Improved Policy Evaluation for Randomized Trials of Algorithmic Resource Allocation

Aditya Mate [1 2] **Bryan Wilder** [3] **Aparna Taneja** [2] **Milind Tambe** [1 2]

Abstract

We consider the task of evaluating policies of algorithmic resource allocation through randomized controlled trials (RCTs). Such policies are tasked with optimizing the utilization of limited intervention resources, with the goal of maximizing the benefits derived. Evaluation of such allocation policies through RCTs proves difficult, notwithstanding the scale of the trial, because the individuals' outcomes are inextricably interlinked through resource constraints controlling the policy decisions. Our key contribution is to present a new estimator leveraging our proposed novel concept, that involves retrospective reshuffling of participants across experimental arms at the end of an RCT. We identify conditions under which such reassignments are permissible and can be leveraged to construct counterfactual trials, whose outcomes can be accurately ascertained, for free. We prove theoretically that such an estimator is more accurate than common estimators based on sample means — we show that it returns an unbiased estimate and simultaneously reduces variance. We demonstrate the value of our approach through empirical experiments on synthetic, semi-synthetic as well as real case study data and show improved estimation accuracy across the board.

1. Introduction

We consider a subclass of randomized controlled trials (RCTs) wherein the goal of the trial is to evaluate the efficacy of an algorithmic resource allocation policy. Such policies recommend an allocation (action), commonly utilizing tools such as reinforcement learning (Galstyan et al., 2004; Tesauro et al., 2006), variations of the multi-armed bandit framework (Jamieson et al., 2014), network optimization (Wilder et al., 2021; Vaswani et al., 2015), etc. As machine learning becomes increasingly widely applied in socially critical settings, such policies have been used to allocate limited resources in a variety of domains, including campaign optimization (Leskovec et al., 2007; Eagle et al., 2010), improving maternal healthcare (Verma et al., 2023), screening for hepatocellular carcinoma (Lee et al., 2019), monitoring tuberculosis patients (Mate et al., 2021), etc. Such a policy may rank all individuals within a group, and offer a scarce resource, such as a home visit by a health worker, to the highest-ranked individuals within the group.

RCTs evaluating these resource allocation policies consist of M experimental arms labeled $\{1, \ldots, M\}$, with each arm consisting of N unique, randomly assigned individuals. The policy in each experimental arm prescribes an allocation of resources while respecting some resource constraints. This process may be either single-shot (allocations made once) or sequential (new allocation decisions made adaptively over a series of rounds). Each policy, dictating its own resource allocation strategy, is evaluated at the end of the trial by analyzing the outcomes data from its corresponding experimental arm. The group-level decision making of resource allocation policies creates new challenges for their experimental analysis. In a standard RCT, the aim is to evaluate the treatment effect – there is no resource constraint, so all participants receive treatment – for the average participant (Angrist & Pischke, 2009). As the outcomes of all participants are independent, analysts can simply compare outcomes in the treatment versus control groups. However, trials of resource allocation policies aim to evaluate the *group* outcome of a set of individuals to whom the policy is applied (some of whom receive the resource, and some who do not). Even when the number of individuals in the trial is large by the standards of a normal RCT, randomized trials of allocation policies can suffer from high variance, leading to noisy and potentially erroneous estimates of the treatment effect.

This high variance stems from two sources. First, the outcomes of individuals within a given arm are correlated because allocation policies typically consider all individuals *jointly* in making an allocation decision (e.g., to respect

[1]School of Engineering and Applied Science, Harvard University, USA [2]Google Research, India [3]Carnegie Mellon University, USA. Correspondence to: Aditya Mate <aditya_mate@g.harvard.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

budget constraints on the total number of individuals who may receive an intervention). This interdependence implies that we only see one independent sample of a policy's performance per RCT, instead of many (in contrast with one sample per participant in a standard RCT). Second, for many individuals the allocation decisions made by different policies being compared may coincide. For instance, no matter the policy applied, many individuals may consistently get screened in, or many may get screened out of receiving a resource. Because these individuals are oblivious to the policy employed and receive identical allocations under different test policies, their outcomes are never truly impacted by the policy employed. Yet their presence generates additional variance in the average outcome of the group due to stochasticity in their outcomes, independent of the policy being evaluated. *To our knowledge, no prior work has proposed strategies to mitigate these sources of variance in RCTs of resource allocation algorithms, despite the increasing use of ML-based policies in socially critical domains.*

An intuitive first attempt at variance reduction would be to share information across arms. For example, we could average out fluctuations in the outcomes of individuals who do not receive an allocation by pooling together such individuals across all of the test policies. Unfortunately, such naive estimators are biased because the distribution of *which* individuals receive a resource (or not) is different across the trial arms – indeed, assessing the comparative efficacy of these distributions is exactly the point of the trial. A more sophisticated strategy would be to exploit overlap between the policies via inverse propensity weighting estimators which reweight participants in one arm to match the actions a different policy would have taken (Kuss et al., 2016; Austin & Stuart, 2015). However, such estimators are themselves subject to notoriously high variance, particularly when the overlap between policies is poor (Zhou et al., 2020). Moreover, as we discuss in Section 4.1, propensity scores are impossible to calculate for sequential intervention settings, where unseen states prevent us from evaluating individual-level propensities for a different policy.

Our key contribution fixing these issues is a novel estimator for the treatment effect of resource allocation policies which exploits overlap between arms in a principled manner, is guaranteed to reduce variance, and is applicable to either single-stage or sequential settings. The main idea is to find a subset of individuals with the following property: if we ran a hypothetical trial where the arm assignments of these individuals were swapped, the allocations made for all individuals remain unchanged. Our estimator identifies such sets of individuals with this property and averages the outcomes of all corresponding hypothetical trials (which are observable because no allocation decisions were altered from the original trial). We make the following contributions: (1) we propose this novel estimator and prove that it produces an unbiased estimate of the treatment effect with a guaranteed reduction in variance compared to the standard estimator, implying that it has strictly smaller average error; (2) we show how this estimator can be implemented in a computationally efficient manner for a class of policies encompassing those most commonly used in real-world allocation decisions; (3) we conduct experiments on three domains leveraging synthetic, semi-synthetic, and real-world data. We show the application of our techniques to real-world case study data illustrating its usefulness in solically critical domains. Across the board, our estimator substantially cuts error (by up to $\sim 70\%$) compared to available estimates.

2. Problem Formulation

RCT Setup: We consider a set of N individuals. Each has a feature vector $\mathbf{x} := [\mathbf{x}_o, \mathbf{x}_u]$, where $\mathbf{x}_o \in \mathcal{R}^o$ denotes the individual's observable features (such as demographics) while $\mathbf{x}_u \in \mathcal{R}^u$ denotes unobservable characteristics which nevertheless influence outcomes. A resource allocation policy π jointly considers all N individuals, their joint matrix of observable features $\mathbf{X}_o \in \mathcal{R}^{N \times o}$, and prescribes an allocation (action) vector $\mathbf{a} := \pi(\mathbf{X}_o) \in \mathcal{A}^N$ where $\mathcal{A}$ denotes the space of possible actions for each individual. The allocations $\mathbf{a}$ must respect resource constraints specific to the domain. On example is a budget constraint $||\mathbf{a}|| \leq B$ capping the total cost of allocated actions. Each individual then stochastically yields an outcome state $s \in \mathcal{S}$, according to an unknown function $P^*(\mathbf{x}, a, s)$ denoting the probability of the individual receiving outcome s. We use $r(s, a)$ to denote the reward accrued by the policymaker from this outcome. We remark that this notation is equivalent to the standard potential outcomes framework (Rubin, 2005) where $r(s, a)$ is the realization of the individual's potential outcome given action a and P governs the joint distribution between $\mathbf{x}$ and the potential outcomes. We adopt the present notation for easy generalization to the sequential setting.

The goal of an RCT is to compare M such resource allocation policies using M randomly constructed experimental arms $C_1, \ldots C_M$ where C_i denotes the set of individuals assigned to the ith arm. We use $\mathfrak{C} := \{C_1, \ldots C_M\}$ to denote this particular assignment of individuals to experimental arms, chosen uniformly at random from $\mathcal{C}$, which denotes the universal set of all possible assignments. At the end of an RCT, the analyst can compare the performance of the M allocation policies by comparing the sum total of rewards accrued by participants within each arm, given as $\text{Eval}(\pi_m) := \sum_{i \in C_m} r\big(s(i), a(i)\big)$.

Sequential RCTs: We also consider RCTs involving a multi-stage resource allocation setup. Such RCTs run for a total of T rounds (the single-stage setting above corresponds to $T = 1$), where allocations $\mathbf{a}_t \in \mathcal{A}^N$ must be made at each timestep $t \in [T]$ subject to potentially time-

dependent resource constraints. At each time t, each individual has a state s belonging to state space $\mathcal{S}$. We use $\mathbf{s}_t$ to denote the N individuals' states at timestep t. The policy π may consider the entire history of previous states and actions to generate a new action allocation as $\mathbf{a}_t \coloneqq \pi(\mathbf{X}_o, \mathbf{s}_{0:t-1}, \mathbf{a}_{1:t-1})$. Similarly, the individual transitions to a new state s_{t+1} according to the probability function $P^*(\mathbf{x}, s_{0:t}, a_{1:t+1}, s_{t+1})$. The state and action trajectories of all N participants are recorded for each of the M experimental arms as matrices: $S_1, \ldots, S_M \in \mathcal{S}^{N\times T+1}$ and $A_1, \ldots, A_M \in \mathcal{A}^{N\times T}$, which facilitates similar computation of $\text{Eval}(\pi_m) \coloneqq \sum_{i\in C_m} r(S[i], A[i])$ as defined for the single-shot setting.

