A federated Learning Framework for Ethical Dynamic Treatment Allocation across Heterogeneous Hospitals

Xenia Konti[1]; Nicoleta J Economou-Zavlanos, PhD[5]; Yi Shen[3]; Giorgos Stamou, PhD[6] ; Armando Bedoya, M.D[4,5]; Michael J. Pencina, PhD[5] ;Chuan Hong, PhD[3] ; Michael M. Zavlanos, PhD[1,2,3]*

[1]Dept. of Computer Science, Duke University

[2]Dept. of Electrical and Computer Engineering, Duke University

[3]Dept. of Mechanical Engineering & Material Sciences, Duke University

[4]Dept. of Biostatistics & Bioinformatics, Duke University

[5]School of Medicine, Duke University

[6]School of Electrical and Computer Engineering, National Technical University of Athens

Keywords: Federated Learning, Causal Inference, Multi-Armed Bandit, treatment allocation

ABSTRACT

Objective. In this paper, we propose an adaptive federated learning framework to learn optimal treatments for individual hospitals that possibly serve different patient populations. The proposed framework can enable the design of more efficient treatment allocation problems.

Methods. We propose a federated treatment recommendation strategy that for each hospital is formulated as a Multi-Armed Bandit (MAB) problem. The process is coordinated by a lead hospital that adaptively learns and transfers Upper Confidence Bounds (UCB) across similar hospitals and Personalized Upper Bounds across heterogeneous hospitals. We test our proposed method on a simulated clinical trial environment created using real Covid-19 data from the Duke University Health System.

Results. Our method relies on collaboration among hospitals, which allows for fewer data samples needed per institution, while protecting the privacy of the individual patient data. At the same time, it ensures fairness of the learned treatments by mitigating possible biases due to differences in the patient populations treated across different hospitals. Finally, our method improves the safety of the learning procedure by reducing the number of patients administered with sub-optimal treatments at each hospital. In the experiments, we show that our proposed method outperforms other state of the art approaches in that it requires up to 36%-75% fewer patient data to learn the optimal treatment for each hospital and administers the optimal treatment to 0.95%-48.6% more patients.

Conclusion. In this paper, we propose an adaptive federated learning strategy for treatment recommendation tasks, that learns optimal treatments for individual hospitals that possibly serve different patient populations, while satisfying privacy, fairness, and safety considerations.

1. INTRODUCTION

1.1 Background

Providing personalized and optimized treatment recommendations is a critical challenge in the healthcare industry, with broad applications. Adaptive treatment strategies are increasingly essential in chronic disease management (e.g., diabetes, hypertension), oncology, mental health care, emergency medicine, post-acute rehabilitation, where treatment efficacy and safety often depend on patient-specific factors such as comorbidities, demographics, and genetic profiles [1, 2]. In clinical trials, personalized allocation can reduce sample size requirements and improve trial efficiency through adaptive randomization [3]. In real-world healthcare delivery, intelligent decision support systems can dynamically recommend treatments based on evolving clinical data and institutional constraints [4]. AI-enabled population health platforms also enable stratified care pathways, helping health systems target interventions to high-risk individuals while reducing unnecessary treatment for lower-risk patients [5]. As artificial intelligence (AI) and machine learning technologies continue to mature, there is a growing opportunity to develop adaptive frameworks that leverage electronic health records (EHR), imaging, genomics, and patient-reported outcomes to generate data-driven treatment plans that are both scalable and personalized, ultimately improving safety, efficacy, and equity in healthcare delivery.

Providing adaptive treatment recommendations in both clinical trials and real-world healthcare settings is typically formulated as an adaptive treatment allocation problem. This refers to a sequential decision-making procedure in which the probability of administering a specific treatment to a patient, changes based on previous observations. Multi-Armed Bandit (MAB) algorithms [6] are particularly suited to solve this problem, e.g., for facilitating treatment allocation [7] and dose-finding [8, 9] or for tailoring treatment strategies [10], as they can effectively navigate the balance between exploration and exploitation of treatments. On top of this, when treatment

allocation strategies need to be personalized to each individual, the contextual variation of MAB is utilized for covariate-based decision making [10, 11, 12]. However, one critical barrier when deploying MAB algorithms is that MAB performance can degrade substantially when applied to small datasets, which often arise in observational healthcare settings or when off-policy evaluation is required [13]. This challenge is further exacerbated by the difficulty of recruiting enough patients in clinical trials, where participation rates are typically low [14, 15]. Thus, while MAB algorithms are well-suited to many healthcare decision-making problems in theory, their effectiveness in practice is often hindered by data scarcity. To mitigate this issue, federated learning methods can be used to leverage distributed knowledge from similar machine learning procedures conducted across different hospitals. Recent advances in federated MAB algorithms [16, 17] have typically involved sharing model parameters and combining information only across similar institutions. In contrast, our framework exchanges only treatment–outcome pairs and returns information that leverages data from both similar and dissimilar hospitals, enabling broader and more robust collaboration. Moreover, the deployment of such approaches in healthcare remains under-explored.

1.2 Significance

Two well-known challenges in the real-world application of treatment recommendation systems are the lack of subject participation and the possible risks for patients due to sub-optimal treatment allocation strategies. To address these challenges, we propose a new cooperative adaptive treatment allocation design that relies on federated learning to allow hospitals to share aggregated information and, as a result, learn their optimal treatment strategies even with fewer participants. Since sharing individual patient data across hospitals raises privacy concerns, we design our framework to learn hospital-specific policies, which aggregate knowledge at the institutional level without exposing individual records. To support effective collaboration under this design, we

incorporate clustering and causal inference techniques [18], which enable knowledge sharing among both similar and heterogeneous hospitals. Finally, these techniques are used in a MAB framework with a focus on upholding patient privacy [19] (since no raw patient-level data are shared across hospitals), ensuring fairness [20], and improving participant safety (as our allocation process prioritizes assigning the most promising treatment rather than randomizing arbitrarily).

To evaluate our proposed design, we focused on a Covid-19 treatment allocation problem. Given the urgent need for new Covid treatments at the time [21], numerous trials were conducted at healthcare facilities around the world. Under those circumstances, federated learning would have provided an excellent opportunity to help coordinate and expedite the discovery of effective Covid treatments. Specifically, we use Covid-19 data available in the Duke Health System to create a simulation environment and test our method's performance, as it is usually done in similar medical applications [22]. We create this environment with a similar strategy as illustrated in [23], using retrospective electronic health records data (EHR) from three Duke hospitals, which allows us to accurately replicate real-world scenarios and rigorously test our model's efficacy and adaptability. Finally, we have designed our method to align with TRIPOD guidelines as described in the supplementary materials. To our knowledge, this is the first instance of a federated learning approach that amalgamates clustering methods and causal inference within an online MAB framework for treatment allocation tasks.

Statement Of Significance

- **Problem:** Multiple hospitals are engaged in the same treatment allocation task, yet these efforts are typically conducted in isolation.
- **What is Already Known:** Prior research has explored adaptive treatment allocation strategies, which aim to optimize patient outcomes during the learning phase by

dynamically adjusting treatment assignments. However, existing approaches do not leverage the fact that similar learning tasks are being independently undertaken at multiple institutions.

- **What this paper Adds:** This paper introduces a novel framework for *Collaborative Adaptive Learning* in the context of treatment allocation. The proposed method facilitates knowledge sharing among homogeneous hospitals (homogeneous hospitals), and even enables knowledge transfer across institutions serving demographically heterogeneous patient populations.
- **Who would benefit from this knowledge in this paper:** Clinicians and healthcare practitioners seeking to implement adaptive treatment strategies with limited patient data. The proposed approach enhances learning efficiency and patient safety by utilizing insights gained collaboratively across institutions.

