. That is, 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 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 $1-\beta$. α 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[9], and $\hat{p}^C, \hat{p}^T$ the observed binary proportions in each group, we have that

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

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

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

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

Appendix D. Additional results

D.1 Discussion of Type I error

Across 1000 repetitions of all simulation settings, we found that AdaGGI and AdaGCPI (with Bonferroni correction) *never* made a Type I error (incorrectly rejecting a true null hypothesis), while GSDS made Type I errors only in setting A ($\boldsymbol{\theta} = [0, 0, 0]$), in the expected $\approx 2.5\%$ of cases. To see whether this is due to the conservativeness of the Bonferroni correction, we reran AdaGGI and AdaGCPI *without* Bonferroni correction, and even then found that Type I errors occured in $\approx 0\%$ of cases. We attribute this observation to the used anytime confidence intervals being unnecessarily conservative as $t \ll \infty$ here. It may be interesting for future work to investigate how to construct less conservative confidence intervals, e.g. by making use of the fact that they only need to allow for at most $B \ll \infty$ peeks at the data.

D.2 Results with normal outcomes (Table 2)

9. 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

$\boldsymbol{\theta}$	Method	Binary %Succ.	$\lvert\mathcal{S}\rvert$	$\frac{t_{stop}}{B}$	$\frac{t_{1g}}{B}$	$\frac{t_{1b}}{B}$	Normal %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	2.4	0.04	0.75		0.51
	AdaGGI	0	0	0.64		0.24	0	0	0.69		**0.25**
	AdaGCPI	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	97.9	0.98	0.63	**0.46**	0.38	96.6	1	0.69	0.52	0.57
	AdaGCPI	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	99	1.00	0.55	0.29	0.59	79	0.87	0.93	0.34	0.57
	AdaGCPI	89	2.28	0.89	0.55	**0.44**	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	99.8	2.27	0.94	**0.36**		99.7	2.06	0.96	**0.4**	
	AdaGCPI	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	100	3	0.49	**0.16**		100	3	0.53	**0.18**	
	AdaGCPI	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) $\lvert\mathcal{S}\rvert$: 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 group (as prop. of budget). (5) t_{1b}/B: Average time it took to discard the *first* bad group (as prop. of budget).

Table 2: Results for simulated clinical trials with binary outcomes (left) and normal outcomes (right) using different treatment effect vectors θ; averaged across 1000 replications.