Index-based policy: We define a specific class of policies – called 'index-based policies' – for which a computationally efficient estimator can be derived. Index-based policies rank individuals according to some scoring rule that determines a prioritization among individuals for allocation of a resource. Such policies encompass most relevant resource allocation policies commonly employed in practice due to their transparency and ease of implementation (Ustun & Rudin, 2019). Note that this class includes seemingly unlikely candidates that allocate resources to individuals cyclically in a set order or benchmarks such as 'control' groups (details in Appendix B). Formally, index-based policies are defined for a binary action space $\mathcal{A} \coloneqq \{0, 1\}$ with a budget constraint allowing at most B individuals to receive the action $a = 1$. The policy computes a time-dependent index $\Upsilon(\mathbf{x}_o, s_{0:t-1}, a_{1:t-1})$ for each individual at each timestep t, based solely on the individual's observable features $\mathbf{x}_o$, trajectory of states s and history of actions a received. The policy π allocates action $a = 1$ to the top B individuals with the largest values of index Υ. At any given time t, we assume the value of Υ is unique for each individual, i.e., the policy induces a total ordering. The key feature of an index-based policy is that Υ is computed independently for each individual based only on their own features and history.

Problem Statement: $\text{Eval}(\pi_m)$ provides a single random sample with which to estimate the *expected* performance of allocation policy π_m, combining the randomness in the assignments $\mathfrak{C} \in \mathcal{C}$ and the randomness in outcomes within each $\mathfrak{C}$, engendered by stochasticity in state transitions. Let $\text{Eval}^*(\pi_m)$ be the expected value of the performance of π_m:

$$\text{Eval}^*(\pi_m) \coloneqq \mathbb{E}_{S_m\sim P^*}\mathbb{E}_{\mathfrak{C}\sim\mathcal{C}}\Big[\text{Eval}(\pi_m)\Big] \tag{1}$$

We assume data available from only a single run of an RCT. Our goal is to build an estimator that estimates $\text{Eval}^*(\pi_m)$ accurately from just the single RCT instance. We remark that our results make no distributional assumptions about the set of individuals in the trial, e.g., we do not require them to be IID from some distribution. Rather, we treat the observed individuals as fixed (nonrandom) and propose techniques that use only the randomness in the assignment process of the RCT. The only assumption required is that individuals' state transitions are independent, akin to the standard stable unit treatment values assumption (SUTVA) in causal inference (Hernán & Robins, 2010). This is aligned with the growing emphasis on design-based inference in causal inference (c.f. (Hudgens & Halloran, 2008; Mukerjee et al., 2018; Abadie et al., 2022)), which allows us to formulate methods which require minimal modeling assumptions.

3. Related Work

Off-Policy Evaluation: The most closely related previous work is the off-policy evaluation (OPE) literature, where the goal is to use samples collected under some baseline policy to inform the evaluation of a new policy (Sutton et al., 2008; Jiang & Li, 2016; Wang et al., 2017). OPE makes frequent use of inverse propensity weighting estimators (Liu & Zhao, 2010; Li et al., 2015); we discuss strategies for constructing such estimators as well as their disadvantages in Section 4.1. To date, the OPE literature has largely focused on individual-level decisions, as opposed to the group-level resource allocation we study. One exception is slate-level OPE, where the policy recommends a ranked list of items to a user (Swaminathan et al., 2017). Slate OPE is most similar to our single-step case, while we develop methods that extend to the multi-step setting. Additionally, slate OPE is motivated by unobservable individual-level rewards, while we assume that individual rewards are observable and the challenge for the single-step setting is that policies may be deterministic (preventing us from using their methods).

Individual treatment rules: Recent work in statistics has studied experimental design and analysis for individual treatment rules (ITR), which are similar to our class of index-based policies (Imai & Li, 2021; Athey & Wager, 2021). The crucial difference is that ITRs make decisions independently for each individual, while in our setting the policy considers the group of individuals jointly, which is required for exact enforcement of constraints such as budgets. Our techniques are motivated by the need to reduce variance when policies can only be evaluated at group level.

Cluster-randomization and interference: Our setting is related to a family of RCTs known as cluster-randomized trials (see (Hayes & Moulton, 2017) for an overview). In such trials, treatment is assigned at a group level (e.g., assignment of classrooms within a school instead of students), just as groups of individuals are assigned to policies in our setting. However, cluster-randomized trials are motivated by the potential for spillover effects, where the outcomes of one unit can influence others. By contrast, in our setting the outcomes of individuals are still independent conditioned on the actions of the policy. Accordingly, there is no need to account for potential correlations as in the interference

literature; instead, we leverage this structured independence to develop lower-variance estimators.

4. Methodology

Our goal is to leverage the overlap in decisions of multiple resource allocation policies to improve our estimate of the reward from deploying each. We start by developing an estimator using inverse propensity weighting – a popular approach typically adopted for such a task – and show how it can be naturally applied to the single-stage setting. However, this natural estimator suffers from two challenges. First, inverse propensity estimators can suffer from notoriously high variance, a phenomenon that we empirically confirm in Section 6. Second, the approach breaks down entirely in the multi-stage setting, where (as detailed below) computation of propensities is impossible due to missing data. We resolve these challenges by developing a more stable "assignment permutation" estimator, which applies to both settings and is guaranteed to reduce estimation error.

4.1. Propensity Scores Approach

In typical off-policy evaluation settings, inverse propensity weighting (IPW) methods reweight samples according to the probability that observed actions would be taken by a given policy. These methods are not immediately applicable to our problem because we do not assume that policies are randomized – indeed, explicit randomization is rare in policies deployed by real-world governments, health systems, etc. When policies are deterministic, the probability that they would yield an alternate action is precisely zero, leaving standard propensity estimators undefined.

We show how to circumvent this issue in the single-step setting by leveraging an alternate source of randomness: the assignment in the trial itself. Calculated over the randomness in the assignment, each individual has some probability of being assigned a given action, denoted as $Pr_{\mathfrak{C},\pi}[a(i) = a]$ (intuitively, whether they receive a resource depends on who else the policy is comparing them to). Formally, exchanging the order of expectations allows us to write $\text{Eval}^*(\pi)$ as:

$$\sum_{i=1}^{N} Pr[i \in C_m] \sum_{a \in \mathcal{A}} Pr_{\mathfrak{C},\pi}[a(i) = a | i \in C_m] \mathbb{E}_s[r(s, a)]$$

Since the assignment $\mathfrak{C}$ is random, $Pr[i \in C_m] = \frac{1}{M}$. Moreoever, in the inner term, conditioning on $i \in C_m$ leaves the other members of C_m distributed uniformly at random. Accordingly, we can estimate $Pr_{\mathfrak{C},\pi}[a(i) = a | i \in C_m]$ for any policy π by drawing repeated samples of the assignment $\mathfrak{C}$ and running π to reveal whether π would have assigned $a(i) = a$ given the group C_m containing individual i. Let $\hat{p}(i, a|\pi)$ denote the fraction of these samples in which $a(i) = a$. A standard IPW estimator for $\text{Eval}(\pi)$ is given by

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{i \in C_m} \frac{\hat{p}(i, a(i)|\pi)}{\hat{p}(i, a(i)|\pi_m)} r(s(i), a(i)). \tag{2}$$

In a sequential setup ($T > 1$), propensity score methods become entirely inapplicable for two reasons. First, standard multi-time step IPW estimators require randomness in the policy, while we assume that policies may be deterministic. We cannot use the alternate approach described above (leveraging randomness in assignments) over multiple time steps, because the marginal probability that individual i receives action $a(i)$ on future steps depends on the state of all other individuals, and we do not have samples of such future states under counterfactual assignments. Second, even if we limited to randomized policies, standard off-policy methods calculate the probability of taking exactly the observed sequence of actions in the observed states. In our case, this requires computing the probability of π selecting the *vector* of actions assigned to each individual, i.e, we have a N-dimensional action space within each time step. Multi-step IPW estimators are already known to suffer from variance which explodes exponentially in T, often rendering them impractical (Li et al., 2015). In our case, their variance would (in the worst case) scale exponentially in N as well.

4.2. Main Contribution: Assignment Permutation

We present a novel approach that counters both challenges to compute a stable, accurate estimator. The key idea behind our estimator is to identify hypothetical trials with counterfactual experimental group assignments, whose reward outcomes can be exactly determined using the given outcomes from the original trial. We leverage the fact that although the state transitions depend only the received allocations, regardless of what policy π chooses those allocations.

As a warm-up, consider a single-shot trial $\mathcal{T}$ in which two individuals i and j are assigned to policies π_i and π_j, that make identical resource allocations a to both individuals, yielding outcomes s_i and s_j respectively. Now consider a hypothetical trial $\mathcal{T}^\dagger$, run exactly identical to $\mathcal{T}$ except that the assignments of i and j are switched. If in $\mathcal{T}^\dagger$, both i and j receive the same allocation a as in $\mathcal{T}$, allocations to other individuals would also remain unaffected, and consequently, all individuals would see identical inputs in both $\mathcal{T}^\dagger$ and $\mathcal{T}$. Thus, the actual sample of outcomes $\mathbf{s}$ in $\mathcal{T}$ is a sample from the same distribution as that induced by $\mathcal{T}^\dagger$. Generalizing this idea, consider a sequential RCT $\mathcal{T}$, in which a subset of individuals' group assignments are permuted to construct a hypothetical trial $\mathcal{T}^\dagger$, which sees new allocations $\mathbf{a}_t^\dagger$ made at time t. If an individual experiences sub-trajectories of states $s_{0:t-1}$ and actions $a_{1:t-1}$ till time $t-1$ that are identical in both $\mathcal{T}^\dagger$ and $\mathcal{T}$, and if the new allocation $a_t^\dagger$ received is also identical to a_t, then original state sample s_t observed in $\mathcal{T}$,

is also a valid sample in $\mathcal{T}^\dagger$, drawn from the same distribution, $P^*(\mathbf{x}, s_{0:t-1}, a_{1:t}, s_t)$. Furthermore, inductively, the entire original state trajectory $s_{0:T}$ of $\mathcal{T}$ can be treated as a valid sample for $\mathcal{T}^\dagger$ if $\forall\, t \in [T]$ the input sub-trajectory $s_{0:t-1}$, produces new allocations, $a_t^\dagger$ that are identical to a_t.