2. MATERIALS AND METHODS

2.1 Problem Set Up

We consider a collection of H non-identical hospitals serving different patient populations in terms of both patient distributions and patient numbers. Each hospital $h \in \mathcal{H} = \{1, \cdots, H\}$ conducts a treatment allocation task with the objective to learn a treatment $x_h^* \in \mathcal{X} = \{1, \cdots, X\}$ among X possible treatment options, that is optimal for the average patient in its local population group. We formulate the treatment allocation strategy at every hospital h as a Multi-Armed Bandit (MAB) problem defined by the tuple $(\mathcal{X}, \{Y_h(x)\}_{x \in \mathcal{X}})$, where $x \in \mathcal{X}$ denotes a treatment option and $Y_h(x) = P(y|x)$ is an unknown probability distribution of the treatment outcome y given a treatment $x \in \mathcal{X}$ that depends on the patient population that each hospital serves. Here we assume a binary treatment outcome $y = \{0,1\}$ so that $y = 1$ if a patient responded well to a

treatment and $y = 0$ otherwise. Furthermore, we assume that patient arrivals at hospitals follow a distribution $P(h) \forall h \in \mathcal{H}$. Then, for every new patient that arrives at hospital h, a treatment $x_{h,t} \in \mathcal{X}$ is selected and a reward $y_h(x_{h,t}) \sim Y_h(x_{h,t})$ is observed. The optimal treatment strategy $x_h^* \in \mathcal{X}$ at hospital h is defined as the one that maximizes the expected reward up to time step t, i.e., $x_h^* = argmax_{x \in \mathcal{X}} \mathbb{E}[Y_h(x)]$.

To learn the optimal treatment strategy x_h^* at each local hospital h and accurately estimate the unknown distribution of treatment outcomes $Y_h(x_h^*)$, the MAB needs to identify the most effective treatment by (i) exploring different treatment options and (ii) exploiting the current best treatment $x_{h,t}^*$ to increase confidence in the current best treatment and maximize patient benefit from this treatment. Managing this balance between exploration and exploitation in a MAB problem is key in learning the optimal treatment strategy fast, using as few data samples as possible.

2.2 Proposed Method

To improve the data efficiency of the MAB problem, we propose a federated approach in which hospitals collaborate through a designated lead institution. Each hospital continues to operate its own treatment allocation task locally but shares with the lead hospital the treatment x administered to patient t along with the corresponding binary outcome $y_{t,h}(x)$. Importantly, no patient-level records are exchanged, ensuring that sensitive data remain within each institution and preserving patient privacy. Once the lead hospital has received T data samples across all participating hospitals, a **Communication Round (CR)** is triggered.

During each CR, the lead hospital determines what information to return to every participating hospital. For each hospital h, it identifies a set of hospitals S_h that are sufficiently similar in terms of their estimated optimal treatment and its expected reward via the following criterion:

$$S_h = \left\{ \tilde{h} \in H \;\middle|\; x_h^\star = \; x_{\tilde{h}}^\star \; and \left| Y_h(x_h^\star) - Y_{\tilde{h}}(x_{\tilde{h}}^\star) \right| \leq \epsilon \right\}$$

where:

- $x_h^\star$ is the current estimated optimal treatment at hospital h,
- $Y_h(x_h^\star)$ is its expected reward,
- ϵ is a tunable similarity threshold

This criterion ensures hospitals are clustered only when they share the same optimal treatment and have similar estimated performance.

Once the cluster for each hospital is defined, for each treatment x, the lead hospital aggregates all observed outcomes from the hospitals in that cluster between the current and the previous CR. Let t denote the patient index within each hospital, and $x_{h,t}$ the treatment chosen that patient by hospital h. Then, the **accumulated reward** for treatment x across the hospitals in S_h is

$$y_{acc}^{S_h}(x) = \sum_{\tilde{h}\in S_h} \sum_{t\in CR: x_{\tilde{h},t}=x} y_{t,\tilde{h}}(x),$$

and the **number of times** treatment x has been administered within that cluster is

$$N^{S_h}(x) = \sum_{\tilde{h}\in S_h} |\{t: x_{\tilde{h},t} = x\}|.$$

The **total number of treatments administered** across all treatments in that cluster is then:

$$\overline{N}^{S_h} = \sum_{x\in X} N^{S_h}(x)$$

Using these aggregated quantities, the lead hospital computes the **Aggregate Upper Confidence Bound (AggrUCB)** for each treatment x and each hospital h as:

$$AggrUCB\big(Y_h(x)\big) = \frac{y_{acc}^{S_h}(x)}{N^{S_h}(x)} + c\sqrt{\frac{\log(\overline{N}^{S_h})}{N^{S_h}(x)}}$$

where c is a positive exploration coefficient controlling how aggressively the algorithm balances exploration (trying new treatments) versus exploitation (selecting the currently best-performing treatment). This form computes classical UCB formula **over aggregated cluster-level data**

rather than individual hospital data, which provide hospital h with a more informed Upper Confidence Bound.

For hospitals that are not considered similar according to the previous criterion, the lead hospital computes **Personalized Upper Bounds on Treatment Effects (PUBound)**. These bounds allow safe knowledge transfer across heterogeneous hospitals where patient populations and treatment responses differ.

Based on the causal inference formulation in [18], the lead hospital uses the aggregated information from all the hospitals to compute upper and lower bounds on the expected reward of each treatment:

$$\mathrm{L}_{Causal}(\mathrm{Y}(\mathrm{x})) = \mathrm{Prob}(\mathrm{Treatment} = \mathrm{x}, \mathrm{y} = 1),$$

$$\mathrm{U}_{Causal}\big(\mathrm{Y}(\mathrm{x})\big) = 1 - \mathrm{Prob}(\mathrm{Treatment} = \mathrm{x}, \mathrm{y} = 0),$$

Furthermore, using information on the patient distribution across hospitals, the lead hospital can analyze the expected reward of each treatment for each hospital individually as follows:

$$Y(x) = \sum_{h \in H} P(h) Y_h(x) \quad (1)$$

Based on equation (1) and the causal upper and lower bounds computed before, for each hospital h and treatment x the lead hospital then computes Personalized Upper Bound for $Y_h(x)$ by solving the following optimization problem:

maximize $UBound(Y_h(x))$

subject to $L_{Causal}\big(Y(x)\big) \leq \sum_{h \in H} P(h) UBound\big(Y_h(x)\big) \leq U_{Causal}(Y(x)),$

$$LCB\big(Y_h(x)\big) \leq UBound\big(Y_h(x)\big) \leq UCB\big(Y_h(x)\big) \forall h \in H.$$

where LCB and UCB denote the lower and upper aggregated confidence bounds:

$LCB\big(Y_h(x)\big) = \frac{y_{acc}^{S_h}(x)}{N^{S_h}(x)} - c' \sqrt{\frac{\log(\bar{N}^{S_h})}{N^{S_h}(x)}}$ and $UCB\big(Y_h(x)\big) = \frac{y_{acc}^{S_h}(x)}{N^{S_h}(x)} + c' \sqrt{\frac{\log(\bar{N}^{S_h})}{N^{S_h}(x)}}$.

Here, c′ is a confidence coefficient controlling how conservative the personalized bounds are. This optimization yields $UBound(Y_h(x))$, a Personalized Upper Bound that incorporates knowledge from dissimilar hospitals while preserving privacy.

The results of this analysis—AggrUCB values for clusters of similar hospitals and PUBound values for dissimilar ones—are communicated back to the local institutions. Each hospital integrates these quantities into its own MAB algorithm, replacing the standard UCB with the tighter of the two values. In this way, hospitals avoid unnecessary exploration of treatments already known to perform poorly, while still retaining the ability to discover effective strategies for their unique populations. This iterative process of patient arrivals, local treatment assignment, periodic communication, and federated updating gives rise to our proposed **CausalAdapUCB** algorithm. An illustration of the method is provided in Figure 1, a complete mathematical description is available in the Supplementary Appendix, and our implementation can be accessed at: https://github.com/xeniakonti/FL-Framework-for-treatment-allocation.

Figure 1. The figure describes the federating process suggested in the paper. The hospitals that participate in the system may serve similar or different patient populations, and they model their corresponding treatment allocation task as a Multi-Armed Bandit problem. The federating process is coordinated by a Lead Hospital that is responsible for clustering hospitals according to how similar their patient populations are and then computing and transferring information that is personalized to each hospital individually.

