Electronic Medical Records Assisted Digital Clinical Trial Design

Xinrui Ruan
Fudan University

Jingshen Wang
UC Berkeley

Yingfei Wang
University of Washington

Waverly Wei
UC Berkeley

Abstract

Randomized controlled trials (RCTs) are gold standards for assessing intervention efficacy. Yet, generalizing evidence from classical RCTs can be challenging and sometimes problematic due to their limited external validity under stringent eligibility criteria and inadequate statistical power resulting from limited sample sizes under budgetary constraints. "Digital clinical trial," which utilizes digital technology and electronic medical records (EMRs) to expand eligibility criteria and enhance data collection efficiency, offers a promising concept for solving the above-mentioned conundrums encountered in classical RCTs. In this paper, we propose two novel digital clinical trial design strategies assisted by EMRs collected from diverse patient populations. On the one hand, leveraging digital technologies, our design strategies adaptively modify both the eligibility criteria and treatment assignment mechanism to enhance data collection efficiency. As a result, evidence gathered from our design can possess greater statistical power. On the other hand, since EMRs capture diverse patient populations and provide large sample sizes, our design not only broadens the trial's eligibility criteria but also enhances its statistical power, enabling us to collect more generalizable evidence with boosted statistical power for evaluating intervention efficacy than classical RCTs. We demonstrate the validity and merit of the proposed designs with detailed theoretical investigation, simulation studies, and a synthetic case study.

1 INTRODUCTION

1.1 Motivation and Contribution

In clinical studies, randomized controlled trials (RCTs) are considered gold standards for assessing the safety of new drugs and the efficacy of interventions. Leveraging the randomized treatment assignment mechanism, RCTs can provide reliable causal conclusions. However, traditional clinical trials present two practical challenges that can hamper the generalization of their results. *On the one hand*, although classical RCTs aim to determine the efficacy of a treatment in a general population, the causal effect estimator derived from such trials can significantly differ from this objective. This deviation can arise from restrictive enrollment criteria and challenges in reaching diverse populations. For example, after studying 32 major HIV RCTs, Gandhi *et al.* (2005) has suggested that the eligibility criteria of HIV trials exclude a large proportion of HIV-infected women from being enrolled in the trial. As the trial population is not representative of the general population, the obtained evidence may not be generalizable.

On the other hand, due to high implementation costs, traditional RCTs can only recruit a relatively small number of participants, and treatments are often administered with fixed randomization probability throughout the trial. Not only does the limited sample size restrict the statistical power of confirming the treatment efficacy, but the fixed randomization raises concerns regarding inefficient data usage. Imagine a scenario where, during a trial, outcomes in the placebo arm show little variability but vary widely in the treatment arm. In this case, it is reasonable to assign more future patients to the treatment arm because the increased variability in outcomes suggests a need for further investigation.

With the rise of digital technology and increased availability of electronic medical records (EMRs), "digital clinical trial" is an emerging new concept to solve the above-mentioned challenges. In a joint workshop hosted by the National Institutes of Health and the National Science Foundation in 2019, "Digital clinical

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

trial" refers to utilizing digital technology to broaden patient enrollment and optimize data collection. Intuitively, online platforms can send push notifications to encourage patient enrollment or digitally revise the treatment allocation scheme while monitoring patients' response to the treatment.

In this paper, we propose a novel digital clinical trial design leveraging external EMRs to deliver robust and generalizable causal evidence. Given the rich and diverse information contained in EMR data, we note that our approach is just one of many methods for integrating EMR data, particularly when data on historical subpopulations and outcomes are accessible. We summarize our contributions as follows:

From a methodological perspective, (i) we provide the mathematical formulation of digital clinical trial design assisted by the external EMR data (Section 2), such that the evidence collected from our digital clinical trials can be easily generalized to the broader population. (ii) Our design realizes the promise of digital clinical trials by broadening the enrollment criteria and improving the efficiency of the data collection mechanism (Section 3). These can be achieved by adaptively revising the randomization probability and enrollment criteria in digital platforms. (iii) We propose two flexible design strategies operating under different practical considerations (Section 4.1 and 4.2), allowing practitioners to choose a design based on how closely the treatment effect in the general population aligns with that of the trial population.

From a theoretical perspective, we first demonstrate the merit of designing EMR-assisted digital clinical trials by showcasing the improved efficiency under our design compared with the design without using EMR data (Remark 2). Second, we establish the statistical validity of our design by showing the asymptotic normality of our proposed estimator and the convergence of our derived design strategies (Theorem 1). As our design operates in an adaptive manner and the collected data depend on historical data, we address the challenge of establishing theoretical results of dependent and non-identically distributed data (Theorem 1). Third, we demonstrate that our design can deliver valid statistical inference in a realistic setting where the treatment effect in the general population may not align with that of the trial population (Theorem 2).

From an application perspective, our design demonstrates robust finite sample performance compared to other benchmark designs. In the simulation studies (Section 6), our design achieves higher estimation efficiency than the design without incorporating any EMR data. Furthermore, we provide guidance regarding the performance of our design under different levels of misalignment of treatment effects between the general population and the trial population. In the synthetic case study, we demonstrate that our design can help uncover significant effects by re-designing an HIV cash transfer trial assisted by an external EMR database.

1.2 Related Literature

Our paper is motivated by recent advances in digital clinical trials (Inan *et al.*, 2020; Steinhubl *et al.*, 2019). Digital tools offer convenience, reducing the burden on participants and coordinators, thereby enabling broader enrollment and greater generalizability of the trial (Zhan *et al.*, 2018). Additionally, digital platforms for massive data collection can enhance experimental design and estimation efficiency. However, existing studies do not provide a comprehensive framework for efficient digital clinical trial design (Perez *et al.*, 2019; Garcia and et al., 2022). Furthermore, our proposed digital clinical design leverages external EMR data, which are known to encompass extensive patient information, such as demographics, medical history, and lab tests (Sun *et al.*, 2018). EMR data have gained increasing attention because they can facilitate evidence-based medicine (Frankovich *et al.*, 2011; Hoffman and Williams, 2011) and improve the quality of primary care (Bates *et al.*, 2003; Wang *et al.*, 2003; Ayaad *et al.*, 2019).

Our proposed digital clinical trial design strategy has a connection with adaptive experiment literature. It naturally connects with response-adaptive randomization (RAR) design, which refers to the design that adaptively revises treatment assignment probabilities based on the collected outcomes accrued during the experiment (Eggenberger and Pólya, 1923; Zelen, 1969; Rosenberger and Lachin, 1993; Williamson *et al.*, 2017; Hu and Rosenberger, 2003). Furthermore, it connects with adaptive enrichment design (Simon and Simon, 2013; Thall, 2021; Lai *et al.*, 2019; Stallard, 2023), which uses interim analysis to revise patient enrollment criteria such that the benefitted patient groups can be identified at the end of the trial.

Our estimation of causal effects in digital clinical trials is related to the literature on statistical inference with adaptively collected datasets. For example, Zhang *et al.* (2021) proposed inference strategies based on M-estimation for adaptively collected data. Dimakopoulou *et al.* (2021) discussed online bandits with adaptive inference. Shi *et al.* (2023) introduced a unified statistical inference framework for multi-armed bandits to estimate causal parameters with delayed outcomes. Shi *et al.* (2022) proposed statistical inference for confounded Markov decision processes. Hadad et al. (2021) introduced a class of test statistics for policy evaluation in adaptive experiments.

Our paper also relates to the literature on data integration, which aims to estimate the causal effect of the target population by combining RCTs and observational studies. In cases where observational data only contains covariate information, various methods can be applied, such as stratification (Buchanan et al., 2018), the plug-in g-formula (Dahabreh et al., 2020), inverse probability of sampling weighting (Colnet et al., 2022), and calibration weighting (Lee et al., 2023). When observational data contain treatment and outcome information, Li et al. (2023) and Yang et al. (2020) use the semiparametric efficiency theory to derive a semiparametrically efficient integrative estimator.

2 PROBLEM SETUP

In this section, we will mathematically formulate the challenges and definition of designing digital clinical trials assisted by external EMRs. We will then establish the promise of digital clinical trials assisted by EMR data.