We exploit this concept to retrospectively check for all such possible reassignments, that would lead to the same sequence of output actions given the same input sub-sequence of the state-action trajectory at all times. The implication is that this allows us to uncover and aggregate outcomes from several such additional 'observable counterfactual assignments' (defined below) in estimating the performance of a given test policy. Algorithm 1 outlines the idea for a general M-arm setting. Later, in Algorithm 2 we present an efficient algorithm crafted for handling index-based policies.

Definition 4.1 (Observable Counterfactual Assignment)**.** For an actual assignment $\mathfrak{C}$, we define $\mathfrak{C}^\dagger$ to be an observable counterfactual assignment if in a hypothetical trial with assignments $\mathfrak{C}^\dagger$, for each $t = 1...T$ the actions each policy would assign to each individual are identical to the original actions received, conditioned on the state and action histories $(\mathbf{s}_{0:t-1}, \mathbf{a}_{1:t-1})$ matching up until time $t-1$.

Estimation via Assignment Permutation: Let $\mathcal{C}^\dagger(\mathfrak{C})$ be the set of all 'observable counterfactual assignments' engendered by a single actual experimental assignment $\mathfrak{C}$. Our proposed estimator averages the outcomes of all such observable counterfactuals:

$$\text{Eval}^\dagger(\pi_m) := \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger} \text{Eval}(\pi_m|\mathfrak{C})}{|\mathcal{C}^\dagger|} \tag{3}$$

In Theorem 4.3, we show that this is an unbiased estimator for the true expectation Eval*. The main technical step (Lemma 4.2) is to show that the that $\mathcal{C}^\dagger(\mathfrak{C})$ defines a partition over $\mathcal{C}$ (the set of all possible assignments) where two assignments $\mathfrak{C}_1, \mathfrak{C}_2$ lie in the same part if $\mathcal{C}^\dagger(\mathfrak{C}_1) = \mathcal{C}^\dagger(\mathfrak{C}_2)$. Intuitively, this means that our estimator does not "overweight" any particular counterfactual assignment; it maintains the equal weight that each has in Eval*.

Algorithm 1 Estimation through Assignment Permutation

Input: States : $\{S_1^{N\times T+1}, \ldots, S_M^{N\times T+1}\}$, Actions : $\{A_1^{N\times T}, \ldots A_M^{N\times T}\}$, Assignment, $\mathfrak{C}$: $\{C_1, \ldots, C_M\}$
Output: Eval†

1: Compute $\mathcal{C}^\dagger$, the set of observable counterfactual assignments of $\mathfrak{C}$.
2: Compute $\text{Eval}^\dagger(\pi_m) := \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger} \text{Eval}(\pi_m|\mathfrak{C})}{|\mathcal{C}^\dagger|}$
3: **return** $\text{Eval}^\dagger(\pi_m)$

4.3. Theoretical Results

In this section, we prove theoretically that $\text{Eval}^\dagger(.)$ gives a more accurate estimate because it is unbiased and simultaneously reduces variance. Let $\dagger$ be a homogeneous relation on $\mathcal{C}$, defined as: $\dagger = \{(\mathfrak{C}_1, \mathfrak{C}_2) \in \mathcal{C}\times\mathcal{C} :\ \mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)\}$. Intuitively, $\dagger$ represents existence of a valid reshuffling to arrive at a counterfactual assignment $\mathfrak{C}_2$ from $\mathfrak{C}_1$.

Lemma 4.2. *The relation $\dagger$ is an equivalence relation and the family of sets defined by $\mathcal{C}^\dagger(\cdot)$ forms a partition over $\mathcal{C}$.*

Proof Sketch. To prove that $\dagger$ is an equivalence relation, we show that it is reflexive, symmetric and transitive. Thus $\dagger$ is an equivalence relation over $\mathcal{C}$ and consequently, partitions $\mathcal{C}$ into a family of equivalence classes $\mathcal{C}^\dagger(\cdot)$ such that every element $\mathfrak{C} \in \mathcal{C}$ lies in exactly one partition (Enderton, 1977). We leverage this property to prove unbiasedness:

Theorem 4.3. *Eval$^\dagger(\pi_m)$ is an unbiased estimate of the expected value of the performance, Eval$^*(\pi_m)$ $\forall m \in [M]$, defined in equation 1. i.e.*

$$\mathbb{E}_{S_m\sim P^*}\mathbb{E}_{\mathfrak{C}\sim\mathcal{C}}[\textit{Eval}^\dagger(\pi_m)] = \textit{Eval}^*(\pi_m)\ \forall m \in [M]$$

Proof.

$$\begin{aligned}
\text{Eval}^*(\pi_m) =& \mathbb{E}_{S_m\sim P^*}\Big[\mathbb{E}_{\mathfrak{C}\sim\mathcal{C}}\big[\text{Eval}(\pi_m)\big]\Big] \\
&= \mathbb{E}_{S_m\sim P^*}\Big[\sum_{\mathfrak{C}\,\in\,\mathcal{C}} \text{Prob}[\mathfrak{C}]\cdot\text{Eval}(\pi_m|\mathfrak{C})\Big] \\
&= \mathbb{E}_{S_m\sim P^*}\Big[\frac{1}{|\mathcal{C}|}\sum_{\mathfrak{C}\,\in\,\mathcal{C}} \text{Eval}(\pi_m|\mathfrak{C})\Big] \\
&\qquad (\because \text{all } \mathfrak{C} \text{ equally likely}) \\
&= \mathbb{E}_{S_m\sim P^*}\Bigg[\frac{1}{|\mathcal{C}|}\Big[\sum_{\mathfrak{C}\,\in\,\mathcal{P}_1} \text{Eval}(\pi_m|\mathfrak{C}) + \cdots + \\
&\qquad \sum_{\mathfrak{C}\,\in\,\mathcal{P}_\eta} \text{Eval}(\pi_m|\mathfrak{C})\Big]\Bigg] \\
&\text{where } \{\mathcal{P}_1, \ldots, \mathcal{P}_\eta\} \text{ defines partition of } \mathcal{C} \text{ induced by } \dagger. \\
&= \mathbb{E}_{S_m\sim P^*}\Bigg[\sum_{j\in[\eta]} \frac{|\mathcal{P}_j|}{|\mathcal{C}|}\cdot\frac{1}{|\mathcal{P}_j|}.\Big[\sum_{\mathfrak{C}\,\in\,\mathcal{P}_j} \text{Eval}(\pi_m|\mathfrak{C})\Big]\Bigg] \\
&= \mathbb{E}_{S_m\sim P^*}\Bigg[\sum_{j\in[\eta]} \frac{|\mathcal{P}_j|}{|\mathcal{C}|}\cdot\Big[\text{Eval}^\dagger(\pi_m|\mathfrak{C})\Big]\Bigg] \qquad (\forall\mathfrak{C}\in\mathcal{P}_j) \\
&= \mathbb{E}_{S_m\sim P^*}\Bigg[\sum_{j\in[\eta]} \text{Prob}[\mathcal{P}_j]\cdot\Big[\text{Eval}^\dagger(\pi_m|\mathfrak{C})\Big]\Bigg](\forall\mathfrak{C}\in\mathcal{P}_j) \\
&= \mathbb{E}_{S_m\sim P^*}\Bigg[\sum_{j\in[\eta]}\sum_{\mathfrak{C}\in\mathcal{P}_j} \text{Prob}[\mathfrak{C}]\cdot\Big[\text{Eval}^\dagger(\pi_m|\mathfrak{C})\Big]\Bigg] \\
&\qquad \Big(\because \text{Prob}[\mathcal{P}_j] = \sum_{\mathfrak{C}\in\mathcal{P}_j} \text{Prob}[\mathfrak{C}]\Big) \\
&= \mathbb{E}_{S_m\sim P^*}\mathbb{E}_{\mathfrak{C}\sim\mathcal{C}}[\text{Eval}^\dagger(\pi_m)]
\end{aligned}$$

Theorem 4.4. *The sample variance of our estimator, $Eval^{\dagger}(\pi)$ is smaller than the standard estimator, $Eval(\pi)$:*

$$\mathrm{Var}(Eval(\pi)) - \mathrm{Var}(Eval^{\dagger}(\pi)) =$$

$$\frac{1}{|\mathcal{C}|} \cdot \sum_{j \in [\eta]} \left[\sum_{\mathfrak{C} \in \mathcal{P}_j} Eval^2(\pi|\mathfrak{C}) - \frac{\left(\sum_{\mathfrak{C} \in \mathcal{P}_j} Eval(\pi|\mathfrak{C}) \right)^2}{|\mathcal{P}_j|} \right]$$

≥ 0*, where $\{\mathcal{P}_1, \ldots, \mathcal{P}_\eta\}$ is the partition of $\mathcal{C}$ induced by $\dagger$.*

Proof Sketch. We compute the sample variance by first conditioning over the partition $\mathcal{P}_j$ (of the equivalence sets defined by $\dagger$) that an instance of an assignment, $\mathfrak{C}$ belongs to and then accounting for the variance stemming from the candidate assignments $\mathfrak{C}$ within the partition. Finally, we use the Cauchy-Schwarz inequality to show that the right-hand-side expression in Theorem 4.4 is non-negative. □

The variance contraction expression of Theorem 4.4 reduces to zero if and only if $\mathrm{Eval}(\pi|\mathfrak{C})$ is identical $\forall \mathfrak{C} \in \mathcal{P}_j,\ \forall\, j \in [\eta]$; if different assignments imply different rewards then our estimator exhibits a strict improvement in variance.