2.3. Methods for Comparison

To evaluate the effectiveness of our method, we benchmark it against several existing treatment assignment strategies commonly used in allocation problems, including (1) **RandTrial** [24, 25]: a randomized treatment assignment, where for every patient the treatment is selected randomly, (2) **LocalUCB** [26]: a standard multi-armed bandit algorithm where each hospital independently executes a UCB policy using only its local data, without any form of collaboration, (3) **GlobalUCB**: a fully pooled approach in which all hospitals are treated as a single entity, and a global UCB is computed using data from the entire population, ignoring hospital-level heterogeneity, and (4) **AdapUCB**: a partially collaborative design where a lead hospital computes and transfers only the aggregate UCB value across institutions, without propagating personalized upper bounds tailored to local hospital populations. Together, these methods provide meaningful points of comparison: RandTrial as a traditional clinical baseline, LocalUCB and GlobalUCB as canonical MAB implementations at the local and global levels, and AdapUCB as an intermediate ablation to isolate the contribution of personalized upper bounds. We deploy each of these methods independently within the test environments we design and compare their performance against our proposed approach. Details of the evaluation metrics and simulation environments are provided in the subsequent sections.

2.4. Dataset Description

We use historical data of Duke patients diagnosed and treated for Coronavirus disease 19 (Covid-19) from year 2020 to year 2021, obtained from the Duke Clinical Research Data Mart, which provides access to Duke patient data since the beginning of 2014. Specifically, the dataset consists of 21,482 different patients who had at least one U07.1 (“COVID-19, virus identified”), an ICD-10 diagnosis code or a positive SARS-CoV-2 reverse transcription polymerase chain reaction (PCR) test result recorded within the healthcare data. Every patient in the dataset may have

visited a Duke hospital multiple times. Each patient visit is considered a new patient encounter and is characterized by a unique encounter id. During each encounter, a patient may have received multiple medications. We only consider inpatient encounters, that is patients who were treated solely at Duke hospitals and didn't visit non-Duke hospitals. For every patient encounter, the dataset includes information on the patient's demographic characteristics (e.g. age, race), the dates they were administered Covid-19 medications during their visit, the type of medications, and the treatment outcome, i.e., whether they died and when. Figure 2c shows the patient distribution in terms of their age and race. The treatments used during the period that the data were collected are Treatments = {Ritonavir, Bamlanivimab, Casirivimab-Imdevimab, Remdesivir, Ritonavir Nirmatrelvir, Sotrovimab, Bamlanivimab Etesevimab}.

2.5. Simulation Environment

To simulate an environment consisting of multiple heterogeneous hospitals conducting the same treatment allocation task, we split the data into different groups, each one modeling an individual hospital. We introduce heterogeneity in the hospital populations by biasing each hospital's data with respect to the patients' age. Specifically, 82.8% of patients that visit Hospital 1 are younger than 40, 17% are between 40 and 80, and 0.2% are older than 80. Additionally, 23.3% of patients that visit Hospital 2 are younger than 40, 76.3% are between 40 and 80, and 0.3% are older than 80. Finally, 59.4% of patients that visit Hospital 3 are younger than 40, 26.6% are between 40 and 80, and 13% are older than 80. In this way, Hospital 1 is assigned 49% of the total patient population in the historical dataset, Hospital 2 is assigned 37% of it, and Hospital 3 is assigned 14% of it. The distribution of patients across the three hospitals in terms of their age and race is shown in Figure 2a-2b. To evaluate the different treatment options at the three simulated hospital environments, we use the following two metrics: i) whether any patient died after receiving a treatment and ii) whether they were re-hospitalized.

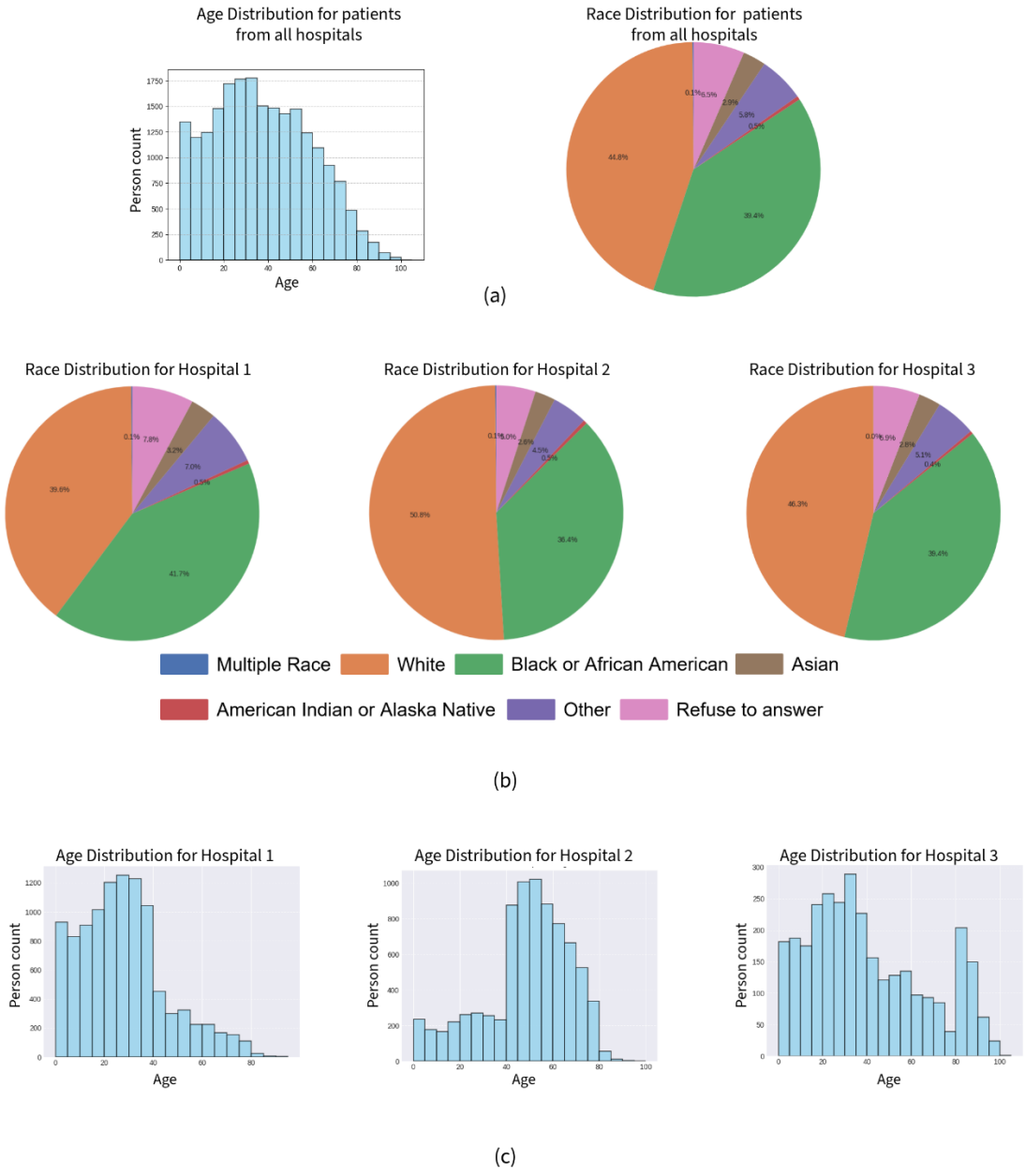


Figure 2. This figure displays the demographic distribution of Duke patients diagnosed with COVID-19, broken down by age and race. Subfigure 2a presents the aggregated age and race distribution across all hospitals. Subfigure 2b shows the race distribution for each of the three

hospitals considered in our experiments, and subfigure 2c shows the corresponding age distribution.