We start by providing the observed EMR data structure under consideration. Suppose EMR data are collected from N patients in a general population. Given a medication or a treatment of interest, we hope to assess its efficacy by investigating its average treatment effect (ATE) measured in this general population. To rigorously quantify this causal effect, for $i = 1, \dots, N$, let $Y_i \in \mathbb{R}$ denote patient i's observed outcome, $D_i \in \{0, 1\}$ the treatment assignment status with $D_i = 0$ being the control and $D_i = 1$ being the treatment, and $X_i \in \mathbb{R}^p$ the patients' covariate information (such as biomarkers and demographics). We note that the EMRs considered in our paper only contain controls collected from the general population. That is, we restrict $D_i = 0$, for $i = 1, \dots, N$, for the EMRs to be integrated into our designs. Following the Neyman-Rubin causal model (Splawa-Neyman et al., 1990; Rubin, 1974), we define $Y_i(d)$ as the potential outcome we would have observed under treatment d, $d \in \{0, 1\}$. To set a clear difference between the EMR data and the trial data to be introduced, we introduce a variable $Z_i \in \{0, 1\}$, where $Z_i = 0$ denotes EMR data source, $Z_i = 1$ denotes trial data source. We are now ready to define our target parameter of interest, which is the ATE in the general population,

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0) | Z_i = 0]. \tag{1}$$

Directly estimating the ATE in the general population τ from EMRs can be challenging for two reasons. On the one hand, EMR data typically only contain the control arm information, making the estimation of causal effect infeasible. On the other hand, even when EMR data contain both treatment and control arms information, the estimation of causal effect can be confounded by unmeasured confounding variables, potentially delivering unreliable causal conclusions.

Clinical trials (or RCTs) seemingly provide a natural remedy to address the above-mentioned challenges that emerged in EMR data because the treatments are randomly assigned to patients and are thus independent of all unmeasured confounders. Nevertheless, as pointed out in Gandhi et al. (2005) and Fehrenbacher et al. (2009), RCTs recruit participants in a restricted population, suggesting the covariate distribution in RCTs may substantially differ from the general population's. Therefore, the sample collected from RCTs might be adopted to characterize a parameter different from τ. Without loss of generality, we denote the target parameter in RCTs as

$$\tau^{\texttt{trial}} = \mathbb{E}[Y_i(1) - Y_i(0) | Z_i = 1] \neq \tau,$$

where $Z_i = 1$ denotes the subject i being collected from RCTs. Furthermore, due to the high costs of clinical trials, not only the sample size of clinical trials is typically limited, but the data collection mechanism with fixed randomization is prevalent, suggesting a potentially inefficient allocation of trial efforts. Given these two concerns, it remains uncertain how effectively traditional RCTs can inform clinical decisions regarding treatment efficacy in the general population.

With the help of digital technology, the concept of "digital clinical trials" has been proposed to revolutionize traditional RCTs (Inan et al., 2020) with the promise of broadening enrollment and optimizing data collection. This is because, rather than conducting fixed randomization and enrollment criteria throughout the trial, digital technologies enable practitioners to use the accrued evidence to flexibly revise the randomization probability as well as the enrollment criteria in multiple stages, potentially without limit. In particular, the randomization probability can be sequentially modified to allocate more (less) treatments to strata exhibiting greater (lower) variability in the outcome variable. This allows practitioners to improve the data collection efficiency, and thus, the statistical power of identifying the desired treatment effect can be potentially elevated.

To provide a comprehensive understanding of how digital clinical trials enhance traditional RCTs, Section 3 will be dedicated to a detailed exploration. In the remainder of this section, we will focus on systematically introducing the data collection mechanisms utilized in digital clinical trials.

Suppose in a digital clinical trial, subjects are enrolled in the trial across $t = 1, \dots, T$ stages, and the total number of subjects enrolled in the trial is $n = \sum_{t=1}^{T} n_t$, where n_t is the number of subjects enrolled in Stage t. For the subjects i at Stage t, we denote the treatment assignment status as $D_{it} \in \{0, 1\}$, and the covariates $X_{it} \in \mathbb{R}^p$. Clinical trials typically involve pre-specified patient subpopulations defined by the covariates; that is, we assume the sample space of covariates can be divided into K non-overlapping regions $\mathcal{X} = \bigcup_{k=1}^{K} S_k$. We denote the total number of individuals enrolled in subpopulation k in the EMR data as N_k. The total number of subjects enrolled in the digital clinical trial in subpopulation k is $n_k = \sum_{t=1}^{T} n_{tk}$. Lastly, we denote the observed outcome as $Y_{it} \in \mathbb{R}$ and assume the outcomes are observed at the end of each stage without delay. Similarly, we can define the potential outcomes as $Y_{it}(d)$, $d \in \{0, 1\}$. Together, we consider the following digital clinical trial data structure:

(Digital clinical trial data). *Across T stages, we observe the dataset: $\{(Y_{it}, D_{it}, X_{it}, Z_{it})_{i=1}^{n_t}\}_{t=1}^{T}$.*

Because digital clinical trials can enroll patients sequentially across multiple stages and revise design strategies at the end of each stage, we have the unique opportunity to revolutionize traditional clinical trials by flexibly revising two features that are typically fixed in traditional trials: (1) enrollment eligibility criteria and (2) treatment assignment probability. Let the enrollment proportion and the treatment assignment probability in each subpopulation at trial Stage t be p_{tk} and e_{tk}, respectively. For $t = 1, \dots, T, k = 1, \dots, K$,

$$\begin{aligned} p_{tk} &: = \mathbb{P}\left(X_{it} \in \mathcal{S}_k \mid Z_{it} = 1\right), \\ e_{tk} &: = \mathbb{P}\left(D_{it} = 1 \mid X_{it} \in \mathcal{S}_k\right). \end{aligned} \tag{2}$$

In our digital clinical trial design, we will adaptively revise p_{tk} and e_{tk}, assisted by external EMR data. By revising these two quantities, we hope to fulfill the promise of digital clinical trials : (1) broaden eligibility criteria, and (2) allocate treatments to efficiently estimate τ in Eq (1).

3 THE PROMISE OF DIGITAL CLINICAL TRIALS

In this section, we shall introduce the objective of our digital clinical trial to showcase the promise of digital clinical trials. The promise of digital clinical trials includes (1) broadening eligibility criteria and (2) improving the estimation efficiency of the ATE. We show that our design fulfills the first promise in Remark 1 under an oracle setting. By "oracle", we refer to the setting where we have perfect knowledge regarding the data-generating distribution, such that the true parameters are known. We demonstrate that our design fulfills the second promise by showing the estimator under our design achieves higher semiparametric efficiency than the design without incorporating EMR data (Remark 2, Section 5). As our design leverages both the EMR data and the clinical trial data, in the oracle setting, we make the "perfect transportability" assumption:

Assumption 1 $\mathbb{E}(Y(d)|X, Z = 0) = \mathbb{E}(Y(d)|X, Z = 1), \ d \in \{0, 1\}$.

Assumption 1 assumes that the conditional expectation of potential outcomes under each treatment arm is the same in the EMR data and the trial data. In Section 4.2, we shall provide a novel design strategy in which perfect transportability is not required.

Our goal is to find the optimal digital clinical trial design assisted by EMR data, such that we can efficiently estimate τ in Eq (1). To estimate τ, we adopt a stratified version of the estimator in Li *et al.* (2023):

$$\begin{aligned} \hat{\tau} &= \sum_{k=1}^{K} \frac{N_k}{N} \hat{\tau}_k, \qquad (3) \\ \hat{\tau}_k &= \hat{\mathbb{E}}[\frac{1 - Z}{1 - \hat{\pi}_k}(\hat{m}_1(X) - \hat{m}_0(X) + \frac{DZ}{\hat{\pi}_k \hat{e}_k}(Y - \hat{m}_1(X)) \\ &\quad - \frac{Z(1 - D) + (1 - Z)\hat{r}(X)}{\hat{\pi}_k (1 - \hat{e}_k) + (1 - \hat{\pi}_k)\hat{r}(X)}(Y - \hat{m}_0(X))], \end{aligned}$$

where all the quantities with subscript k indexes subpopulation k. Here, $\hat{\pi}_k$ is the estimator for $\pi_k = \mathbb{P}(Z = 1|X \in S_k)$, the proportion of subjects from the trial in subpopulation k. $\hat{m}_d(X)$ is the estimator for the conditional expectation $m_d(X) = \mathbb{E}[Y|X, D = d, Z = 1]$. $\hat{e}_k$ is the estimator for $e_k = \mathbb{P}(D = 1|X \in \mathcal{S}_k)$, the treatment assignment probability in subpopulation k. $\hat{r}(X)$ is the estimator for the variance ratio of the potential outcome under the control arm in the two data sources: $r(X) = \frac{\mathbb{V}(Y(0)|X,Z=1)}{\mathbb{V}(Y(0)|X,Z=0)}$. Let the asymptotic variance of $\hat{\tau}$ under enrichment proportions $\boldsymbol{p} = (p_1, \dots, p_K)$ and treatment allocations $\boldsymbol{e} = (e_1, \dots, e_K)$ as $\mathbb{V}[\hat{\tau}(\boldsymbol{p}, \boldsymbol{e})]$.