Corollary 4.5. *Let $\mathcal{P}^*$ denote the part of the partition that the instance of assignment $\mathfrak{C}$, under consideration belongs to. Let N_{perms} denote the number of observable counterfactual assignments belonging to $\mathcal{P}^*$ (thus yielding $|\mathcal{P}^*| = N_{perms} + 1$). Then given the data observed, the lower bound on the variance reduction achieved by our estimator $Eval^{\dagger}(\pi)$ can be given as:*

$$\mathrm{Var}(Eval(\pi)) - \mathrm{Var}(Eval^{\dagger}(\pi)) \geq$$

$$\frac{1}{|\mathcal{C}|} \cdot \left[\sum_{\mathfrak{C} \in \mathcal{P}^*} Eval^2(\pi|\mathfrak{C}) - \frac{\left(\sum_{\mathfrak{C} \in \mathcal{P}^*} Eval(\pi|\mathfrak{C}) \right)^2}{N_{perms} + 1} \right]$$

Proof. The lower bound on variance reduction is obtained by simply considering the worst case in which we set $|\mathcal{P}^*| = N_{perms} + 1$ for the current part and $|\mathcal{P}_j| = 1$ for all other parts in the partition. Plugging this back in the expression of Theorem 4.4 yields the final result.

5. Efficient Swapping Algorithm

Identifying $\mathcal{C}^{\dagger}$ exhaustively in Algorithm 1 involves iterating through every possible assignment in $\mathcal{C}$ and running the policy to determine if the assignment belongs to $\mathcal{C}^{\dagger}$. However, the number of possible assignments $|\mathcal{C}|$ grows exponentially with N, making full enumeration infeasible. We show how this computational bottleneck can be circumvented for index-based policies with a modified estimator denoted $\mathrm{Eval}^{\dagger}_{\Upsilon}(\cdot)$. This estimator implicitly averages over a subset of the possible permutations in $\mathcal{C}^{\dagger}$, trading off some variance reduction for computational efficiency.

For ease of exposition, here we consider RCTs with two experimental arms ($M = 2$) employing allocation policies π_0 and π_1 respectively. We aim to estimate $\mathrm{Eval}^*(\pi_j)_{j=\{0,1\}}$. Intuitively, instead of working with the space of all possible assignments, we instead consider all individuals participating in the trial and to identify non-overlapping groups of individuals $\{\mathbf{G}_k\} \subset (C_0 \cup C_1)$ that satisfy certain desirable properties. Specifically, we intend to find sets of 'compatible' individuals such that any subset of individuals within each group $\{\mathbf{G}_k\}$ can be mutually swapped to arrive at either an unchanged assignment or a valid observable counterfactual assignment in $\mathcal{C}^{\dagger}(\mathfrak{C})$. Our intention is to compute an estimate by replacing the original reward of every individual $i \in \mathbf{G}_k$ by the average of rewards of all individuals in $\mathbf{G}_k$, for every such group $\mathbf{G}_k$ (justification in Theorem 5.1). We identify these groups by checking for two eligibility conditions pertaining to swappability of individuals.

The first eligibility condition for swapping two individuals i and j is that their original allocations $a(t)$ must be identical to each other $\forall t$, to continue to satisfy the resource constraints in both arms after the swap. To check for this condition, we partition individuals into super-groups $\{\bar{\mathbf{G}}_1, \ldots, \bar{\mathbf{G}}_\kappa\}$, putting all individuals experiencing the same action vector $a(t) \in \mathcal{A}^T$, in the same super-group, where κ denotes the number of such super-groups. All individuals within each $\bar{\mathbf{G}}_k$ satisfy this first eligibility condition for being included in group $\mathbf{G}_k$. For convenience, we let $\phi\colon C_0 \cup C_1 \to [\kappa]$ denote a many-to-one map identifying the super-group $\bar{\mathbf{G}}_{\phi(i)}$ that an individual i belongs to.

The second eligibility condition for swapping an individual is that their new allocation $a^{\dagger}(t)$ under the new policy must be identical to the original $a(t)$, for the same sequence of input states as in the original trial. For each individual i, we use a binary-valued variable $\Lambda_i \in \{0, 1\}$ to indicate satisfaction of this second condition. We introduce and exploit the 'index-threshold' property here to verify this condition efficiently. We define an index threshold $\tau_j(t)$ as the smallest value among indices $\Upsilon(t)$ of individuals in C_j at time t, that get picked to receive the allocation $a = 1$ under policy π_j. To enable efficient computation, we only allow swaps within a group $\mathbf{G}_k$ that maintain the index thresholds $\tau_j(t)$ at the same values as the original. We implement this constraint by setting $\Lambda_i = 0$ for all individuals exactly at the index threshold. Furthermore, for other individuals i, Λ_i can be cheaply determined by just verifying if the index $\Upsilon_i^{\pi_j}(t)$ lies to the same side of threshold $\tau_j(t)\ \forall\, t \in [T]$ and for $j \in \{0, 1\}$. To summarize

$$\Lambda_i = \begin{cases} 1 & \text{if } \prod_{j=0}^{j=1} (\Upsilon_i^{\pi_j}(t) - \tau_j(t)) > 0\ \forall\, t \in [T] \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Intuitively, $\Lambda_i = 1$ means that $a^{\dagger}_{1:T}(i) = a_{1:T}(i)$ and indicates that individual i satisfies the second eligibility condition. Taking an intersection of both conditions, we form group $\mathbf{G}_k$ by including all individuals $i \in \bar{\mathbf{G}}_k$ that have

$\Lambda_i = 1$. For each group $\mathbf{G}_k$, we compute the reward of a representative average individual as:

$$\tilde{r}_k := \frac{1}{|\mathbf{G}_k|} \sum_{i \in \mathbf{G}_k} r(S[i], A[i]) \tag{5}$$

In computation of the final estimate $\text{Eval}^\dagger_\Upsilon(\pi_j)$, we consider all individuals in the arm C_j, but replace the reward of swappable individuals among those (i.e. whose $\Lambda_i = 1$) by $\tilde{r}_{\phi(i)}$. We leave the rewards of other individuals unchanged. Finally we compute $\text{Eval}^\dagger_\Upsilon(\pi_j)$ by summing up as:

$$\text{Eval}^\dagger_\Upsilon(\pi_j) = \sum_{i \in C_j} \Big(\Lambda_i \tilde{r}_{\phi(i)} + (1 - \Lambda_i) r(S[i], A[i]) \Big) \tag{6}$$

Our theoretical analysis of this estimator establishes that it corresponds to an instance of the general permutation estimator which averages over a subset of the assignments in $\mathcal{C}^\dagger$ (instead of the entire set). The main idea is that we can sub-partition $\mathcal{C}^\dagger$ into sets with the same value of the index threshold (shown formally in Lemma A.2). We denote the part of $\mathcal{C}^\dagger$ where the index thresholds are the same as in the actual trial as $\mathcal{C}^\dagger_\Upsilon$. Each permutation of individuals within groups $\{\mathbf{G}_k\}$ corresponds to an assignment in $\mathcal{C}^\dagger_\Upsilon(\mathfrak{C})$, and the final estimator averages over all such assignments:

Theorem 5.1. *$\text{Eval}^\dagger_\Upsilon(\cdot)$ computed as per Equation 6 computes the average of $\text{Eval}(\pi|\mathfrak{C})$ over all assignments in $\mathcal{C}^\dagger_\Upsilon$. i.e. $\text{Eval}^\dagger_\Upsilon(\pi) = \frac{\sum_{\mathfrak{C} \in \mathcal{C}^\dagger_\Upsilon} \text{Eval}(\pi|\mathfrak{C})}{|\mathcal{C}^\dagger_\Upsilon|}$*

Proof Sketch. The key to showing that the two are equivalent is in interpreting the summation of (modified) rewards over individuals in the view of average group rewards over assignments. Mathematically, starting from the definition of $\text{Eval}^\dagger_\Upsilon(\pi_m)$, the key lies in moving the summation operation over assignments in $\mathcal{C}^\dagger_\Upsilon$ from outside the $\text{Eval}^\dagger_\Upsilon()$ term to inside, applying it individually on each contributing participant. □

From this, $\text{Eval}^\dagger_\Upsilon(\pi_j)$ is easily shown to inherit the desirable properties of the general permutation estimator, e.g., Corollary A.3 proves that it is also an unbiased estimator of $\text{Eval}^*(\pi_j)$. The tradeoff is a slight sacrifice in variance contraction as it yields smaller partitions $\mathcal{P}_j$ of $\mathcal{C}$ (as defined in Theorem 4.4), since we discard assignments with a different threshold. However, working with $\mathcal{C}^\dagger_\Upsilon$ enables a computationally efficient algorithm for index policies, avoiding the exponential runtime of the general estimator.