Name of Treatment	Hospital 1	Hospital 2	Hospital 3
Real Data with Death Event Reward			
Casirivimab-Imdevimab	0.988	0.981	0.96
Remdesivir	0.86	0.81	0.76
Real Data with Death Event or Readmission Reward			
Casirivimab-Imdevimab	0.63	0.62	0.6
Remdesivir	0.79	0.75	0.7
Bamlanivimab-Etesevimab	0.61	0.54	0.55
Artificial Simulation Environment			
Treatment 1	0.3	0.9	0.2
Treatment 2	0.8	0.5	0.7
Treatment 3	0.6	0.1	0.9

Table 1. Summary of the expected reward of each treatment for each hospital for different outcomes, with artificial hospital segmentation and optimal medication strategies.

2.6. Real Data Simulation with Death Event as a reward metric

In this case, we consider the death event metric, and every time a treatment is administered to a patient, we examine whether the patient died within 7 to 30 days after receiving the treatment. If the patient dies, we assign the treatment a reward 0, and a reward 1 otherwise. We note that in the initial dataset, most of the treatments were not administered frequently enough for accurate reward estimation. In fact, only three of the seven treatments—Casirivimab-Imdevimab, Remdesivir, and Bamlanivimab-Etesevimab—had sufficient data. Casirivimab-Imdevimab and Bamlanivimab-Etesevimab showed similar to each other and superior to Remdesivir success rates. From a clinical perspective Casirivimab–Imdevimab and Bamlanivimab–Etesevimab

are both dual monoclonal antibody therapies authorized for the treatment of mild-to-moderate COVID-19 in high-risk outpatients, sharing similar mechanisms of action, administration routes, and clinical use contexts. They were often used interchangeably based on availability and were considered functionally equivalent by treatment guidelines during their authorized periods. A real-world effectiveness study by Wynn et al. (2022) found an 86% probability of equivalence between the two combinations in terms of hospital-free days by day 28, supporting their clinical comparability in routine care settings [27]. Consequently, we combined Casirivimab-Imdevimab and Casirivimab-Imdevimab into one treatment strategy (by combining their data) so that this combined treatment strategy is now the new optimal one. As a result, in this case, the simulation environment consists of three hospitals and two different treatment options.

2.7. Real Data Simulation with Death Event or Readmission as a reward metric

We use both the death and readmission events as evaluation metrics for treatments, focusing on the patient encounter id. Specifically, for each patient encounter, we define the reward as 0 if there is a readmission or death event within 7 to 30 days from the last time the patient was administered a treatment during this encounter, and 1 otherwise. During an encounter, a patient is often administered multiple treatments. For each one of these treatments, we create a treatment-reward pair using the reward of the encounter. Finally, we use the treatment-reward pairs to update the estimated rewards of the treatments used in this encounter. Like Experiment 1, our dataset sufficiently covers only three of the seven treatments—Casirivimab-Imdevimab, Remdesivir, and Bamlanivimab-Etesevimab. Consequently, we created a simulation environment that includes three hospitals and three different treatment strategies.

For each one of the above two treatment outcomes, Table 1 summarizes the success rate of each medication and at each hospital. We observe that the same medication is consistently better across all three hospitals and the success rates are similar (although not the same). This is expected since the data were collected from the same controlled environment within the Duke Health System. Yet, the same medication cannot always be expected to perform best for all patient populations, especially if they are significantly different, e.g., at hospitals across different parts of the world.

2.8. Artificial Simulation Environment

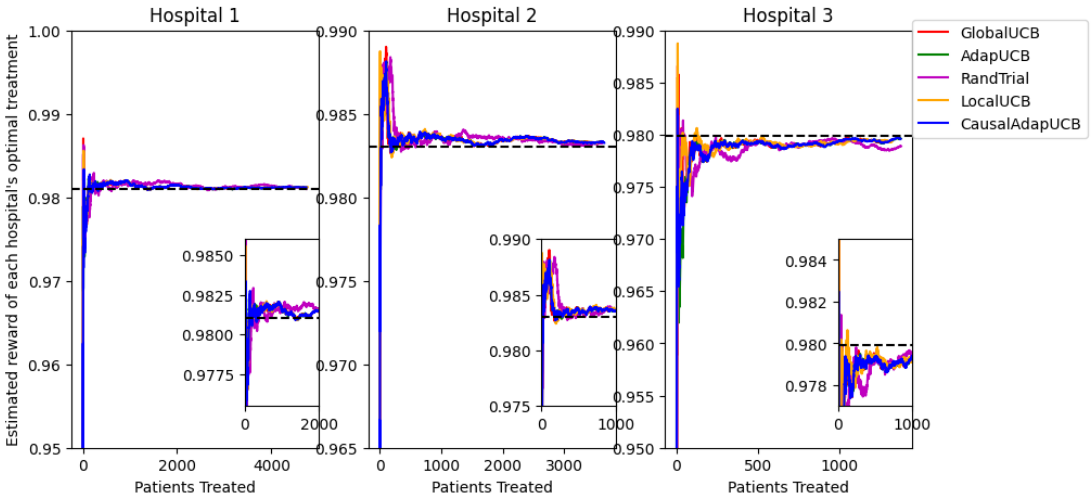
To introduce heterogeneity in treatment strategies across hospitals, we created a third artificial simulation environment, not based on real data from Duke Clinics. In this artificial environment, we simulate three hospitals, mirroring the patient populations from the first two real-data environments, with each hospital needing to choose from three different treatment options. We assume each hospital has a distinct optimal treatment: Treatment 1 is best for Hospital 2, Treatment 2 for Hospital 1, and Treatment 3 for Hospital 3. This artificial environment is shown at the bottom part of Table 1.

Evaluation Metric	Mathematical Definition	Analytical Description
$estimated\_reward(x)$	$\frac{\#\ of\ times\ treatment\ x\ had\ positive\ reward}{\#\ of\ times\ treatment\ x\ was\ administered}$	expected reward of the optimal treatment on the target patient population at each hospital defined as the success probability of the optimal treatment
$success\_rate$	$\frac{\#\ of\ patients\ administered\ optimal\ treatment}{total\ \#\ of\ patients}$	the ratio between the number of patients that were treated with the optimal treatment divided by the total number of patients.
$regret$	$T \cdot \mathbb{E}[y(x^*)] - \sum_{t=1}^{T} \mathbb{E}[y(x_t)]$	the difference between the reward that could have been achieved using the optimal treatment strategy and the reward that was achieved using the treatment implemented by the MAB algorithm.

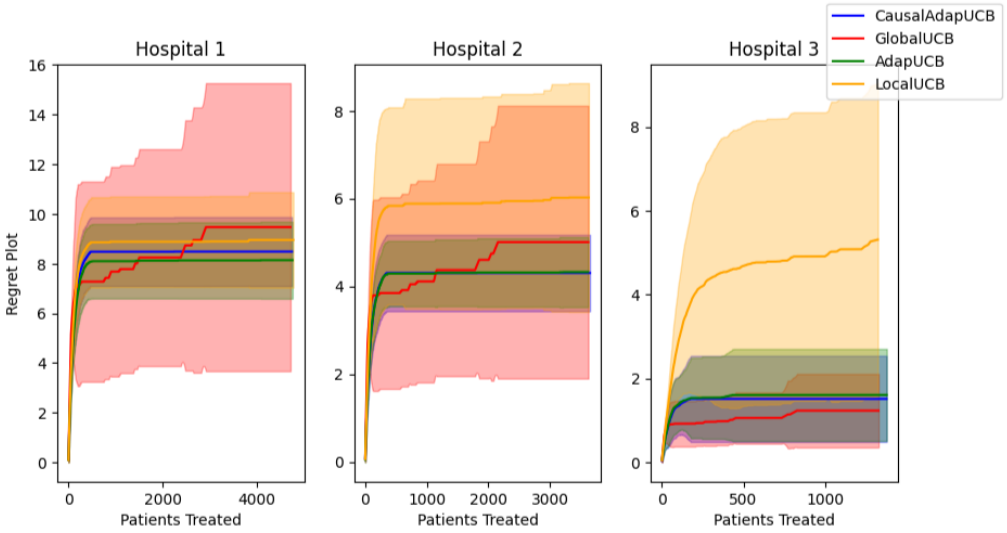
Table 2. Evaluation metrics used for the comparison of the different methods. (1) The estimated reward ($estimated\_reward$) compares the considered methods in terms of their ability to correctly identify the optimal treatment as well as in terms of their ability to estimate its expected reward. (2) The regret ($regret$) measures the ability of a method to minimize the number of errors, i.e., the number of times an incorrect treatment is administered to a patient. (3) Finally, the success rate ($success\_rate$) of each method. The better the method the higher the $success\_rate$ and the lower the regret of the trial.