Our design goal is to minimize the variance of estimating τ by working with the optimization Problem 1. By solving this oracle problem, we can find the oracle treatment allocation $\boldsymbol{e}^* = (e_1^*, ..., e_K^*)$ and the oracle subpopulation proportion $\boldsymbol{p}^* = (p_1^*, ..., p_K^*)$. Here, $c_1 \in (0, 1/2)$, p_{0k} is the subpopulation proportion in the EMR data, and p_k is the subpopulation proportion in the trial. κ_0 is the asymptotic proportion of EMR data when combining EMR data and the trial data, and e_k is the treatment allocation in the trial.

Problem 1.

$$\begin{aligned}\min_{\boldsymbol{p},\boldsymbol{e}}\ & \mathbb{V}[\hat{\tau}(\boldsymbol{p},\boldsymbol{e})]\\ &= \sum_{k=1}^{K} p_{0k}^2\Big\{\frac{(\tau_k-\tau)^2}{\kappa_0 p_{0k}+(1-\kappa_0)p_k}\\ &+\frac{\sigma_{1k}^2}{(1-\kappa_0)p_k e_k}+\frac{\sigma_{0k}^2}{(1-\kappa_0)p_k\,(1-e_k)+\kappa_0 p_{0k} r_k}\Big\},\\ \text{s.t. } & \sum_{k=1}^{K} p_k = 1,\ p_k>0,\ k=1,\dots,K\\ & c_1 \leqslant e_k \leqslant 1-c_1,\ k=1,\dots,K\end{aligned}$$

$\tau_k = \mathbb{E}[Y(1)-Y(0)|X\in S_k, Z=0]$ and $\sigma_{dk}^2 = \mathbb{E}\left[\mathbb{V}[Y(d)\mid X, Z=1]\mid X\in S_k\right]$, $d\in\{0,1\}$, denote the treatment effect and the associated variance in subpopulation k, respectively. Lastly, $r_k = \mathbb{E}\left[r(X)\mid X\in S_k\right]$, where $r(X)=\frac{\mathbb{V}(Y(0)|X,Z=1)}{\mathbb{V}(Y(0)|X,Z=0)}$. As the solution to **Problem 1** does not have a closed-form expression, we provide some insights on the oracle solution in a simplified setting in Remark 1.

Remark 1 (Oracle solution in the presence of binary outcomes) *Assume the outcome Y is binary. Under Assumption 1, we have $r_k=1$, hence the oracle solutions are*

$$p_k^* = \left[\frac{w_k}{(1-\kappa_0)\sum_{k=1}^K w_k p_{0k}}-\frac{\kappa_0}{1-\kappa_0}\right]p_{0k},$$
$$e_k^* = \frac{\sigma_{1k}\left(1+\frac{\kappa_0 p_{0k}}{(1-\kappa_0)p_k^*}\right)}{\sigma_{0k}+\sigma_{1k}},$$

where $w_k = \left((\sigma_{1k}+\sigma_{0k})^2+(\tau_k-\tau)^2\right)^{1/2}$.

The oracle subpopulation enrichment proportion p_k^* suggests that when σ_{dk} and τ_k do not differ too much across K subpopulations, $p_k^*\approx p_{0k}$, implying that the oracle subpopulation enrichment proportion in the digital clinical trial mimics the subpopulation proportions in the EMR data. Such an oracle enrichment strategy is particularly beneficial as the trial population can match with the general population, making the evidence derived from digital clinical trials widely generalizable. Furthermore, to understand the oracle treatment assignment probability e_k^*, we can compare it with the classical Neyman allocation approach (Neyman, 1992), where $e_k^{\text{Neyman}} = \sigma_{1k}/(\sigma_{0k}+\sigma_{1k})$. Compared with Neyman allocation, our oracle treatment assignment probability leans toward enrolling more subjects into the treatment arm in subpopulation k. We conjecture that this is because the EMR dataset already provides rich information regarding the control arm. When no EMR data are included, that is $N=0$; our oracle treatment assignment strategy aligns with the Neyman allocation. As suggested by the oracle designs, we reach the promise of the digital clinical trial by broadening the eligibility criteria and estimating the ATE efficiently with adaptive treatment allocation. In Section 4, we shall formally introduce our digital clinical trial design strategies in the non-oracle setting.

4 DIGITAL CLINICAL TRIAL DESIGN STRATEGIES IN TWO SETTINGS

In this section, we propose two novel digital clinical trial designs assisted by EMR data. The first design in Section 4.1 assumes perfect transportability (Assumption 1), while the second design in Section 4.2 is robust to imperfect transportability (violation of Assumption 1). As the parameters τ_k and σ_{dk} in the oracle problem (**Problem 1**) are typically unknown in practice, we propose to sequentially learn the unknown parameters from the digital clinical trials. Based on the accrued data, we refine our understanding of the optimal design strategies (p_k^*, e_k^*). We consider a multi-stage digital clinical trial, where the sample size in each stage, n_t, is large, and the number of stages, T, is small.

4.1 Design under Perfect Transportability

In this section, we shall illustrate our first digital clinical trial design under perfect transportability in Algorithm 1.

In Stage 1 (line 1-2), since we have no prior information about the distribution in the trial population, Stage 1 serves as an exploration stage. We estimate the unknown parameters from both EMR data and the first-stage trial data. The form of the estimators can be found in Supplementary Materials Section A. In Stage t (line 4-6), we find the optimal designs by solving **Problem 2**. The form of $\hat{\tau}_k^{(t-1)}, \hat{\tau}^{(t-1)}, (\hat{\sigma}_{1k}^{(t-1)})^2, (\hat{\sigma}_{0k}^{(t-1)})^2$ and $\hat{r}_k^{(t-1)}$ can be found in the Supplementary Materials. In addition, we rescale the subpopulation proportion and propensity score by the sample sizes of the previous stages: $\tilde{p}_{tk}^* = \frac{1}{n_t}\left(\hat{p}_{tk}^* n - \sum_{s=1}^{t-1} n_{sk}\right)$ and $\tilde{e}_{tk}^* = \frac{1}{n_{tk}}\left(\hat{e}_{tk}^*\sum_{s=1}^{t} n_{sk} - \sum_{s=1}^{t-1} n_{sk}e_{sk}^*\right)$. After Stage T (line 8-9), we construct the final-stage treatment effect estimator $\hat{\tau}$ by Eq (3), and the associated variance estimators as $\hat{\mathbb{V}} =$

$\sum_{k=1}^{K} \frac{N_k^2}{N^2} \left\{ \frac{(\hat{\tau}_k - \hat{\tau})^2}{N_k + n\hat{p}_k} + \frac{(\hat{\sigma}_{1k})^2}{n\hat{p}_k\hat{e}_k} + \frac{(\hat{\sigma}_{0k})^2}{n\hat{p}_k(1-\hat{e}_k)+N_k\hat{r}_k} \right\}$.

Problem 2.

$$\min_{\boldsymbol{p},\boldsymbol{e}} \sum_{k=1}^{K} \frac{N_k^2}{N^2} \left\{ \frac{\left(\hat{\tau}_k^{(t-1)} - \hat{\tau}^{(t-1)}\right)^2}{N_k + np_k} + \frac{\left(\hat{\sigma}_{1k}^{(t-1)}\right)^2}{np_k e_k} + \frac{\left(\hat{\sigma}_{0k}^{(t-1)}\right)^2}{np_k\left(1-e_k\right) + N_k \hat{r}_k^{(t-1)}} \right\},$$

$$\text{s.t.} \sum_{k=1}^{K} p_k = 1, p_k > 0,\ c_1 \leqslant e_k \leqslant 1 - c_1.$$

Lastly, we can construct the two-sided α-level confidence interval for $\hat{\tau}$ as

$$\left[\hat{\tau} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\hat{\mathbb{V}}}\right] \tag{4}$$

4.2 Design under Imperfect Transportability

In practice, it is imperative to carefully assess the impact of violating the perfect transportability assumption and adopt a digital clinical design that is more robust to such a violation. In this section, we propose another digital clinical trial design under imperfect transportability.

First, we characterize the "imperfect transportability" using a bias function, following the notation introduced by Dahabreh *et al.* (2023):

$$u(d, X) = \mathbb{E}\left[Y(d) \mid X, Z=1\right] - \mathbb{E}\left[Y(d) \mid X, Z=0\right],$$
$$\delta(X) = u(1, X) - u(0, X),$$

where $d \in \{0, 1\}$. $u(d, X)$ is the bias function, and $\delta(X)$ denotes the difference of bias functions. It is infeasible to identify $\delta(X)$ because we cannot obtain any information about $\mathbb{E}[Y(1)|X, Z=0]$ from EMR data in the absence of treatment arm information. For simplicity, Dahabreh *et al.* (2023) suggests to use functions that do not depend on covariates, such that

$$u(0, X) \equiv u(0),\ \delta(X) \equiv \delta.$$

Hence, we can have a pair of sensitivity parameters $(u(0), \delta)$. In our setting, we can also consider subpopulation sensitivity parameters $(u(0)_k, \delta_k)$, $k = 1, \ldots, K$.