6. Empirical Evaluation

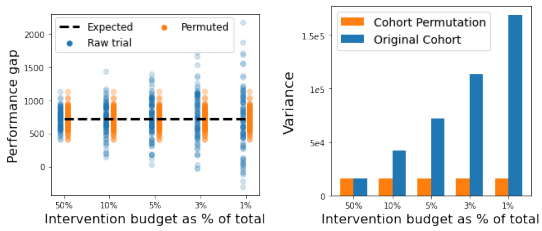
We test our proposed methodology empirically on several datasets: (1) synthetic example (2) semi-synthetic tuberculosis medication adherence monitoring data and (3) real-world

Algorithm 2 Reshuffling between index based policies

Input: $\mathfrak{C} := \{C_0, C_1\}$, States: $S_0, S_1 \in \mathcal{S}^{N \times T+1}$, Actions: $A_0, A_1 \in \{0,1\}^{N \times T}$, Indexes: $\Upsilon^{\pi_0}, \Upsilon^{\pi_1} \in \mathbb{R}^{2N \times T}$

Output: Estimates: $\text{Eval}^\dagger_\Upsilon$

1: Group individuals according to action vectors into $\{\bar{\mathbf{G}}_k\}$ and determine $\phi(i)$ $\forall$ individuals $i \in C_0 \cup C_1$.
2: Determine $\forall t \in [T]$, index thresholds on both arms as: $\bar{\tau}_j(t) := \min\left\{\Upsilon_{\pi_j}(i,t) \,|A_j[i,t] = 1,\ i \in [N]\right\}$
3: Determine Λ_i $\forall$ individuals i according to Equation 4.
4: Compute group $\mathbf{G}_k$ as: $\{i \in \bar{\mathbf{G}}_k \,|\Lambda_i = 1\}, \forall k \in [\kappa]$
5: For each group $\mathbf{G}_k$, compute the average reward of a representative individual as per Equation 5.
6: Compute $\text{Eval}^\dagger_\Upsilon(\pi_j)$ per Equation 6



(a) Total engagements (b) Sample Variance

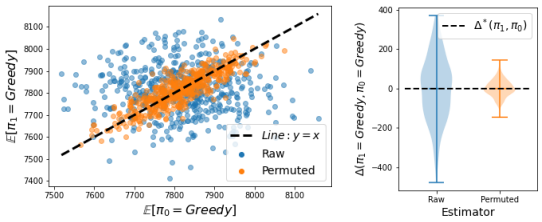
Figure 1. Estimates and sample variance in synthetic domain.

field trial data from an intervention for a maternal healthcare. We consider a state space $\mathcal{S} = \{0, 1\}$, respectively representing an 'undesirable' and a 'desirable' state (of health, program engagement, etc.). The action space $\mathcal{A} = \{0, 1\}$, denotes 'no delivery' or 'delivery' of an intervention. We assume a budget constraint, limiting the total number of interventions per time step. The reward function is defined as $r(s_{0:T}, a_{1:T}) = \sum_t s_t$, translating to an objective of maximizing the total time spent by individuals in state $s = 1$.

6.1. Synthetic Dataset

This setup consists of three types of individuals characterized by their $P-$ matrices. P_1 and P_2 are designed such that it is always optimal to intervene on P_1 individuals consistently, whereas intervening on P_2 individuals is strictly sub-optimal (details in Appendix C.1). P_3 individuals are unaffected by interventions, with transition dynamics independent of the action received. We consider two test policies π_1 and π_2, which are designed such that policy π_j always chooses individuals of type P_j to intervene on, when available, making π_1 the optimal policy. We simulate 300 P_1 individuals, 300 P_2 individuals and $(300^*\eta\ P_3)$ individuals and set an intervention budget of 300 per timestep for $T = 20$ timesteps. We measure the performance lift of π_1 against π_2 as: $\Delta := \Delta(\pi_1, \pi_2) = \text{Eval}(\pi_1) - \text{Eval}(\pi_2)$.

In Figure 1a, we vary the budget on the x-axis. Each blue



(a) (b)

Figure 2. Multi-step setting. (a) The permuted estimates (orange) are closer to the true expectation (black line) than the raw estimates (blue) (b) Assignment permutation reduces variance.

dot in the scatter plot shows one independent RCT instance and measures the raw difference in rewards Δ on the y-axis. Applying assignment permutation maps each blue dot to an orange dot. The black dashed line marks the expected value of the performance lift. Visually, both colors are centered on the black line, as both estimators are unbiased. The assignment-permuted estimates lie closer to the expected value than the raw estimates, indicating a smaller sample variance. Quantitatively, we measure the sample variance of Δ on the y-axis in Figure 1b. Variance reduces sharply upon applying assignment permutation – for instance, at a budget level of 3%, our approach cuts the variance by $7\times$, from 11.3×10^4 to 1.6×10^4. The intuition is that both π_1 and π_2 overlap in their decision to not intervene upon P_3 individuals. However, their final rewards are based partly on which P_3 individual gets (randomly) assigned to which group, independent of the underlying policies. Assignment permutation counters this randomness by averaging over alternate assignments of the P_3 individuals.

6.2. Semi-synthetic evaluation with tuberculosis dataset

We use real tuberculosis medication adherence monitoring data, consisting of daily records of patients in Mumbai, India, obtained from (Killian et al., 2019) and simulate patient behavior by estimating the P matrix. More details can be found in the appendix.

We consider two policies: a "Whittle index" policy (Whittle, 1988) that attempts to maximize long-run reward, and a greedy policy which optimizes an estimate of next-step reward. We simulate $N = 1000$ patients in each arm and vary the budget constraint. We consider both the multi-step setting ($T = 10$) and single step ($T = 1$). We compare three estimation methods: "Raw" is the naive average of outcomes in each arm, "Permuted" our proposed estimator, and "IPW" the inverse propensity estimator from Section 4.1 (available only for $T = 1$). In practice, we find it necessary to trim propensity scores for IPW (Zhou et al., 2020) to the range [0.01, 0.99], since extreme values lead to very large variance. This introduces a slight bias, visible in Figure 3b.

Table 1 shows the sample variance of estimates returned by each method. For unbiased methods (Raw, Permuted) the variance is also their mean squared error; this holds approximately for IPW due to trimming. Table 1 includes an additional column labeled '$n-$value'. To benchmark the improvement produced by our method, this gives the minimum number of *independent* RCTs that would need to be run (and averaged over) to match the sample variance achieved by assignment permutation (computed by simulation).

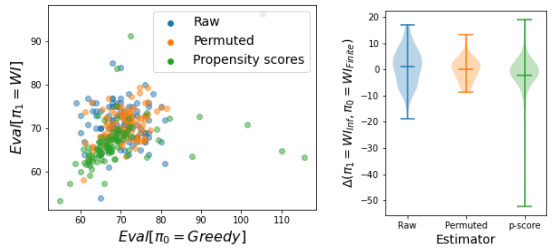
For all comparisons and parameter settings, we find that our assignment permutation estimator produces a substantial improvement in variance. Indeed, achieving a comparably precise estimate using the naive raw estimator would require running anywhere from 2 to *13* independent RCTs. This underscores the importance of variance reduction – running RCTs is hugely costly and assigns many individuals to suboptimal policies; improved analysis allows us to draw comparably precise conclusions at dramatically lower cost.

Figure 2 illustrates this improvement in a single example where both trial arms are the Greedy policy and so the expected difference in rewards is exactly zero (with $B = 3\%$ and $T = 10$). Each dot in Figure 2a corresponds to a single instance of a trial, with the x- and y-axes giving total engagements in the two arms. The black dashed line ($y = x$) denotes the expectation, $\Delta = 0$. The blue dots, representing raw measurements, have a wider spread around the black line than the orange dots obtained via assignment permutation. Figure 2b, shows a violin plot of the sample difference in rewards between the two arms. Both violins are centered on the zero line, reflecting that the estimators are unbiased. The violin corresponding to assignment permutation is more compact, indicating lower sample variance.

In the single-step setting, the IPW estimator returns mixed results: it has the best variance for small values of the budget, but actually performs *worse* than the naive raw estimator for larger budgets. Essentially, the overlap between two policies becomes smaller as the budget increases because they agree only on the few highest-priority individuals. Low overlap translates into extreme propensity scores, inflating variance. Figures 3b and 3a show an illustration, where IPW often produces an improvement but is susceptible to large outliers. However, when overlap is high and we operate only in the single-step setting, IPW can be a valuable option.

6.3. Case study: Real-world Trial

Our method is directly applicable to real-world settings; we show this by considering an actual large-scale RCT reported in (Mate et al., 2022) evaluating a Restless Multi-Armed Bandit-based algorithm for resource allocation in a maternal and child healthcare. The data consists of 23,000 real-world beneficiaries, randomly split between three groups for the trial: RMAB algorithm, baseline algorithm and a control



(a) (b)

Figure 3. Illustration of results for single-step setting

Table 1: Sample variance in Measured Performance Lift

T	B	π_1 V π_0	RAW	PERMUTED	IPW	n-VAL
1	3%	π_{WI} V π_{GR}	49.09	4.94	**0.48**	9
1	10%	π_{WI} V π_{GR}	49.86	15.11	**6.66**	3
1	25%	π_{WI} V π_{GR}	49.45	**19.94**	78.12	2
10	3%	π_{WI} V π_{WI}	2381	**916**	NA	3
10	3%	π_{WI} V π_{GR}	2348	**728**	NA	4
10	3%	π_{GR} V π_{GR}	26356	**1860**	NA	13
10	10%	π_{GR} V π_{GR}	25983	**3808**	NA	7
10	25%	π_{GR} V π_{GR}	23619	**5477**	NA	5

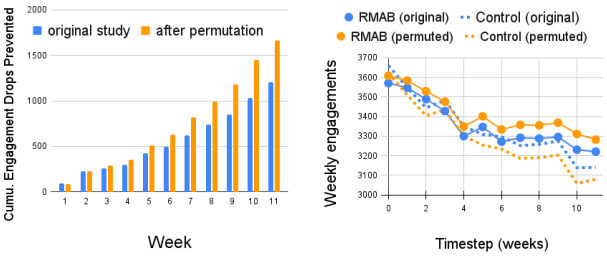


Figure 4. Impact of permutation estimator on real-world data.

group, which sees no interventions. Real-world health workers delivered interventions recommended by the algorithms. We consider the performance lift of the RMAB algorithm in improving engagement with the program in comparison to the control group and apply our proposed permutation algorithm to the originally reported raw results. Figure 4 (left) plots the total engagement numbers on the y-axis as a function of time (in weeks) on the x-axis. Figure 4 (right) computes the lift provided by the RMAB algorithm, as defined in (Mate et al., 2022) on the y-axis. Our findings suggest that the performance lift of RMAB algorithm is larger than originally reported and by week 7, RMAB is estimated to prevent 815 engagement drops, vs the originally reported 622. Since this is real data, the true values are unknown. However, this case study provides evidence that the variance reduction provided by our estimator can be significant in practice.