2.9. Evaluation Metrics

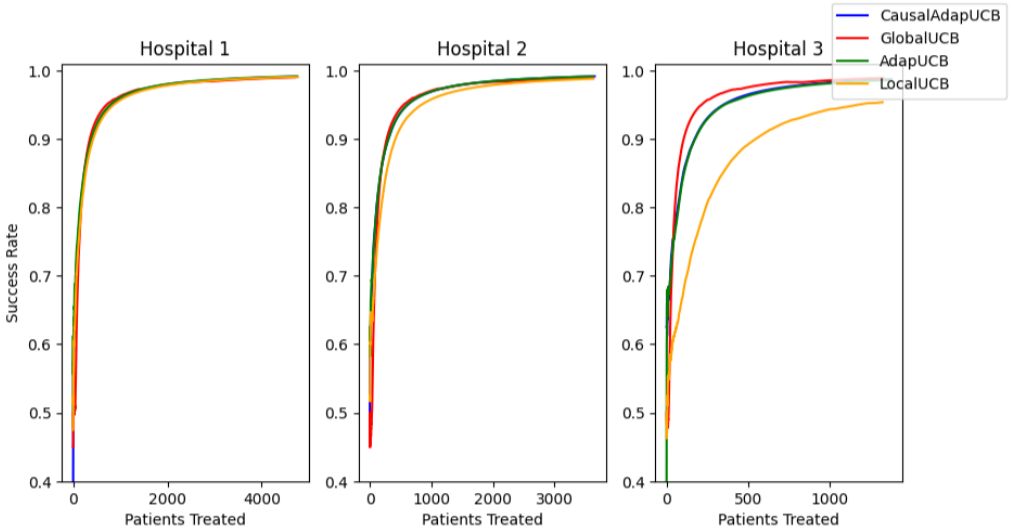
We compare our proposed method with the baselines adopted currently in treatment allocation tasks in terms of the metrics: (1) $estimated\_reward$, (2) $regret$ and (3) $success\_rate$ that are described in Table 2.



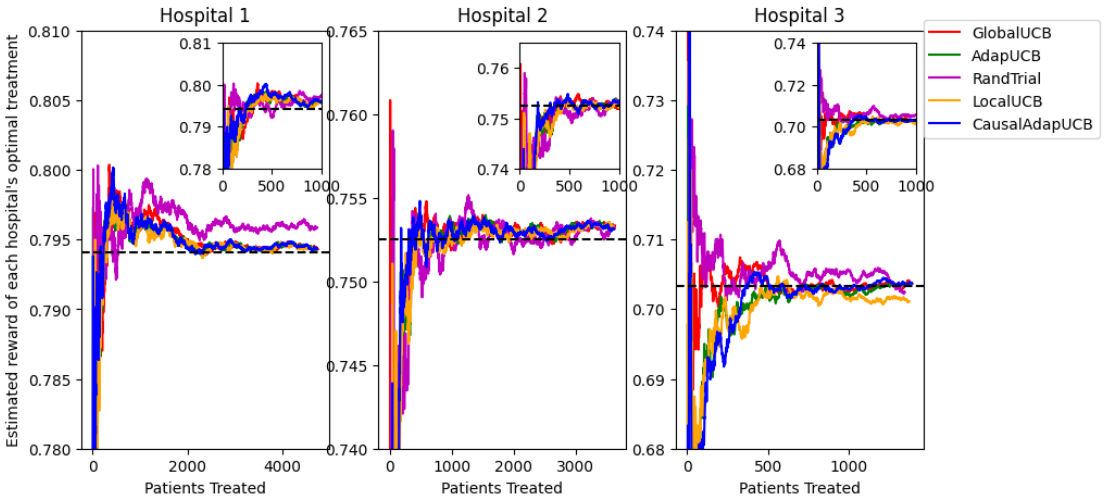
(a) Estimated reward of the optimal arm in each hospital for Experiment 1



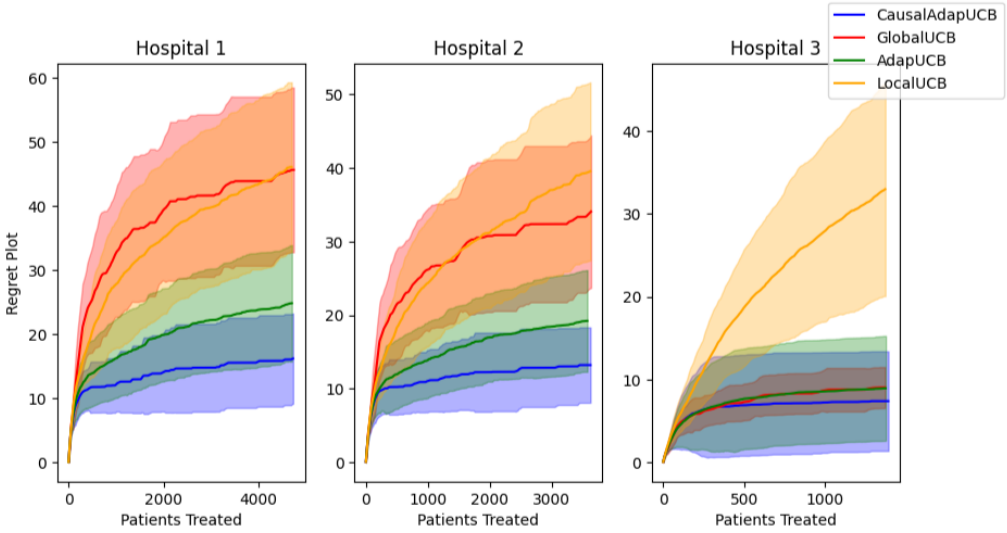
(b) Regret Plot for Experiment 1



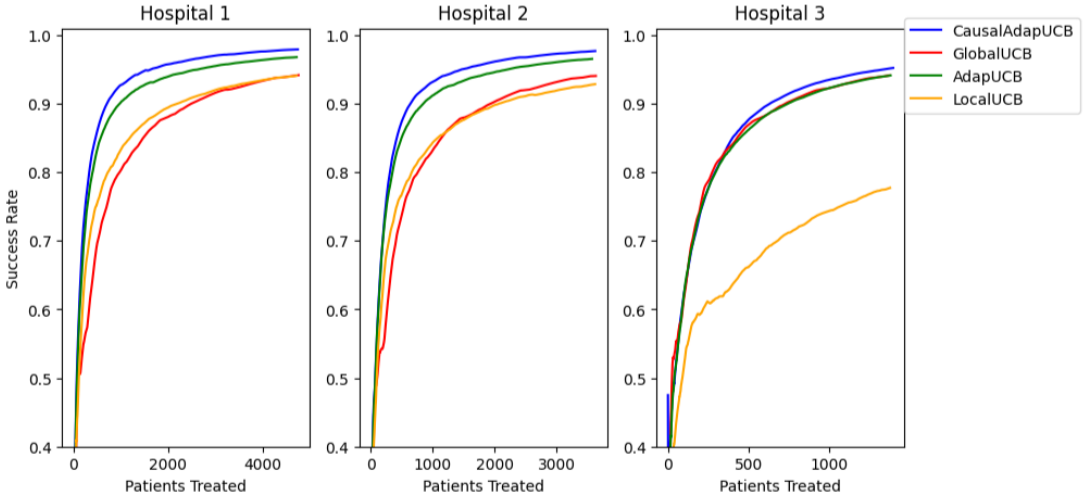
(c) Success rate for each hospital for Experiment 1