Second, we consider a new objective function, mean squared error (MSE), to account for the trade-off between efficiency and unbiasedness of estimating τ. This is because in the presence of imperfect transportability, the estimator in Eq (3) is inconsistent. We propose a minimax framework to design an experiment that is robust to the potential violation of the transportability assumption. Similar to **Problem 2**, we can find the optimal designs by solving optimization **Problem 3**, where Γ_0 and Γ_1 are the sensitivity bounds to quantify the range of bias functions. We present the algorithms and considerations for choosing the sensitivity bounds in Supplementary Material Section B. As **Problem 3** does not have a closed-form solution, we can solve the problem by numerical methods.

Lastly, we conduct statistical inference based on the combined EMR data and digital clinical trial data. However, when there exists imperfect transportability, the traditional Gaussian-type confidence intervals in Eq (4) are no longer invalid. Instead, we employ a percentile bootstrap approach (Zhao *et al.*, 2019) to conduct valid inference. This approach effectively addresses both the bias arising from the violation of perfect transportability (Assumption 1) and the inherent estimation uncertainty without requiring an explicit characterization of the asymptotic distributions of the estimators. Note that our approach is different from Zhao *et al.* (2019), as we take bootstrap samples within each stratum and obtain the final bound by weighted average. Our approach is summarized in Algorithm 3.

To start, for fixed Γ_0, Γ_1, we choose a pair of sensitivity parameters $(u(0), \delta)$ such that $|u(0)| \leqslant \Gamma_0, |\delta| \leqslant \Gamma_1$. We define the k-th bias-calibrated treatment effect estimand under the specific parameter $(u(0), \delta)$ as $\tau_k^{(u(0),\delta)} = \widetilde{\tau}_k - \text{Bias}\left(\widetilde{\tau}_k, \tau_k\right)$, and the bias-calibrated estimator as $\hat{\tau}_k^{(u(0),\delta)} = \hat{\tau}_k - \widehat{\text{Bias}}\left(\widetilde{\tau}_k, \tau_k\right)$. We take B bootstrap samples from a subset of the collected data where $X_{it} \in S_k$. For every $(u(0), \delta)$, we can construct a confidence interval for $\tau_k^{(u(0),\delta)}$ via percentile bootstrap: $\left[L_k^{(u(0),\delta)}, U_k^{(u(0),\delta)}\right] = \left[Q_{\alpha/2}\left(\hat{\tau}_k^{*(u(0),\delta)}\right), Q_{1-\alpha/2}\left(\hat{\tau}_k^{*(u(0),\delta)}\right)\right]$, where $\hat{\tau}_k^{*(u(0),\delta)}$ is estimated from bootstrap sample $\{1, 2, \ldots, B\}$, and $Q_{\alpha/2}$ is the $\alpha/2$-percentile of the bootstrap distribution. In Section 5, we will show that $\left[L_k^{(u(0),\delta)}, U_k^{(u(0),\delta)}\right]$ is an asymptotically valid confidence interval for ${\tau_k}^{(u(0),\delta)}$ under sensitivity parameters $(u(0), \delta)$. Lastly, we can construct a $1-\alpha$ confidence interval for τ as $\left[Q_{\alpha/2}\left(\inf_{u(0),\delta} \hat{\tau}^{*(u(0),\delta)}\right), Q_{1-\alpha/2}\left(\sup_{u(0),\delta} \hat{\tau}^{*(u(0),\delta)}\right)\right]$.

5 THEORETICAL INVESTIGATION

In this section, we investigate the theoretical properties of our proposed digital clinical trial design under the perfect transportability in Theorem 1 and Remark 2 and the design under the imperfect transportability in Theorem 2. We start with listing an additional assumption.

Assumption 2 *The sample sizes $N, n_t \to \infty$, such that $\frac{n_t}{N+n} \to \kappa_t$, where $0 < \kappa_t < 1, 0 \leqslant t \leqslant T$.*

Lemma 1 *Under Assumptions 1-2, **Problem 1** has unique solutions $(\boldsymbol{p}^*, \boldsymbol{e}^*)$.*

Lemma 1 says that our solutions to the oracle problem are unique, which can be proved by separating **Problem 1** into two iterative subproblems and showing the convexity of each subproblem. Building onto Lemma 1, we establish the consistency of our first design (Section 4.1) and the asymptotic normality of our proposed estimator under the perfect transportability assumption.

Theorem 1 *Under Assumptions 1-2 and assume all working models are correctly specified, we have $\hat{p}^*_{kt} \xrightarrow{p} p^*_k$, $\hat{e}^*_{kt} \xrightarrow{p} e^*_k$, where $t = 2, \ldots, T$, $k = 1, \ldots, K$. Under regularity conditions described in Theorems 2.6 and 3.4 of Newey and McFadden (1994), we have*

$$\sqrt{N+n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}^*)$$

where $\mathbb{V}^* = \sum_{k=1}^K p_{0k}^2 \left\{ \frac{(\tau_k - \tau)^2}{\kappa_0 p_{0k} + (1-\kappa_0) p_k^*} + \frac{\sigma_{1k}^2}{(1-\kappa_0) p_k^* e_k^*} + \frac{\sigma_{0k}^2}{(1-\kappa_0) p_k^* (1-e_k^*) + \kappa_0 p_{0k} r_k} \right\}$.

Theorem 1 shows that, on the one hand, our design strategies are consistent with the oracle design strategies. This suggests that our design operates in a similar manner as the oracle design, thus it can realize the promise of digital clinical trials as specified in Section 3. On the other hand, our proposed design-based estimator $\hat{\tau}$ asymptotically converges to a Gaussian distribution and it validates our constructed confidence interval in Eq (4).

Problem 3.

$$\min_{\boldsymbol{p},\boldsymbol{e}} \max_{\delta, u(0)} \sum_{k=1}^K \frac{N_k^2}{N^2} \left\{ \frac{\left(\hat{\tau}_k^{(t-1)} - \hat{\tau}^{(t-1)}\right)^2}{N_k + np_k} + \frac{\left(\hat{\sigma}_{1k}^{(t-1)}\right)^2}{np_k e_k} + \frac{\left(\hat{\sigma}_{0k}^{(t-1)}\right)^2}{np_k(1-e_k) + N_k \hat{r}_k^{(t-1)}} + \left(\delta + \frac{N_k \hat{r}_k^{(t-1)}}{np_k(1-e_k) + N_k \hat{r}_k^{(t-1)}} u(0)\right)^2 \right\},$$

$$\text{s.t. } |\delta| \leqslant \Gamma_0, |u(0)| \leqslant \Gamma_1,$$

$$\sum_{k=1}^K p_k = 1, p_k > 0, \; c_1 \leqslant e_k \leqslant 1 - c_1,$$

Algorithm 1 Digital clinical trial design strategy

```
Stage 1 (Initialization):
Enroll $n_1$ participants, set the subpopulation proportion as $p_{1k} = \frac{1}{K}$, and propensity score $e_{1k} = \frac{1}{2}$, $k = 1, \ldots, K$.
Compute $\hat{\tau}_k^{(1)}, \hat{\tau}^{(1)}, \left(\hat{\sigma}_{1k}^{(1)}\right)^2, \left(\hat{\sigma}_{0k}^{(1)}\right)^2$ and $\hat{r}_k^{(1)}$ from $\{(Y_{is}, X_{is}, D_{is}, Z_{is})_{i=1}^{n_s}\}_{s=0}^{1}$.
Stage $t$ (Multi-stage adaptive experiment):
for $t \to 2$ to $T$ do
    Obtain $(\hat{p}^*_{tk}, \hat{e}^*_{tk})$ by solving Problem 2.
    Enroll $n_t$ participants, set the subpopulation proportion and propensity score as $(\tilde{p}^*_{tk}, \tilde{e}^*_{tk})$, by rescaling $(\hat{p}^*_{tk}, \hat{e}^*_{tk})$.
    Update $\hat{\tau}_k^{(t-1)}, \hat{\tau}^{(t-1)}, \left(\hat{\sigma}_{1k}^{(t-1)}\right)^2, \left(\hat{\sigma}_{0k}^{(t-1)}\right)^2$ and $\hat{r}_k^{(t-1)}$ from $\{(Y_{is}, X_{is}, D_{is}, Z_{is})_{i=1}^{n_s}\}_{s=0}^{t-1}$.
end for
Stage $T$ (Inference):
Update $\hat{\tau}_k, \hat{\tau}, (\hat{\sigma}_{1k})^2, (\hat{\sigma}_{0k})^2$ and $\hat{r}_k$ from $\{(Y_{is}, X_{is}, D_{is}, Z_{is})_{i=1}^{n_s}\}_{s=0}^{T}$.
Construct a two-sided $\alpha$-level confidence interval for $\hat{\tau}$ in Eq (4).
```

Algorithm 2 Percentile Bootstrap Algorithm

```
(Bootstrap sampling)
for $b = 1$ to $B$ do
    for $k = 1$ to $K$ do
        Generate a bootstrap sample from a subset of the collected data where $X_{it} \in S_k$.
        Compute the bounds of potential point estimates in $k$-th strata:
        $\left[\hat{\tau}_k^{(b)} \pm \max_{u(0)_k, \delta_k} \widehat{Bias}^{(b)}(\tilde{\tau}_k, \tau_k)\right]$
    end for
    Compute weighted average bounds over $K$ strata.
end for
From $B$ bootstrapped sample bounds, construct CI based on $\alpha/2$ and $1 - \alpha/2$ percentile.
```

Since the design strategy for Stage t is dependent on the historical data, we do not have the classic i.i.d assumption. To address this challenge, we leverage the empirical process arguments similar to Hahn *et al.* (2011) to establish the asymptotic normality result. To provide more insights regarding the benefits of our design, we compare the asymptotic efficiency gain of our EMR data-assisted digital clinical trial for estimating $\hat{\tau}$ with the benchmark digital clinical trial without using EMR data in Remark 2.