7. Conclusion

We address the critical gap of mitigating the error in evaluation of resource allocation policies through RCTs. We propose a new estimator using a novel concept based on the idea of retrospective reassignment of participants to experimental arms. We prove that our estimator is unbiased while simultaneously reducing sample variance, and hence reduces error. Through empirical tests on multiple data sets – including a real-world dataset in a socially critical domain – we show that our approach cuts error by as much as 70% and from a single given RCT, can achieve benefits equivalent of running upto 13 *independent* RCTs in parallel.

Acknowledgements

This work was completed in part while the first author (A.M) was an intern at Google. This research was sponsored by the Army Research Office and was accomplished under grant number: W911NF-17-1-0370. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. The authors would also like to thank Luke Miatrix, Akanksha Saran and Eric Rice for valuable inputs during the project.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2022.

Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Athey, S. and Wager, S. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Austin, P. C. and Stuart, E. A. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

Eagle, N., Macy, M., and Claxton, R. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010. doi: 10.1126/science.1186605.

Enderton, H. B. *Elements of set theory*. Academic press, 1977.

Galstyan, A., Czajkowski, K., and Lerman, K. Resource allocation in the grid using reinforcement learning. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, volume 1, pp. 1314–1315. IEEE Computer Society, 2004.

Hayes, R. J. and Moulton, L. H. *Cluster randomised trials*. Chapman and Hall/CRC, 2017.

Hernán, M. A. and Robins, J. M. Causal inference, 2010.

Hudgens, M. G. and Halloran, M. E. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

Imai, K. and Li, M. L. Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*, pp. 1–15, 2021.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pp. 423–439. PMLR, 2014.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Killian, J. A., Wilder, B., Sharma, A., Choudhary, V., Dilkina, B., and Tambe, M. Learning to prescribe interventions for tuberculosis patients using digital adherence data. In *KDD*, 2019.

Kuss, O., Blettner, M., and Börgermann, J. Propensity score: An alternative method of analyzing treatment effects: Part 23 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 113(35-36): 597, 2016.

Lee, E., Lavieri, M. S., and Volk, M. Optimal screening for hepatocellular carcinoma: A restless bandit model. *Manufacturing & Service Operations Management*, 21 (1):198–212, 2019.

Leskovec, J., Adamic, L. A., and Huberman, B. A. The dynamics of viral marketing. *ACM Trans. Web*, 1(1): 5–es, may 2007. ISSN 1559-1131. doi: 10.1145/1232722.1232727. URL https://doi.org/10.1145/1232722.1232727.

Li, L., Munos, R., and Szepesvári, C. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pp. 608–616. PMLR, 2015.

Liu, K. and Zhao, Q. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, pp. 5547–5567, 2010.

Mate, A., Killian, J., Xu, H., Perrault, A., and Tambe, M. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Mate, A., Perrault, A., and Tambe, M. Risk-aware interventions in public health: Planning with restless multi-armed bandits. *Autonomous Agents and Multii-Agent Systems (AAMAS)*, 2021.

Mate, A., Madaan, L., Taneja, A., Madhiwalla, N., Verma, S., Singh, G., Hegde, A., Varakantham, P., and Tambe, M. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. 2022.

Mukerjee, R., Dasgupta, T., and Rubin, D. B. Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113(522):868–881, 2018.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Sutton, R. S., Maei, H., and Szepesvári, C. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. *Advances in neural information processing systems*, 21, 2008.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

Tesauro, G., Jong, N. K., Das, R., and Bennani, M. N. A hybrid reinforcement learning approach to autonomic resource allocation. In *2006 IEEE International Conference on Autonomic Computing*, pp. 65–73. IEEE, 2006.

Ustun, B. and Rudin, C. Learning optimized risk scores. *J. Mach. Learn. Res.*, 20(150):1–75, 2019.

Vaswani, S., Lakshmanan, L., Schmidt, M., et al. Influence maximization with bandits. *arXiv preprint arXiv:1503.00024*, 2015.

Verma, S., Singh, G., Mate, A. S., Verma, P., Gorantala, S., Madhiwalla, N., Hegde, A., Thakkar, D. H., Jain, M., Tambe, M. S., et al. Deployed saheli: Field optimization of intelligent rmab for maternal and child care. 2023.

Wang, Y.-X., Agarwal, A., and Dudık, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597. PMLR, 2017.

Whittle, P. Restless bandits: Activity allocation in a changing world. *J. Appl. Probab.*, 25(A):287–298, 1988.

Wilder, B., Onasch-Vera, L., Diguiseppi, G., Petering, R., Hill, C., Yadav, A., Rice, E., and Tambe, M. Clinical trial of an ai-augmented intervention for hiv prevention in youth experiencing homelessness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14948–14956, 2021.

Zhou, Y., Matsouaka, R. A., and Thomas, L. Propensity score weighting under limited overlap and model misspecification. *Statistical methods in medical research*, 29 (12):3721–3756, 2020.

A. Complete Proofs to Theoretical Results

A.1. Proof of Lemma 4.2

Lemma A.1. *The relation $\dagger$ is an equivalence relation and the family of sets defined by $\mathcal{C}^\dagger(\cdot)$ forms a partition over $\mathcal{C}$.*

Proof. To prove that $\dagger$ is an equivalence relation, we show that it is reflexive, symmetric and transitive. $\dagger$ is reflexive because $\forall \mathfrak{C} \in \mathcal{C}, \mathfrak{C} \in \mathcal{C}^\dagger(\mathfrak{C})$ by definition. Furthermore, $\dagger$ is also trivially symmetric because if $\mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$, then by definition, the allocations received by all individuals at all times are identical under both $\mathfrak{C}_1$ and $\mathfrak{C}_2$. Hence $\mathfrak{C}_1 \in \mathcal{C}^\dagger(\mathfrak{C}_2)$. Finally, $\dagger$ is also transitive because if all allocations received by all individuals at all times are identical in $\mathfrak{C}_1$ and $\mathfrak{C}_2$ as well as in $\mathfrak{C}_2$ and $\mathfrak{C}_3$, that means the allocations are also identical in $\mathfrak{C}_1$ and $\mathfrak{C}_3$. Thus formally, if $\mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$ and $\mathfrak{C}_3 \in \mathcal{C}^\dagger(\mathfrak{C}_2)$, then $\mathfrak{C}_3 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$. Thus $\dagger$ is an equivalence relation over $\mathcal{C}$ and consequently, partitions $\mathcal{C}$ into a family of equivalence classes $\mathcal{C}^\dagger(\cdot)$ such that every element $\mathfrak{C} \in \mathcal{C}$ lies in exactly one partition (Enderton, 1977). □

A.2. Proof of Theorem 4.3

Theorem 4.3. *$Eval^\dagger(\pi_m)$ is an unbiased estimate of the expected value of the performance, $Eval^*(\pi_m)$ $\forall m \in [M]$, defined in equation 1. i.e.*

$$\mathbb{E}_{S_m \sim P^*} \mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}[Eval^\dagger(\pi_m)] = Eval^*(\pi_m) \; \forall m \in [M]$$

Proof.

$$\begin{aligned}
\text{Eval}^*(\pi_m) =& \mathbb{E}_{S_m \sim P^*}\Big[\mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}\big[\text{Eval}(\pi_m)\big]\Big] \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\sum_{\mathfrak{C} \in \mathcal{C}} \text{Prob}[\mathfrak{C}] \cdot \text{Eval}(\pi_m|\mathfrak{C})\Big] \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\frac{1}{|\mathcal{C}|}\sum_{\mathfrak{C} \in \mathcal{C}} \text{Eval}(\pi_m|\mathfrak{C})\Big] && (\because \text{all } \mathfrak{C} \text{ equally likely}) \\
=& \mathbb{E}_{S_m \sim P^*}\Bigg[\frac{1}{|\mathcal{C}|}\Big[\sum_{\mathfrak{C} \in \mathcal{P}_1} \text{Eval}(\pi_m|\mathfrak{C}) + \cdots + \sum_{\mathfrak{C} \in \mathcal{P}_\eta} \text{Eval}(\pi_m|\mathfrak{C})\Big]\Bigg] \\
& \text{where } \{\mathcal{P}_1, \ldots, \mathcal{P}_\eta\} \text{ defines partition of } \mathcal{C} \text{ induced by } \dagger. \\
=& \mathbb{E}_{S_m \sim P^*}\Bigg[\sum_{j \in [\eta]} \frac{|\mathcal{P}_j|}{|\mathcal{C}|} \cdot \frac{1}{|\mathcal{P}_j|} \cdot \Big[\sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi_m|\mathfrak{C})\Big]\Bigg] \\
=& \mathbb{E}_{S_m \sim P^*}\Bigg[\sum_{j \in [\eta]} \frac{|\mathcal{P}_j|}{|\mathcal{C}|} \cdot \Big[\text{Eval}^\dagger(\pi_m|\mathfrak{C})\Big]\Bigg] && (\forall \mathfrak{C} \in \mathcal{P}_j) \\
=& \mathbb{E}_{S_m \sim P^*}\Bigg[\sum_{j \in [\eta]} \text{Prob}[\mathcal{P}_j] \cdot \Big[\text{Eval}^\dagger(\pi_m|\mathfrak{C})\Big]\Bigg] && (\forall \mathfrak{C} \in \mathcal{P}_j) \\
=& \mathbb{E}_{S_m \sim P^*}\Bigg[\sum_{j \in [\eta]} \sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Prob}[\mathfrak{C}] \cdot \Big[\text{Eval}^\dagger(\pi_m|\mathfrak{C})\Big]\Bigg] && \Big(\because \text{Prob}[\mathcal{P}_j] = \sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Prob}[\mathfrak{C}]\Big) \\
=& \mathbb{E}_{S_m \sim P^*}\mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}[\text{Eval}^\dagger(\pi_m)]
\end{aligned}$$