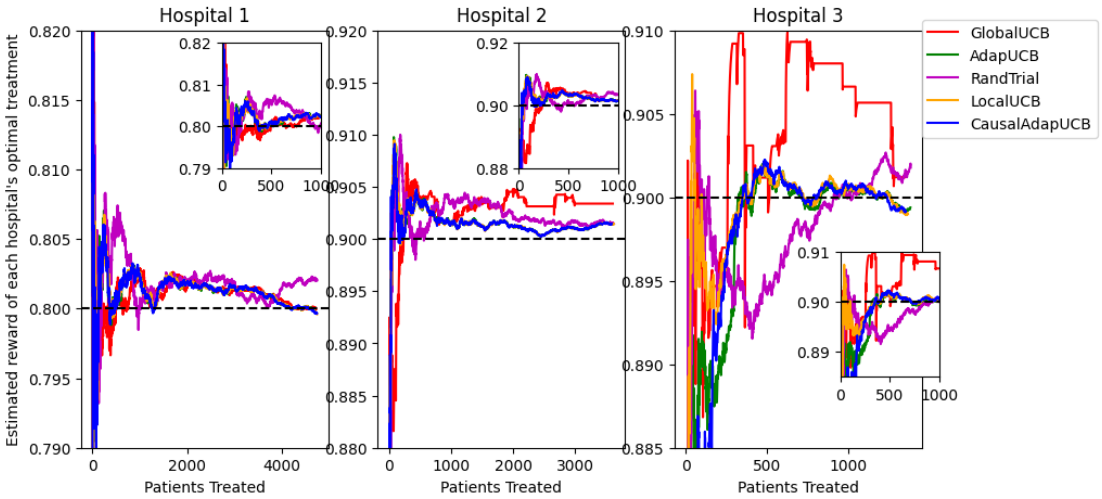
(d) Estimated reward of the optimal arm in each hospital for Experiment 2



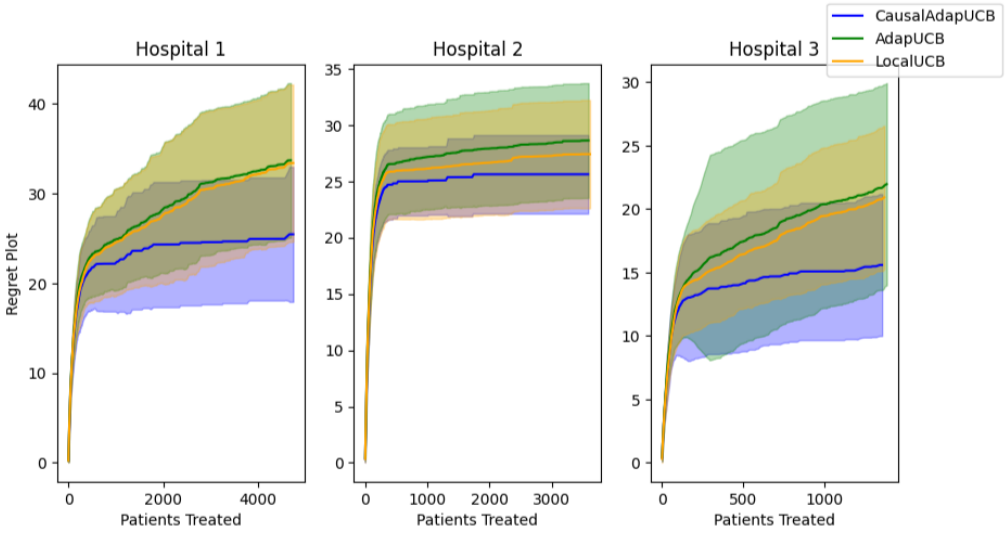
(e) Regret Plot for Experiment 2



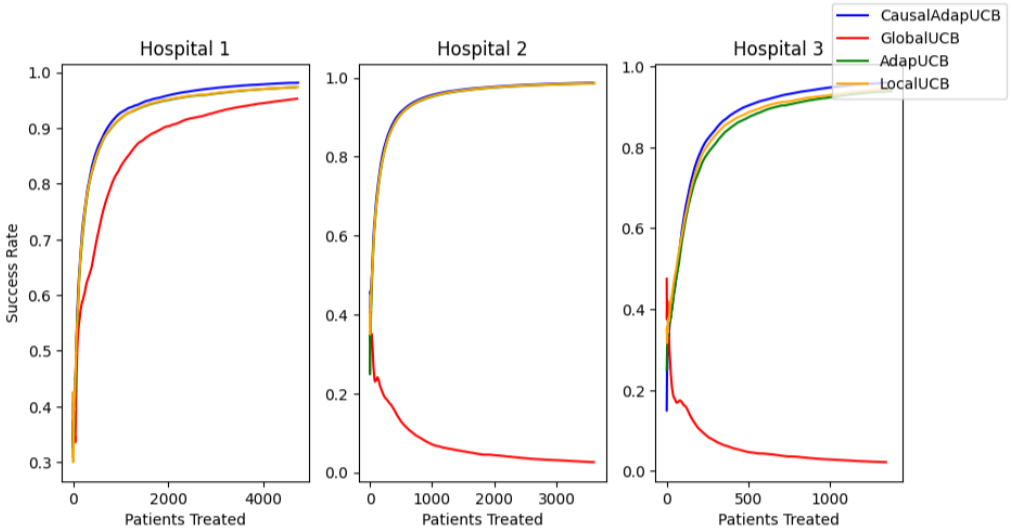
(f) Success rate for each hospital for Experiment 2



(g) Estimated reward of the optimal arm in each hospital for Experiment 3



(h) Regret Plot for Experiment 3



(i) Success rate for each hospital for Experiment 3

Figure 3. Plots of experimental results. Subfigures (a)-(c) show results for the Real Data Simulation with the Death Event as a reward metric, subfigures (d)-(f) show results for the Real Data Simulation with the Death Event or Readmission as a reward metric, and subfigures (g)-(i) show results for the Artificial Simulation Environment.

3. RESULTS

3.1. Numerical Analysis

In the **Real Data Simulation with the Death Event as a reward metric**, the methods evaluated are all very accurate in estimating the true expected reward associated with the optimal treatment in every hospital, as depicted in Figure 3a. The approaches **CausalAdapUCB**, **AdapUCB**, and **GlobalUCB** demonstrate significant efficiency, requiring 75% fewer patients than **LocalUCB** across all hospitals to successfully identify the optimal treatment and achieve convergence in the regret metric, a finding illustrated in Figure 3b. When it comes to the accumulated regret, all methods perform almost the same, with **GlobalUCB** having slightly higher mean regret but very high variance in two of the three hospitals. Furthermore, when examining the success rates, shown in Figure 3c, it is evident that the transfer learning strategies, **CausalAdapUCB**, **AdapUCB**, and **GlobalUCB**, slightly surpass **LocalUCB**, by administering the optimal treatment to an additional 1.1% of the 10,000 patients treated over all the hospitals. Note that, in this experiment, each treatment allocation task only considers two treatment options. The number of arms/treatments in a MAB problem affects the sampling complexity and thus the difficulty in finding the optimal solution. Fewer treatments mean more patients per treatment, allowing quicker estimation of rewards and fewer patients administered with sub-optimal treatments. This is why in this case, all MAB-based algorithms, including **LocalUCB**, perform similarly across hospitals.

In the **Real Data Simulation with the Death Event or Readmission** as a reward metric, all UCB-based methods estimate the expected optimal reward with near-perfect accuracy, as shown in Figure 3d, with **CausalAdapUCB** demonstrating a marginal superiority in performance over its counterparts across different hospitals. Notably, **CausalAdapUCB** exhibits the minimal regret among all hospitals, a result highlighted in Figure 3e. This method also shows 36% reduction in the number of patients required compared to AdapUCB and up to 62% fewer compared to LocalUCB and GlobalUCB to learn the optimal treatment strategy. Furthermore, **CausalAdapUCB** achieves the highest success rates, successfully administering the optimal treatment to 0.95% more patients than AdapUCB and 7.6% more than LocalUCB and GlobalUCB**,** as detailed in Figure 3f.