Remark 2 *Let $(\boldsymbol{p}^{*\prime}, \boldsymbol{e}^{*\prime})$ be the oracle solution corresponding to the benchmark design strategy without* *integrating EMR data. The asymptotic variance of the benchmark design-based estimator can be written as* $\mathbb{V}^{*\prime} = \sum_{k=1}^{k} \frac{p_{0k}^2}{p_k^{*\prime}} \left\{ (\tau_k - \tau)^2 + \frac{\sigma_{1k}^2}{e_k^{*\prime}} + \frac{\sigma_{0k}^2}{1-e_k^{*\prime}} \right\}$. *We can show that* $(N+n)^{-1}\mathbb{V}^* < n^{-1}\mathbb{V}^{*\prime}$.

Remark 2 suggests that our design has efficiency gain by incorporating EMR data. Lastly, we investigate the theoretical properties of our second design under the imperfect transportability assumption in Section 4.2.

Theorem 2 *Under the assumption that* $r(x;\hat{\gamma})$ *is correctly specified and uniformly bounded in* x *and the regularity conditions described in Theorems 2.6 and 3.4 of Newey and McFadden (1994), for fixed sensitivity parameters* $(u(0),\delta)$*, we have*

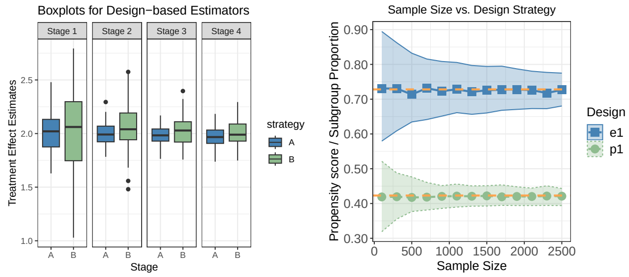
$$\limsup_{N_k,n_k\to\infty} \mathbb{P}\left(\tau_k^{(u(0),\delta)} < L_k^{(u(0),\delta)}\right) \leqslant \frac{\alpha}{2},$$

$$\limsup_{N_k,n_k\to\infty} \mathbb{P}\left(\tau_k^{(u(0),\delta)} > U_k^{(u(0),\delta)}\right) \leqslant \frac{\alpha}{2}.$$

Theorem 2 says that the proposed percentile bootstrap confidence interval in section 4.2 delivers valid inference. The proofs can be found in the Supplementary Materials, which involve proving the asymptotic normality of $\hat{\tau}_k^{(u(0),\delta)}$ and the bootstrap sample estimator $\hat{\tau}_k^{*(u(0),\delta)}$.

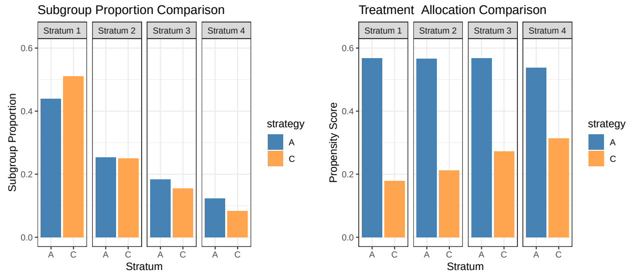
6 SIMULATION STUDIES

In this section, we conduct simulation studies to investigate our proposed designs. Simulation setups are provided in the Supplementary Materials Section F. We consider three designs: (A) our proposed design under **Problem 2**, (B) a design that does not integrate EMR data, and (C) our proposed design under **Problem 3**. Under the perfect transportability assumption, we compare designs A and B in Figure 1. Under the imperfect transportability, we compare designs A and C in Figure 2 and verify the statistical validity of our percentile bootstrap procedure in Table 1. We further compare the mean squared errors (MSE) of designs B and C in Figure 3.



(a) (b)

Figure 1: (a) The efficiency comparison between designs A and B. (b) The convergence of our design strategies to the oracle designs.



(a) (b)

Figure 2: Design strategies comparison between designs A and C.

Figure 1 (a) demonstrates that by incorporating external EMR data, our proposed design A has higher estimation efficiency compared to the benchmark design B, which validates the theoretical insights in Remark 2. Figure 1 (b) shows that our design strategies converge to the oracle design strategies as the sample size increases. Figure 2 demonstrates that under imperfect transportability, the oracle treatment allocation e^* in design C is lower than that under design A. Furthermore, in design A, the values of p_k^* closely resemble the subpopulation proportions p_{0k} in the EMR population, whereas in design C, there is a noticeable deviation from this resemblance. Figure 2 suggests that design C is more robust to the violation of the perfect transportability assumption. Regarding the choice of sensitivity bounds, Table 1 suggests design C is more sensitive to Γ_0 than to Γ_1. However, as long as the sensitivity bounds (Γ_0, Γ_1) can control the bias parameters $(u(0), \delta)$, the percentile bootstrap confidence interval has high coverage probabilities.

$(u(0),\delta)$	(Γ_0,Γ_1)	Coverage	CI
$(0,0)$	$(0,0)$	0.958	$[1.828, 2.149]$
(0.5,0.5)	(0,0)	0.005	$[2.356, 2.826]$
	(0.5,0)	0.140	$[2.287, 3.168]$
	(0,0.5)	0.811	$[1.871, 3.304]$
	(0.5,0.5)	0.987	$[1.777, 3.682]$
	(1,0.5)	0.997	$[1.604, 3.870]$
	(0.5,1)	1	$[1.236, 4.072]$

Table 1: Percentile bootstrap confidence intervals under various tuning parameters.

In Figure 3, we investigate the potential benefits of leveraging external EMR data, particularly in the presence of imperfect transportability. The results indicate that design C outperforms benchmark design B regarding MSE when the violation of the transportability assumption is not severe. However, under severe violation, design C may experience efficiency loss compared to design B. Based on the simulation results, we recommended practitioners first assess the transportability assumption (as described in Supplementary Materials Section G) and then consider adopting design C if the violation is not too severe.

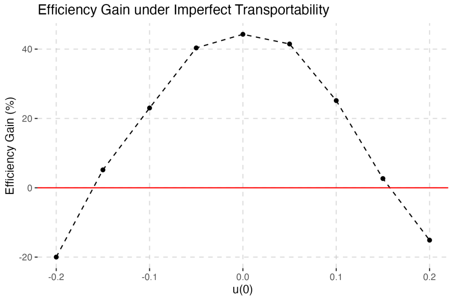


Figure 3: Efficiency (MSE) comparison between designs B and C under imperfect transportability.

7 A SYNTHETIC CASE STUDY

In this case study, we aim to investigate the treatment effect of a cash transfer program on the viral suppression rate in HIV patients leveraging an external EMR database. This case study showcases the performance of our design in a realistic setting, as there is no perfect transportability between the trial data and the EMR data. The assessment of the transportability assumption can be found in Supplementary Materials Section G. We consider a Tanzania EMR database containing $295,961$ patient records from 2015 to 2023. The original trial was a cash transfer RCT conducted in Tanzania where subjects are randomized to either the treatment or control arm (Fahey *et al.*, 2020). The treatment is "receiving cash transfer", and the control is "not receiving any cash transfer". The outcome is Viral load suppression, which is a binary biomarker for HIV status, where Viral load < 1000 ($Y = 1$) and Viral load $\geqslant 1000$ ($Y = 0$). We consider two subpopulations defined by biological sex: Males ($k = 1$) and females ($k = 2$). The detailed synthetic data-generating process is provided in Supplementary Materials Section G.