□

A.3. Proof of Theorem 4.4

Theorem 4.4. *The sample variance of our estimator, $Eval^\dagger(\pi)$ is smaller than the standard estimator, $Eval(\pi)$:*

$$\text{Var}(Eval(\pi)) - \text{Var}(Eval^\dagger(\pi)) =$$
$$\frac{1}{|\mathcal{C}|} \cdot \sum_{j \in [\eta]} \Bigg[\sum_{\mathfrak{C} \in \mathcal{P}_j} Eval^2(\pi|\mathfrak{C}) - \frac{\Big(\sum_{\mathfrak{C} \in \mathcal{P}_j} Eval(\pi|\mathfrak{C})\Big)^2}{|\mathcal{P}_j|}\Bigg]$$

≥ 0, *where* $\{\mathcal{P}_1, \ldots, \mathcal{P}_\eta\}$ *is the partition of* $\mathcal{C}$ *induced by* $\dagger$.

Proof. We compute the sample variance by first conditioning over the partition $\mathcal{P}_j$ (of the equivalence sets defined by $\dagger$) that an instance of an assignment, $\mathfrak{C}$ belongs to and then accounting for the variance stemming from the candidate assignments $\mathfrak{C}$ within the partition. Thus we get:

$$\begin{aligned}
\mathrm{Var}(\mathrm{Eval}(\pi)) &= \frac{1}{|\mathcal{C}|}\sum_{\mathfrak{C}\in\mathcal{C}}\big(\mathrm{Eval}(\pi|\mathfrak{C}) - \mathrm{Eval}^*(\pi)\big)^2 \\
&= \frac{1}{|\mathcal{C}|}\sum_{\mathrm{j}\in[\eta]}\sum_{\mathfrak{C}\in\mathcal{P}_j}\big(\mathrm{Eval}(\pi|\mathfrak{C}) - \mathrm{Eval}^*(\pi)\big)^2 \\
&= \frac{1}{|\mathcal{C}|}\sum_{\mathrm{j}\in[\eta]}\Bigg(\sum_{\mathfrak{C}\in\mathcal{P}_j}\big(\mathrm{Eval}(\pi|\mathfrak{C})\big)^2 - 2\,\mathrm{Eval}^*(\pi)\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C}) + |\mathcal{P}_j|\big(\mathrm{Eval}^*(\pi)\big)^2\Bigg)
\end{aligned} \tag{7}$$

Similarly, we compute the variance of our estimator $\mathrm{Eval}^\dagger$ as:

$$\begin{aligned}
\mathrm{Var}(\mathrm{Eval}^\dagger(\pi)) &= \frac{1}{|\mathcal{C}|}\sum_{\mathrm{j}\in[\eta]}\sum_{\mathfrak{C}\in\mathcal{P}_j}\big(\mathrm{Eval}^\dagger(\pi|\mathfrak{C}) - \mathrm{Eval}^*(\pi)\big)^2 \\
&= \frac{1}{|\mathcal{C}|}\sum_{\mathrm{j}\in[\eta]}\left(|\mathcal{P}_j|\cdot\left\{\frac{\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})}{|\mathcal{P}_j|} - \mathrm{Eval}^*(\pi)\right\}^2\right) \\
&= \frac{1}{|\mathcal{C}|}\sum_{\mathrm{j}\in[\eta]}\left(|\mathcal{P}_j|\cdot\left\{\left(\frac{\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})}{|\mathcal{P}_j|}\right)^2 - 2\,\mathrm{Eval}^*(\pi)\left(\frac{\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})}{|\mathcal{P}_j|}\right) + \Big(\mathrm{Eval}^*(\pi)\Big)^2\right\}\right) \\
&= \frac{1}{|\mathcal{C}|}\sum_{\mathrm{j}\in[\eta]}\left\{\frac{\Big(\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})\Big)^2}{|\mathcal{P}_j|} - 2\,\mathrm{Eval}^*(\pi)\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C}) + |\mathcal{P}_j|\big(\mathrm{Eval}^*(\pi)\big)^2\right\}
\end{aligned} \tag{8}$$

Subtracting expression in Equation 8 from the expression in Equation 7 gives:

$$\mathrm{Var}(\mathrm{Eval}(\pi)) - \mathrm{Var}(\mathrm{Eval}^\dagger(\pi)) = \frac{1}{|\mathcal{C}|}\sum_{\mathrm{j}\in[\eta]}\left(\sum_{\mathfrak{C}\in\mathcal{P}_j}\big(\mathrm{Eval}(\pi|\mathfrak{C})\big)^2 - \frac{\Big(\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})\Big)^2}{|\mathcal{P}_j|}\right) \tag{9}$$

We can show that the expression for variance contraction derived in Equation 9 is non-negative as a direct consequence of the Cauchy-Schwarz inequality. The Cauchy-Schwarz inequality states that for two vectors $\mathbf{u}$ and $\mathbf{v}$, $|\langle\mathbf{u},\mathbf{v}\rangle|^2 \leq \langle\mathbf{u},\mathbf{u}\rangle\cdot\langle\mathbf{v},\mathbf{v}\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the inner product. Setting $\mathbf{u} = \underbrace{\Big[\mathrm{Eval}(\pi|\mathfrak{C}_1), \ldots, \mathrm{Eval}(\pi|\mathfrak{C}_{|\mathcal{P}_j|})\Big]}_{|\mathcal{P}_j|\text{entries}}$) and $\mathbf{v} = \underbrace{\Big[1, \ldots, 1\Big]}_{|\mathcal{P}_j|\ 1's}$ yields the desired result: $\mathrm{Var}(\mathrm{Eval}(\pi)) - \mathrm{Var}(\mathrm{Eval}^\dagger(\pi)) \geq 0$. □

A.4. Efficient Algorithm

Lemma A.2. *The relation* $\dagger_\Upsilon$ *is an equivalence relation over both* $\mathcal{C}$ *as well as each set in the family* $\mathcal{C}^\dagger(\cdot)$ *and the family of sets defined by* $\mathcal{C}^\dagger_\Upsilon(\cdot)$ *forms a partition over* $\mathcal{C}$.

Proof. Similar to Lemma 4.2, we prove that $\dagger_\Upsilon$ is an equivalence relation by showing that it is reflexive, symmetric and transitive. $\dagger_\Upsilon$ is reflexive because $\forall\mathfrak{C}\in\mathcal{C}, \mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon(\mathfrak{C})$ by definition. Furthermore, $\dagger$ is also trivially symmetric because if $\mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$, then by definition, the index thresholds, as well as the allocations received by all individuals at all times, are identical under both $\mathfrak{C}_1$ and $\mathfrak{C}_2$. Hence $\mathfrak{C}_1 \in \mathcal{C}^\dagger(\mathfrak{C}_2)$. Finally, $\dagger$ is also transitive because if all index threshold and allocations received by all individuals at all times are identical in $\mathfrak{C}_1$ and $\mathfrak{C}_2$ as well as in $\mathfrak{C}_2$ and $\mathfrak{C}_3$, that means the same are also identical in $\mathfrak{C}_1$ and $\mathfrak{C}_3$. Thus formally, if $\mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$ and $\mathfrak{C}_3 \in \mathcal{C}^\dagger(\mathfrak{C}_2)$, then $\mathfrak{C}_3 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$. Thus $\dagger_\Upsilon$ is an equivalence relation over $\mathcal{C}$ and consequently, partitions $\mathcal{C}$ into a family of equivalence classes $\mathcal{C}^\dagger(\cdot)$ such that every element $\mathfrak{C}\in\mathcal{C}$ lies

in exactly one partition (Enderton, 1977). Similar reasoning also shows that $\dagger_\Upsilon$ is an equivalence relation over each set in the family $\mathcal{C}^\dagger(\cdot)$. □

Corollary A.3. *$Eval^\dagger_\Upsilon(\pi_m)$ is an unbiased estimate of the expected value of the performance, $Eval^*(\pi)$, defined in equation 1. i.e. $\mathbb{E}_{S_m\sim P^*}\mathbb{E}_{\mathfrak{C}\sim\mathcal{C}}[Eval^\dagger(\pi_m)] = Eval^*(\pi_m)\ \forall m \in [M]$*

Proof. Using Lemma A.2, we apply similar arguments as Theorem 4.3 on the partition defined by $\dagger_\Upsilon$ to show that $\text{Eval}^\dagger_\Upsilon$ yields an unbiased estimate. □

Theorem 5.1. *$Eval^\dagger_\Upsilon(\cdot)$ computed as per Equation 6 computes the average of $Eval(\pi|\mathfrak{C})$ over all assignments in $\mathcal{C}^\dagger_\Upsilon$. i.e. $Eval^\dagger_\Upsilon(\pi) = \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon} Eval(\pi|\mathfrak{C})}{|\mathcal{C}^\dagger_\Upsilon|}$*