Note that in both Real Data experiments discussed above, the optimal treatment is the same for all three hospitals and the expected reward of this treatment is also similar across hospitals. Therefore, cooperation among the hospitals with any transfer-learning based method (i.e., **GlobalUCB**, **AdapUCB**, and **CausalAdapUCB**) naturally outperforms learning treatment strategies independently at each hospital (i.e., **LocalUCB**). Even in this case, **CausalAdapUCB** still manages to outperform the other methods with respect to both the number of patients needed to learn the optimal treatment and the number of patients treated with the optimal treatment.

Finally, in the **Artificial Simulation Environment, RandTrial** and **GlobalUCB** face challenges in accurately estimating the reward of the optimal treatment, particularly for hospitals that have fewer participants in the trials (Hospital 3), as depicted in Figure 3g. Moreover, **GlobalUCB** is notably unable to achieve convergence at Hospital 3. In contrast, the remaining three methods exhibit comparable performance levels, with **CausalAdapUCB** outperforming the others. In terms of regret, **CausalAdapUCB** stands out significantly; it requires 66% fewer patients to reach convergence, a substantial efficiency improvement highlighted in Figure 3h. Additionally, in

measuring success rates across hospitals, **CausalAdapUCB** administered the optimal treatment to 1.3% more patients compared to AdapUCB, 0.3% compared to **LocalUCB** and 48.6% more patients than **GlobalUCB**, as shown in Figure 3i.

The Artificial Simulation Environment was developed to simulate a scenario with highly heterogeneous patient populations across hospitals in terms of the optimal treatment strategy. In this case, **GlobalUCB** that computes a common UCB for all hospitals, introduces bias towards the hospital that serves the largest patient population (Hospital 1). This bias is evident in Hospitals 2 and 3 (hospitals with fewer patients) that settle for a sub-optimal treatment. Even Hospital 1, shows lower success rate since about half of the patients used to train the global model come from hospitals with different patient demographics. Consequently, **GlobalUCB** needs more patients to identify the optimal treatment for Hospital 1, as seen in Figure 3f and 3i. On the contrary, **CausalAdapUCB** and **AdapUCB** can handle heterogeneity across hospitals and, therefore, outperform the other methods by returning fairer treatment strategies adapted to individual hospital populations with **CausalAdapUCB** demonstrating the best performance.

Note that, in all the experiments above, Hospital 3, that has the fewest participants, benefits the most from collaborating with other hospitals, since all collaborative methods—**GlobalUCB**, **AdapUCB**, and **CausalAdapUCB**—outperform **LocalUCB**. Only in the case of highly heterogeneous patient populations (the Artificial Simulation), does **CausalAdapUCB** show significant advantage over the rest methods for Hospital 3. On the other hand, for hospitals with smaller populations (Hospitals 2 and 3) **CausalAdapUCB** outperforms the other transfer learning based methods.

Ablation Comparison between CausalAdapUCB and AdapUCB. To further assess the specific contribution of the causal inference component, we compare CausalAdapUCB directly with

AdapUCB across all experiments. While both methods leverage cross-hospital information adaptively, only CausalAdapUCB integrates PUBounds to constrain transfer from heterogeneous hospitals. This causal regularization yields consistently higher success rates and faster convergence, particularly in settings with pronounced heterogeneity, as in the Artificial Simulation Environment. The observed gains—ranging from 0.95% to 1.3% in success rate, for the homogeneous and heterogeneous cases accordingly, and up to 36% reduction in patient data requirements compared to AdapUCB—quantify the added value of causal knowledge transfer in guiding treatment selection under non-uniform hospital populations.

3.2. Discussion

A key challenge often encountered in treatment recommendation tasks is the lack of sufficient numbers of participants needed to learn an optimal treatment strategy. In this work, we address this challenge by designing a federated learning method that leverages collaboration across multiple hospitals conducting individual learning procedures to compensate for limited numbers of participants. Specifically, collaboration between hospitals increases the information they can obtain on the available treatments beyond what is possible with local data, leading to fewer participants needed in each hospital. The proposed federation is coordinated by a lead hospital that only has access to aggregate local hospital data without any patient sensitive information, thus protecting patient privacy [19]. The effect of collaboration can be observed in the two real-data simulations, where the number of participants needed by collaborative methods (**GlobalUCB, AdapUCB, CausalAdapUCB**) is lower compared to that needed by non-collaborative methods (**LocalUCB, RandTrial**). Therefore, federated learning allows to learn the optimal treatment at each hospital faster. In the case of clinical trials for example, this could also lead to significant cost savings related to the recruitment of trial participants.

Along with sufficient subject participation, equally important is fairness of the learned optimal treatment strategies across all patients, meaning that these treatment strategies are free of any biases caused by any subgroup of patients' populations. Fairness becomes even more important in collaborative procedures where information can be shared across hospitals and, as a result, biases can be introduced in the learned treatments. In our experiments, we observe that **GlobalUCB** that computes one common UCB for all hospitals suffers in terms of fairness, since it learns treatment strategies biased towards hospitals that serve larger patient populations. This limitation of **GlobalUCB** is addressed by **AdapUCB** that clusters similar hospitals allowing only those to collaborate. In this way, **AdapUCB** does not introduce bias in the learned optimal treatment strategies (like **LocalUCB**), however, since it uses data from fewer hospitals, it requires more participants to learn the optimal treatment strategies. To ensure fairness of the learned treatment strategies, like **AdapUCB**, our **CausalAdapUCB** method clusters similar hospitals that can safely share information with each other, but it also employs causal inference to transfer knowledge on treatment effects (in the form of PUBounds) across heterogeneous hospitals without introducing bias. Specifically, **CausalAdapUCB** transfers more conservative PUBounds to hospitals with fewer patient participants, which protects these hospitals from learning treatment strategies that are biased towards hospitals with larger patient populations. If the hospitals are similar, the use of PUBounds in **CausalAdapUCB** offers no measurable advantage over **GlobalUCB** and **AdapUCB**. However, if the hospitals are heterogeneous, **CausalAdapUCB** allows small hospitals to learn unbiased treatment strategies even though they have fewer patient participants. On the other hand, PUBounds transferred by **CausalAdapUCB** to larger hospitals with more patient participants are less conservative, allowing these hospitals to refine their exploration of treatment options and learn an unbiased optimal treatment strategy faster, using fewer patient participants. As a result, **CausalAdapUCB** outperforms the other methods in terms of both learning fairer treatment strategies and needing fewer participants to learn the optimal ones.

Finally, it is critical that during the learning process of a treatment recommendation task we maximize patients' safety. Randomized treatment allocation increases the risk to participants, since more subjects are administered with sub-optimal treatments. In contrast, **CausalAdapUCB** is designed to adapt its treatment recommendations as more data are collected. In our experiments, we show that this adaptive nature of the algorithm reduces the number of patients administered with ineffective treatments, regardless of how similar or heterogeneous the hospitals are, or how many treatment options are explored during the process. Therefore, **CausalAdapUCB** improves the safety of these learning procedures compared to randomized approaches or other state-of-the-art MAB methods.

From a translational perspective, the proposed framework can naturally serve as a coordination mechanism for multi-site adaptive clinical trials or learning health system initiatives, enabling hospitals to collaboratively update treatment allocation strategies using real-world evidence while retaining patient-level data locally.

Limitations.

Deployment considerations. Practical deployment of the proposed framework would require addressing several operational constraints. Communication overhead increases with the number of participating hospitals and the frequency of communication rounds, and in real-world settings this can be managed by tuning the communication interval or triggering updates only when uncertainty is sufficiently high. In addition, although raw patient-level data are never exchanged, the transfer of aggregated statistics still requires appropriate security safeguards, including authenticated and encrypted communication, access control and auditing at the coordinating institution to limit information leakage.