As the perfect transportability assumption does not hold, we consider the second design strategy proposed in Section 4.2 and set a covariate-dependent sensitivity bound $\Gamma_0 = (0.1, 0)$. We follow a simulation setup similar to Section 6 to mimic multi-stage clinical trials. We construct the confidence interval by percentile bootstrap in Figure 4 with respect to various choices of Γ_1. When Γ_1 is large ($\Gamma_1 > 0.4$), we obtain a rather conservative confidence interval, reflecting our belief of strong unmeasured confounding. Otherwise, we obtain a narrower confidence interval, suggesting that the cash transfer program can significantly improve the proportion of patients recovered from HIV. Figure 4 not only provides realistic guidance for practitioners to decide the scale of sensitive bound based on domain knowledge but also demonstrates the estimation efficiency under our design in the presence of moderately imperfect transportability.

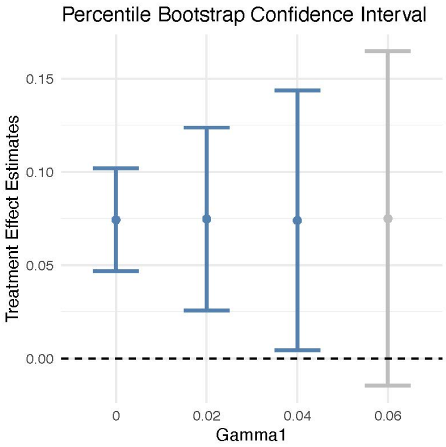


Figure 4: Percentile bootstrap confidence interval with respect to various choices of Γ_1.

8 DISCUSSION

In this paper, we propose two digital clinical trial designs leveraging external EMR data. While recognizing the diversity of patient information available in traditional EMRs, our approach selectively incorporates only those elements of patient covariates and outcomes that align with the data collection requirements of a digital clinical trial. We look forward to exploring the possibilities of integrating additional information from EMRs into future trial designs.

Acknowledgements

We thank the reviewers for their insightful and constructive feedback, which has greatly helped us in improving the paper. The second author acknowledges the support NSF-DMS-2239047 and NIH-R01MH125746 awards.

References

Ayaad, O., Alloubani, A., ALhajaa, E. A., Farhan, M., Abuseif, S., Al Hroub, A., and Akhu-Zaheya, L. (2019). "The role of electronic medical records in improving the quality of health care services: Comparative study," *International journal of medical informatics*, *127*, 63–67.

Bates, D. W., Ebell, M., Gotlieb, E., Zapp, J., and Mullins, H. (2003). "A proposal for electronic medical records in US primary care," *Journal of the American Medical Informatics Association*, *10*(1), 1–10.

Bickel, P. J. and Freedman, D. A. (1981). "Some asymptotic theory for the bootstrap," *The annals of statistics*, *9*(6), 1196–1217.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. (2018). "Generalizing evidence from randomized trials using inverse probability of sampling weights," *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(4), 1193–1209.

Colnet, B., Josse, J., Varoquaux, G., and Scornet, E. (2022). "Causal effect on a target population: a sensitivity analysis to handle missing covariates," *Journal of Causal Inference*, *10*(1), 372–414.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernan, M. A. (2020). "Extending inferences from a randomized trial to a new target population," *Statistics in medicine*, *39*(14), 1999–2014.

Dahabreh, I. J., Robins, J. M., Haneuse, S. J.-P., Saeed, I., Robertson, S. E., Stuart, E. A., and Hernán, M. A. (2023). "Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population," *Statistics in Medicine*.

Dimakopoulou, M., Ren, Z., and Zhou, Z. (2021). "Online multi-armed bandits with adaptive inference," *Advances in Neural Information Processing Systems*, *34*, 1939–1951.

Eggenberger, F. and Pólya, G. (1923). "Über die statistik verketteter vorgänge," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, *3*(4), 279–289.

Fahey, C. A., Njau, P. F., Katabaro, E., Mfaume, R. S., Ulenga, N., Mwenda, N., Bradshaw, P. T., Dow, W. H., Padian, N. S., Jewell, N. P. *et al.* (2020). "Financial incentives to promote retention in care and viral suppression in adults with HIV initiating antiretroviral therapy in Tanzania: a three-arm randomised controlled trial," *The Lancet HIV*, *7*(11), e762–e771.

Fehrenbacher, L., Ackerson, L., and Somkin, C. (2009). "Randomized clinical trial eligibility rates for chemotherapy (CT) and antiangiogenic therapy (AAT) in a population-based cohort of newly diagnosed non-small cell lung cancer (NSCLC) patients," *Journal of Clinical Oncology*, *27*(15_suppl), 6538–6538.

Frankovich, J., Longhurst, C. A., and Sutherland, S. M. (2011). "Evidence-based medicine in the EMR era," *N Engl J Med*, *365*(19), 1758–1759.

Gandhi, M., Ameli, N., Bacchetti, P., Sharp, G. B., French, A. L., and Young, M. (2005). "Eligibility criteria for HIV clinical trials and generalizability of results: The gap between published reports and study protocols," *AIDS*, *19*, 1885–1896.

Garcia, A. and et al. (2022). "Lessons learned in the Apple Heart Study and implications for the data management of future digital clinical trials," *Journal of Biopharmaceutical Statistics*, *32*(3), 496–510.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2021). "Confidence intervals for policy evaluation in adaptive experiments," *Proceedings of the national academy of sciences*, *118*(15), e2014602118.

Hahn, J., Hirano, K., and Karlan, D. (2011). "Adaptive experimental design using the propensity score," *Journal of Business & Economic Statistics*, *29*(1), 96–108.

Hoffman, M. A. and Williams, M. S. (2011). "Electronic medical records and personalized medicine," *Human genetics*, *130*, 33–39.

Hu, F. and Rosenberger, W. F. (2003). "Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons," *Journal of the American Statistical Association*, *98*(463), 671–678.

Inan, O., Tenaerts, P., Prindiville, S., Reynolds, H., Dizon, D., Cooper-Arnold, K., Turakhia, M., Pletcher, M., Preston, K., Krumholz, H. *et al.* (2020). "Digitizing clinical trials," *NPJ digital medicine*, *3*(1), 101.

Kallus, N. and Zhou, A. (2021). "Minimax-optimal policy learning under unobserved confounding," *Management Science*, *67*(5), 2870–2890.

Lai, T. L., Lavori, P. W., and Tsang, K. W. (2019). "Adaptive enrichment designs for confirmatory trials," *Statistics in medicine*, *38*(4), 613–624.

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D., and Cai, J. (2023). "Improving trial generalizability using observational studies," *Biometrics*, *79*(2), 1213–1225.

Li, X., Miao, W., Lu, F., and Zhou, X.-H. (2023). "Improving efficiency of inference in clinical trials with external control data," *Biometrics*, *79*(1), 394–403.

Luedtke, A., Carone, M., and van der Laan, M. J. (2019). "An omnibus non-parametric test of equality in distribution for unknown functions," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *81*(1), 75–99.

Newey, W. K. and McFadden, D. (1994). "Large sample estimation and hypothesis testing," *Handbook of econometrics*, *4*, 2111–2245.

Neyman, J. (1992). "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection," In *Breakthroughs in Statistics: Methodology and Distribution*: Springer, 123–150.

Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., Balasubramanian, V., Russo, A. M., Rajmane, A., Cheung, L. *et al.* (2019). "Large-scale assessment of a smartwatch to identify atrial fibrillation," *New England Journal of Medicine*, *381*(20), 1909–1917.

Rosenberger, W. F. and Lachin, J. M. (1993). "The use of response-adaptive designs in clinical trials," *Controlled clinical trials*, *14*(6), 471–484.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, *66*(5), 688.

Shi, C., Zhu, J., Ye, S., Luo, S., Zhu, H., and Song, R. (2022). "Off-policy confidence interval estimation with confounded Markov decision process," *Journal of the American Statistical Association*, 1–12.

Shi, L., Wang, J., and Wu, T. (2023). "Statistical Inference on Multi-armed Bandits with Delayed Feedback."

Simon, N. and Simon, R. (2013). "Adaptive enrichment designs for clinical trials," *Biostatistics*, *14*(4), 613–625.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). "On the application of probability theory to agricultural experiments. Essay on principles. Section 9.," *Statistical Science*, 465–472.

Stallard, N. (2023). "Adaptive enrichment designs with a continuous biomarker," *Biometrics*, *79*(1), 9–19.

Steinhubl, S. R., Wolff-Hughes, D. L., Nilsen, W., Iturriaga, E., and Califf, R. M. (2019). "Digital clinical trials: creating a vision for the future," *NPJ digital medicine*, *2*(1), 126.

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., and Wang, G. (2018). "Data processing and text mining technologies on electronic medical records: a review," *Journal of healthcare engineering*, *2018*.