Proof. The key to showing that the two are equivalent is in interpreting the summation of (modified) rewards over individuals in the view of average group rewards over assignments. Mathematically, starting from the definition of $\text{Eval}^\dagger_\Upsilon(\pi_m)$, the key lies in moving the summation operation over assignments in $\mathcal{C}^\dagger_\Upsilon$ from outside the $\text{Eval}^\dagger_\Upsilon()$ term to inside, applying it individually on each contributing participant. Formally, we can rewrite the expression of $\text{Eval}^\dagger_\Upsilon(\pi_m)$ as:

$$\text{Eval}^\dagger_\Upsilon(\pi_j) = \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon} \text{Eval}(\pi_j|\mathfrak{C})}{|\mathcal{C}^\dagger_\Upsilon|} \tag{10}$$

$$= \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon} \sum_{i\in C_j} r(S[i], A[i])}{|\mathcal{C}^\dagger_\Upsilon|} \tag{11}$$

$$= \sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon} \frac{\sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k} \mathbb{1}_{\{i\in C_j\}}\cdot r(S[i], A[i]) + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i])}{|\mathcal{C}^\dagger_\Upsilon|} \tag{12}$$

(splitting the summation over groups $\mathbf{G}_k$ and other individuals that can't be swapped) (13)

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon} \frac{\mathbb{1}_{\{i\in C_j\}}\cdot r(S[i], A[i])}{|\mathcal{C}^\dagger_\Upsilon|}\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{14}$$

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon} \frac{\mathbb{1}_{\{i\in C_j\}}}{|\mathcal{C}^\dagger_\Upsilon|}\cdot r(S[i], A[i])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{15}$$

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\Pr(i\in C_j|\mathcal{C}^\dagger_\Upsilon)\cdot r(S[i], A[i])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{16}$$

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\frac{|\{\iota : \iota\in(\mathbf{G}_{\phi(i)}\cap C_j)\}|}{|\mathbf{G}_{\phi(i)}|}\cdot r(S[i, 0:T], A[i, 1:T])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{17}$$

$$= \sum_{k\in\kappa}|\{\iota : \iota\in(\mathbf{G}_{\phi(i)}\cap C_j)\}|\cdot\sum_{i\in\mathbf{G}_k}\left[\frac{1}{|\mathbf{G}_{\phi(i)}|}\cdot r(S[i, 0:T], A[i, 1:T])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{18}$$

$$= \sum_{k\in\kappa}|\{\iota : \iota\in(\mathbf{G}_{\phi(i)}\cap C_j)\}|\cdot\tilde{r}_k + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i]) \tag{19}$$

$$= \sum_{i\in C_j}\tilde{r}_{\phi(i)}\cdot\Lambda_i + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i], A[i], 1:T) \tag{20}$$

□

Table 1. Sample variance in Measured Performance Lift

T	B	π_1 V π_0	RAW	PERMUTED	IPW	n-VAL
1	3%	π_{WI} V π_{GR}	49.09	4.94	**0.48**	9
1	10%	π_{WI} V π_{GR}	49.86	15.11	**6.66**	3
1	25%	π_{WI} V π_{GR}	49.45	**19.94**	78.12	2
10	3%	π_{WI} V π_{WI}	2381	**916**	NA	3
10	3%	π_{WI} V π_{GR}	2348	**728**	NA	4
10	3%	π_{GR} V π_{GR}	26356	**1860**	NA	13
10	10%	π_{GR} V π_{GR}	25983	**3808**	NA	7
10	25%	π_{GR} V π_{GR}	23619	**5477**	NA	5

B. Casting Resource Allocation Policies as Index Policies

Control: A control group that sees no interventions can be handled by using any randomly generated index matrix with finite entries. Setting the index threshold $\Upsilon_i = \infty \; \forall \; i$ ensures that no individual assigned the control policy gets picked for intervention.

Round Robin: Common policies such as 'round robin', that operate by selecting individuals cyclically for intervention in a set order can also be represented as index policies. The index for each individual at each time step, can be determined in two stages. First, we consider the feature used for ranking the N individuals and we start by setting $\Upsilon_i(t) := r$, for $r \in \{1, \dots N\}$ where r denotes the priority rank of the individual (highest rank picked first). Next, each time an individual receives an action $a = 1$, we want to push them at the bottom of the queue, so we subtract N from their index for all future timesteps, repeating this process for each instance of $a = 1$.

C. Additional Experimental Results

C.1. Synthetic Data Generation in Section 6.1 Explained

$$P^{p,1} = \begin{bmatrix} 0.97 & 0.03 \\ 0.03 & 0.97 \end{bmatrix} \; P^{a,1} = \begin{bmatrix} 0.96 & 0.04 \\ 0.01 & 0.99 \end{bmatrix}$$
$$P^{p,2} = \begin{bmatrix} 0.25 & 0.75 \\ 0.03 & 0.97 \end{bmatrix} \; P^{a,2} = \begin{bmatrix} 0.23 & 0.77 \\ 0.01 & 0.99 \end{bmatrix}$$

Figure 5. Probability values forming the matrix P_1 and P_2

We reproduce the transition probabilities P_1 and P_2 used in our simulation, adopted from (Mate et al., 2020) in Figure 5. Each P comprises of a set of probabilities under each of the two actions ($a = 0$, denoted as 'p', for passive and $a = 1$ denoted as 'a', for active).

Intuition is that P_1 has a very small $P^a_{0,1}$ and $P^p_{0,1}$ and is thus difficult to revive once it enters state $s = 0$, even with an intervention ($a = 1$), making it important to keep intervening to stop the individual from ever entering $s = 0$. On the other hand, P_2 has a large $P^p_{0,1}$, making it self-correcting, meaning the individual is likely to return to $s = 1$ quickly even without intervention.

C.2. Single-shot RCTs

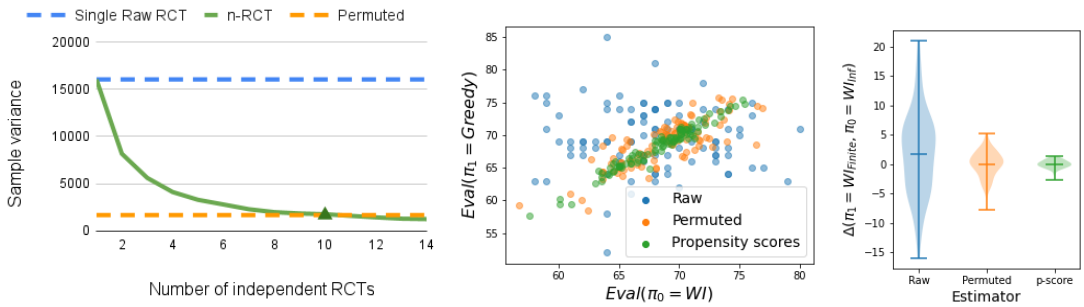


Figure 6. (Left)Variance reduces by running and averaging over n-independently run trials. (Center, Right)Single shot setup

C.3. Sequential RCTs

We run more comparisons using $N = 100$ individuals per arm, simulating 500 instances of trials for $T = 10$ timesteps. The $n-$values are listed in Table 1.

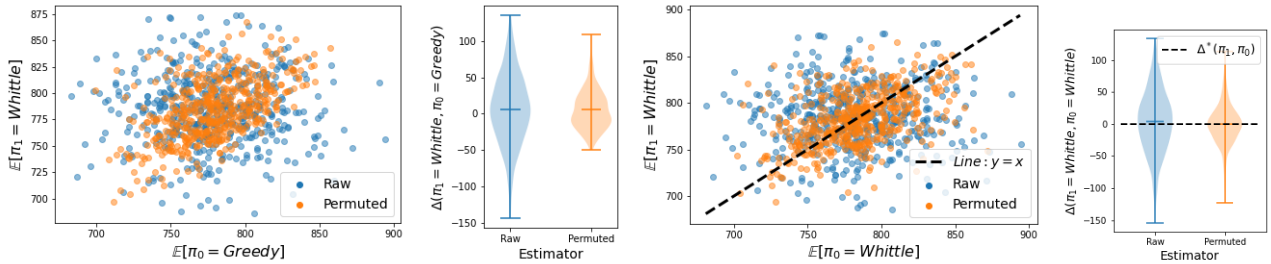


Figure 7. Whittle vs Greedy (left two panels) and Whittle vs whittle (right two panels)

Setup: Greedy vs Greedy For more granular analysis, we consider the state trajectories of individuals participating in the trial. This uses $N = 1000$ individuals per arm and simulates 30 instances of trials for $T = 10$ timesteps. We see that orange trajectories (after permutation) is closer to the expected value than blue trajectories (blue)

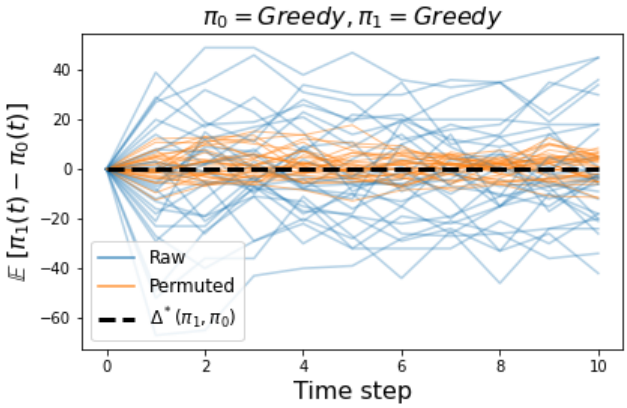


Figure 8. Time trajectories of Greedy v Greedy