Our proposed method is an online algorithm suitable for treatment recommendation settings, where patients arrive sequentially and decisions must be made dynamically as new outcomes are

observed. Because the algorithm updates its estimates iteratively after each treatment–outcome pair, it requires an interactive setting that mimics real-time decision-making. As such, it cannot be directly applied to static retrospective datasets without significant modifications, since those datasets do not capture the sequential feedback loop that drives the algorithm. Consequently, we evaluated the algorithm using a simulation environment. This simulator, constructed using real Covid-19 data, closely replicates real-world scenarios. As a result, the findings from our simulations provide valuable information on how our algorithm would have performed if it had been deployed during the treatment recommendation task at the time. The development of federated treatment allocation strategies that can directly learn from historical data and do not require real-time interaction with patients is of great interest and currently part of our future work. Moreover, sharing patients' sensitive information across hospitals is a violation of their privacy, and as a result designing a personalized treatment allocation strategy for every patient in the federated setting is a demanding task. The safest first step towards this direction is to design personalized strategies for each hospital. We therefore work under the assumption that patients of the same hospital have similar reactions to every treatment. In this setting, we show that at the population level our treatment allocation policy performs optimally, since it manages to decrease the number of patients treated with sub-optimal treatments and affect individual patient outcomes. Extending our work to patient-level frameworks that design personalized treatment allocation strategies for each individual patient while protecting their privacy constraints is part of our future work.

4. CONCLUSION

In this paper, we propose an adaptive federated learning strategy for treatment treatment recommendation tasks. We formulate the problem as a federated Multi-Armed Bandit problem where hospitals that potentially serve different patient populations cooperate to learn their local

optimal treatments. We test our proposed approach for treatment allocation tasks on a simulated clinical trial environment created using real Covid-19 data from the Duke University Health System and show that it outperforms other state of the art methods, by learning the optimal treatment for each hospital faster and with fewer number of patient-participants, while also satisfying privacy, fairness, and safety requirements.

CONTRIBUTORS

XK led all aspects of the Review: conceptualization, data curation (title and abstract screening, full-text screening, and data extraction), and writing (original draft, review, and editing); MZ was involved in conceptualization, supervision, data curation (title and abstract screening, full-text screening), and writing (original draft, review, and editing); CH was involved in conceptualization, data curation (title and abstract screening, full-text screening), and writing (original draft, review, and editing); YS was involved in conceptualization, and writing (original draft, review, and editing); NEM,GS,MP,AB were involved in conceptualization. All authors have had full access to the data and accept responsibility to submit for publication.

DISCLOSURE OF THE USE OF CHATGPT

ChatGPT was employed exclusively to improve the readability and language of this manuscript, with strict human oversight. It did not participate in any scientific tasks or analyses, nor was it granted authorship. The authors maintain full responsibility for the manuscript's integrity and accuracy, having carefully examined the AI-enhanced output to adhere to ethical scientific reporting standards.

ACKNOWLEDGEMENTS

This work is supported in part by The Duke Endowment (TDE) under grant #7262-SP and by the Onassis Foundation under award #ZT033-1/2023-2024.

REFERENCES

[1] Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 25, 44–56.

[2] Esteva, A., Robicquet, A., Ramsundar, B. *et al.* (2019). A guide to deep learning in healthcare. *Nat Med* **25**, 24–29

[3] Villar SS, Bowden J, Wason J. (2015).Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. Statistical Science 30(2):199-215.

[4] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347-1358.

[5] Shickel B, Tighe PJ, Bihorac A, Rashidi P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE *Journal of Biomedical and Health Informatics*,Sep;22(5):1589-1604.

[6] Kuleshov, Volodymyr & Precup, Doina. (2014). Algorithms for multi-armed bandit problems. Journal of Machine Learning Research. 1.

[7] Bouneffouf, D., Rish, I., and Aggarwal, C. (2020). Survey on Applications of Multi-Armed and Contextual Bandits. In 2020 IEEE Congress on Evolutionary Computation (CEC) (IEEE), pp. 1–8.

[8] Kojima M. Application of multi-armed bandits to dose-finding clinical designs. Artif Intell Med. 2023 Dec;146:102713. doi: 10.1016/j.artmed.2023.102713. Epub 2023 Nov 13. PMID: 38042600.

[9] Aziz, M., Kaufmann, E., and Riviere, M.-K. (2021). On Multi-Armed Bandit Designs for Dose-Finding Trials. J. Mach. Learn. Res. *22*, 1–38.

[10] Varatharajah Y, Berry B. A Contextual-Bandit-Based Approach for Informed Decision-Making in Clinical Trials. Life (Basel). 2022 Aug 21;12(8):1277. doi: 10.3390/life12081277. PMID: 36013456; PMCID: PMC9410371.

[11] Shao, Y. (2024). Personalized clinical trial based on multi-armed bandit algorithms with covariates. In International Conference on Algorithms, Software Engineering, and Network Security (ACM). https://doi.org/10.1145/3677182.3677185.

[12] Alban, Andres and Chick, Stephen E. and Zoumpoulis, Spyros, Learning Personalized Treatment Strategies with Predictive and Prognostic Covariates in Adaptive Clinical Trials (January 24, 2025). INSEAD Working Paper No. 2025/13/TOM/DSC, Available at SSRN: https://ssrn.com/abstract=4160045 or http://dx.doi.org/10.2139/ssrn.4160045.

[13] Gottesman, O., Johansson, F., Komorowski, M. *et al.* Guidelines for reinforcement learning in healthcare. *Nat Med* **25**, 16–18 (2019). https://doi.org/10.1038/s41591-018-0310-5.

[14] Ganz, P.A. (1990). Clinical trials. Concerns of the patient and the public. Cancer *65*, 2394–2399.

[15] Chu, S.H., Kim, E.J., Jeong, S.H., and Park, G.L. (2015). Factors associated with willingness to participate in clinical trials: a nationwide survey study. BMC Public Health *15*, 10.

[16] Shi, C., Shen, C., and Yang, J. (13--15 Apr 2021). Federated Multi-armed Bandits with Personalization. In Proceedings of The 24th International Conference on Artificial Intelligence

and Statistics Proceedings of Machine Learning Research., A. Banerjee and K. Fukumizu, eds. (PMLR), pp. 2917–2925.

[17] Ghosh, A., Sankararaman, A., and Ramchandran, K. (2021). Adaptive Clustering and Personalization in Multi-Agent Stochastic Linear Bandits. arXiv [stat.ML].

[18] Zhang, J., and Bareinboim, E. (2017). Transfer learning in multi-armed bandits: A causal approach. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (International Joint Conferences on Artificial Intelligence Organization). 10.24963/ijcai.2017/186.

[19] Tucker, K., Branson, J., Dilleen, M., Hollis, S., Loughlin, P., Nixon, M.J., and Williams, Z. (2016). Protecting patient privacy when sharing patient-level data from clinical trials. BMC Med. Res. Methodol. *16 Suppl 1*, 77.

[20] Wawira Gichoya, J., McCoy, L.G., Celi, L.A., and Ghassemi, M. (2021). Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Health Care Inform *28*. 10.1136/bmjhci-2020-100289.

[21] P Kim, *et al.* Therapy for Early COVID-19 - A Critical Need. *Journal of the America Medical Association* DOI: DOI:10.1001/jama.2020.22813 (2020).

[22] Chen, Z., Zhang, H., Guo, Y., George, T.J., Prosperi, M., Hogan, W.R., He, Z., Shenkman, E.A., Wang, F., and Bian, J. (2021). Exploring the feasibility of using real-world data from a large clinical data research network to simulate clinical trials of Alzheimer's disease. NPJ Digit Med *4*, 84.

[23] Holford, N.H., Kimko, H.C., Monteleone, J.P., and Peck, C.C. (2000). Simulation of clinical trials. Annu. Rev. Pharmacol. Toxicol. *40*, 209–234.

[24] Zelen, M. (1974). The randomization and stratification of patients to clinical trials. J. Chronic Dis. *27*, 365–375.

[25] Broglio, K. (2018). Randomization in Clinical Trials: Permuted Blocks and Stratification. JAMA *319*, 2223–2224.

[26] Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem. Machine learning. 2002;47:235-56.

[27] McCreary, E. K., Bariola, J. R., Minnier, T. E., Wadas, R. J., Shovel, J. A., Albin, D., ... & Huang, D. T. (2022). The comparative effectiveness of COVID-19 monoclonal antibodies: a learning health system randomized clinical trial. *Contemporary Clinical Trials*, *119*, 106822.