Thall, P. F. (2021). "Adaptive enrichment designs in clinical trials," *Annual review of statistics and its application*, *8*, 393–411.

Wang, S. J., Middleton, B., Prosser, L. A., Bardon, C. G., Spurr, C. D., Carchidi, P. J., Kittler, A. F., Goldszer, R. C., Fairchild, D. G., Sussman, A. J. *et al.* (2003). "A cost-benefit analysis of electronic medical records in primary care," *The American journal of medicine*, *114*(5), 397–403.

Williamson, S. F., Jacko, P., Villar, S. S., and Jaki, T. (2017). "A Bayesian adaptive design for clinical trials in rare diseases," *Computational statistics & data analysis*, *113*, 136–153.

Yang, S., Zeng, D., and Wang, X. (2020). "Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding," *arXiv preprint arXiv:2007.12922*.

Zelen, M. (1969). "Play the winner rule and the controlled clinical trial," *Journal of the American Statistical Association*, *64*(325), 131–146.

Zhan, A., Mohan, S., Tarolli, C., Schneider, R. B., Adams, J. L., Sharma, S., Elson, M. J., Spear, K. L., Glidden, A. M., Little, M. A. *et al.* (2018). "Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score," *JAMA neurology*, *75*(7), 876–880.

Zhang, K., Janson, L., and Murphy, S. (2021). "Statistical inference with m-estimators on adaptively collected data," *Advances in neural information processing systems*, *34*, 7460–7471.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). "Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *81*(4), 735–761.

Checklist

1. For all models and algorithms presented, check if you include:
   - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   - (b) Complete proofs of all theoretical results. [Yes]
   - (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   - (a) Citations of the creator If your work uses existing assets. [Yes]
   - (b) The license information of the assets, if applicable. [Not Applicable]
   - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   - (d) Information about consent from data providers/curators. [Yes]
   - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Electronic Medical Records Assisted Digital Clinical Trial Design: Supplementary Materials

A The form of estimators in Stage t

In stage t ($2 \leqslant t \leqslant T$), we update parameters $\tau_k, \tau, \sigma_{1k}, \sigma_{0k}$ and r_k in the optimization problem by the following estimates obtained from past information $\{(Y_{is}, X_{is}, D_{is}, Z_{is}), i = 1, \cdots, n_s\}_{s=0}^{t-1}$.

$$\begin{aligned}
\hat{\tau}_k^{(t-1)} = & \frac{1}{\sum_{s=0}^{t-1}\sum_{i=1}^{n_s} 1\{X_{is} \in S_k\}} \times \sum_{s=0}^{t-1}\sum_{i=1}^{n_s} 1\{X_{is} \in S_k\} \\
& \times \Bigg[\frac{(1 - Z_{is})}{1 - \hat{\pi}_k^{(t-1)}} \left(\hat{m}_1^{(t-1)}(X_{is}) - \hat{m}_0^{(t-1)}(X_{is}) \right) \\
& + \frac{D_{is} Z_{is}}{\hat{\pi}_k^{(t-1)} \hat{e}_k^{(t-1)}} \left(Y_{is} - \hat{m}_1^{(t-1)}(X_{is}) \right) \\
& - \frac{Z_{is}(1 - D_{is}) + (1 - Z_{is})\hat{r}^{(t-1)}(X_{is})}{\hat{\pi}_k^{(t-1)}(1 - \hat{e}_k^{(t-1)}) + (1 - \hat{\pi}_k^{(t-1)})\hat{r}^{(t-1)}(X_{is})} \left(Y_{is} - \hat{m}_0^{(t-1)}(X_{is}) \right) \Bigg]
\end{aligned}$$

$$\hat{\tau}^{(t-1)} = \sum_{k=1}^{K} \frac{N_k}{N} \hat{\tau}_k^{(t-1)}$$

For $d \in \{0, 1\}$ and stratum k, we have the variance estimator

$$(\hat{\sigma}_{dk}^{(t-1)})^2 = \frac{\sum_{s=1}^{t-1}\sum_{i=1}^{n_s} \left(Y_{is} 1\{D_{is} = d\} 1\{X_{is} \in S_k\} - \frac{1}{\sum_{s=1}^{t-1} n_s} \sum_{s=1}^{t-1}\sum_{i=1}^{n_s} \left(Y_{is} 1\{D_{is} = d\} 1\{X_{is} \in S_k\} \right) \right)^2}{\sum_{s=1}^{t-1}\sum_{i=1}^{n_s} 1\{X_{is} \in S_k\} 1\{D_{is} = d\}}$$

And the variance ratio estimator

$$\hat{r}_k^{(t-1)} = (\hat{\sigma}_{0k}^{(t-1)})^2 / (\hat{\sigma}_{0k}'^{(t-1)})^2$$

Where $(\hat{\sigma}_{0k}'^{(t-1)})^2$ is the estimate of $V(Y(0) \mid X, Z = 0)$:

$$(\hat{\sigma}_{0k}'^{(t-1)})^2 = \frac{\sum_{i=1}^{N} \left(Y_i 1\{D_{i0} = d\} 1\{X_{i0} \in S_k\} - \frac{1}{N} \sum_{i=1}^{N} \left(Y_{i0} 1\{D_{i0} = d\} 1\{X_{i0} \in S_k\} \right) \right)^2}{\sum_{i=1}^{N} 1\{X_{i0} \in S_k\} 1\{D_{i0} = d\}}$$

$\hat{\pi}_k^{(t-1)}$ is the estimated proportion of individuals enrolled in the trial among subgroup k:

$$\hat{\pi}_k^{(t-1)} = \left(\sum_{s=0}^{t-1}\sum_{i=1}^{n_s} 1\{X_{is} \in S_k\} \right)^{-1} \times \sum_{s=0}^{t-1}\sum_{i=1}^{n_s} Z_{is} 1\{X_{is} \in S_k\}$$

$\hat{e}_k^{(t-1)}$ is the estimated propensity score from the trial data

$$\hat{e}_k^{(t-1)} = \left(\sum_{s=1}^{t-1}\sum_{i=1}^{n_s} 1\{X_{is} \in S_k\} \right)^{-1} \times \sum_{s=1}^{t-1}\sum_{i=1}^{n_s} D_{is} 1\{X_{is} \in S_k\}$$

and $\hat{m}_d^{(t-1)}(X_{is})$, $\hat{r}^{(t-1)}(X_{is})$ are the regression / nonparametric estimators based on past information $\{(Y_{is}, X_{is}, D_{is}, Z_{is}), i = 1, \cdots, n_s\}_{s=0}^{t-1}$.

B Choosing bounds in practice and alternative approach to address Problem 3

In **Problem 3**, we propose a minimax framework to design an experiment that is robust to imperfect transportability. The choice of appropriate sensitivity bounds (Γ_0 and Γ_1) is a critical aspect of this framework as they quantify the range of bias functions. The choice of sensitivity bounds is a delicate balance. Choosing excessively large sensitivity bounds leads to better uniform control over a larger range of bias functions, but it can become overly conservative if the actual bias is constrained by narrower bounds, resulting in an efficiency loss. Conversely, choosing overly narrow sensitivity bounds yields worse uniform control of imperfect transportability. This creates a trade-off that needs to be carefully navigated.

The ideal approach to choosing sensitivity bounds is to leverage background knowledge about the trial and the general population. When the background knowledge is limited, a practical strategy for setting Γ_0 is

$$\Gamma_0 = \sup_{X \in \mathcal{X}} \mid \hat{\mathbb{E}}[Y \mid X, D = 0, Z = 1] - \hat{\mathbb{E}}[Y \mid X, Z = 0] \mid$$

where $\hat{\mathbb{E}}[Y \mid X, D = 0, Z = z]$ is the regression-based estimator. For Γ_1, as no treatment data is available in the EMR data, we employ a calibration algorithm inspired by the calibration plot introduced by Kallus and Zhou (2021). The basic idea of the algorithm is to find Γ_1 that minimizes the expected MSE. In the absence of prior information about the true bound Γ_1^*, we assume Γ_1^* is uniformly distributed over a wide range $\left[\underline{\Gamma_1}, \overline{\Gamma}_1\right]$ (line 1). Our calibration algorithm for choosing Γ_1 is outlined in Algorithm 3.

Algorithm 3 Calibration Algorithm

```
1: Choose a sequence of possible $\Gamma_1$ candidates: $\Gamma_1^{(1)}, \Gamma_1^{(2)}, \Gamma_1^{(3)}, \ldots \Gamma_1^{(M)}$ evenly from $\left[\underline{\Gamma_1}, \overline{\Gamma_1}\right]$.
2: for $m, m' = 1$ to $M$ do
3:     Treat $\Gamma_1^{(m')}$ as the true bound $\Gamma_1^*$.
4:     Calculate the minimax MSE using the solved optimal design $\left(p_k^*\left(\Gamma_1^{(m)}\right), e_k^*\left(\Gamma_1^{(m)}\right)\right)$ and true bound $\Gamma_1^{(m')}$: $\mathrm{MSE}\left(p_k^*\left(\Gamma_1^{(m)}\right), e_k^*\left(\Gamma_1^{(m)}\right), \Gamma_1^{(m')}\right)$.
5: end for
6: Find $\Gamma_1^{(m)}$ that minimizes the empirical expectation:
       $$\Gamma_1^* = \mathrm{argmin}_{\Gamma_1^{(m)}} \frac{1}{M} \sum_{m'=1}^{M} \mathrm{MSE}\left(p_k^*\left(\Gamma_1^{(m)}\right), e_k^*\left(\Gamma_1^{(m)}\right), \Gamma_1^{(m')}\right)$$
```

Problem 3 does not have a closed-form solution. To gain a better understanding of the solutions, we propose an alternative approach by decomposing the objective function into two components: variance and squared bias. We optimize the two parts separately and then combine the solutions by taking a weighted average to arrive at the final solution. Given that the e_k solution for minimizing squared bias is zero, we can express the solution for the weighted average propensity score as

$$e_k^* = (1 - \lambda) \frac{\sigma_{1k}\left(1 + \frac{\kappa_0 p_{0k}}{(1-\kappa_0) p_k^*} r_k\right)}{\sigma_{0k} + \sigma_{1k}}$$

Compared to the oracle solution under perfect transportability, this outcome manifests as a shrinkage solution that tends toward zero. This behavior is due to the fact that when there exists imperfect transportability, the trial data can be considered "contaminated" when estimating τ, and we need to enroll fewer treatment units to control estimation bias. Furthermore, it's important to note that the degree of shrinkage is positively correlated with factors such as p_{0k}, Γ_0, and Γ_1.

C Proof of Lemma 1

In **Problem 1**, we define oracle treatment allocation e_k^* and subgroup proportion p_k^* as the solution of an optimization problem. Since the constraints of two optimization variables:

$$\mathcal{E}_1 = \left\{ p_k : \sum_{k=1}^{K} p_k = 1, p_k > 0, k = 1, \ldots, K \right\}$$
$$\mathcal{E}_2 = \{ e_k : \quad c_1 \leqslant e_k \leqslant 1 - c_1, k = 1, \ldots, K \}$$

are both convex, we can apply Von Newman's alternating projection Lemma and separate the **Problem 1** into two iterative subproblems, which are easier to solve. In the first subproblem, we optimize e_k given p_k :

$$\begin{aligned} &\min_{e_k} \mathbb{V}\left(e_k ; p_k\right) \\ &\text{s.t. } c_1 \leqslant e_k \leqslant 1 - c_1, k = 1, \ldots, K \end{aligned}$$

The Hessian matrix is

$$H_1 = \operatorname{diag}\left\{ 2p_{0k^2} \left(\frac{\sigma_{1k}^2}{(1-\kappa_0) p_k e_k^3} + \frac{(1-\kappa_0)^2 p_k^2 {\sigma_{0k}}^2}{\left((1-\kappa_0) p_k \left(1 - e_k\right) + \kappa_0 p_{0k} r_k\right)^3} \right) \right\}_{k=1,\ldots,K}$$

which is apparently positive definite since the diagonal elements of the diagonal matrix are all positive. Hence $V\left(e_k ; p_k\right)$ is strictly convex, and the subproblem has a unique solution

$$e_k^*\left(p_k\right) = \frac{\sigma_{1k}}{\sigma_{1k} + \sigma_{0k}} \left(1 + \frac{\kappa_0 p_{0k}}{(1-\kappa_0) p_k} r_k \right), k = 1, \ldots, K$$

Now we plug in $e_k^*\left(p_k\right)$ to the second subproblem and solve p_k :

$$\begin{aligned} &\min_{p_k} \mathbb{V}\left(p_k\right) \\ &\text{s.t. } \sum_{k=1}^{k} p_k = 1, p_k > 0, k = 1, \ldots, K \end{aligned}$$

Similarly, we can derive the Hessian matrix to show $\mathbb{V}\left(p_k\right)$ is strictly convex

$$H_2 = \operatorname{diag}\left\{ 2(1-\kappa_0)^2 p_{0k}^2 \left(\frac{\left(\tau_k - \tau\right)^2}{\left(\kappa_0 p_{0k} + (1-\kappa_0) p_k\right)^3} + \frac{\left(\sigma_{1k} + \sigma_{0k}\right)^2}{\left(\kappa_0 p_{0k} r_k + (1-\kappa_0) p_k\right)^3} \right) \right\}_{k=1,\ldots,K}$$

Hence, the oracle solution is unique. Similarly, we can show that the sample analogs of oracle problems also have unique solutions.

D Proof of asymptotic properties of design strategies (Theorem 1)

We begin by establishing several lemmas.

Lemma 2 (Convergence of quantities) *Following the approach in Hahn et al. (2011), we make the following assumptions for simplicity in our analysis:*

(i) N and n_t tend to infinity, with $\frac{n_t}{N+n} \rightarrow \kappa_t$ and $\frac{N}{N+n} \rightarrow \kappa_0$, where $n = \sum_{t=1}^{T} n_t$ and $0 < \kappa < 1$.

(ii) The covariates X_i follow a multinomial distribution with finite support.

(iii) Let $e'_{tk} = plim \tilde{e}_{tk}^*$ and $p'_{tk} = plim \tilde{p}_{tk}^*$. We assume that $\tilde{e}_{tk}^* = e'_{tk} + O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\widetilde{p}_{tk}^* = p'_{tk} + O_p\left(\frac{1}{\sqrt{n}}\right)$. Also, $e'_k = \frac{\sum_{t=1}^{T} n_t p'_{tk} e'_{tk}}{\sum_{t=1}^{T} n_t p'_{tk}}$ and $p'_k = \frac{\sum_{t=1}^{T} n_t p'_{tk}}{\sum_{t=1}^{T} n_t}$.

(iv) All working models are correctly specified.

Under these assumptions, we can conclude:

$$\begin{aligned}
m_d(x) - \hat{m}_d(x) &= O_p\left(n^{-1/2}\right) \\
\mathbb{V}_d(x) - \hat{\mathbb{V}}_d(x) &= O_p\left(n^{-1/2}\right) \\
r(x) - \hat{r}(x) &= O_p\left(n^{-1/2}\right) \\
\hat{p}_{tk} - \hat{p}^*_{tk} &= O_p\left(n^{-1/2}\right) \\
\hat{e}_{tk} - \hat{e}^*_{tk} &= O_p\left(n^{-1/2}\right)
\end{aligned}$$

where $m_d(X) = \mathbb{E}[Y \mid X, D = d, Z = 1], \mathbb{V}_d(X) = \mathbb{V}(Y \mid X, D = d, Z = 1), d \in \{0, 1\}$.

Note that e'_{tk} and p'_{tk} are defined as probability limits, which do not necessarily equal the oracle solutions e^*_k and p^*_k. The second assumption ensures we can directly employ a "sample-mean version" estimator to estimate the correct model rather than relying on nonparametric identification. For $\hat{m}_0(x)$, we use all the control units from external data and T-stage trial data (to enhance efficiency). For other estimators, we exclusively use T-stage trial data.

Proof. To establish $m_1(x) - \hat{m}_1(x) = O_p\left(n^{-1/2}\right)$, we utilize empirical process arguments as presented in Hahn *et al.* (2011). We treat the treatment indicators as independent and identically distributed, generated from a uniform distribution, i.e., $D_{it} = 1\left\{U_{it} \leqslant \widetilde{e}^*_{kt}\right\}$, and $1\left\{X_{it} = x\right\}$ as $1\left\{U_{it} \leqslant p^*_{kt} P\left(X = x \mid X \in S_k\right)\right\}$. It's worth noting that p^*_{kt} is correlated with previous data, whereas $\mathbb{P}\left(X = x \mid X \in S_k\right)$ is not.

$$\begin{aligned}
\hat{m}_1(x) =& \frac{\sum_{t=1}^{T}\sum_{i=1}^{n_t} D_{it} Y_{it} 1\left\{X_{it} = x\right\}}{\sum_{t=1}^{T}\sum_{i=1}^{n_t} D_{it} 1\left\{X_{it} = x\right\}} \\
=& \left(\frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1} Y_{i1} 1\left\{U_{i1} \leqslant \widetilde{e}^*_{k1}\right\} 1\left\{U_{i1} \leqslant \widetilde{p}^*_{k1}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)\right\}\right. \\
&\left. + \frac{n_2}{n}\frac{1}{n_2}\sum_{i=1}^{n_2} Y_{i2} 1\left\{U_{i2} \leqslant \widetilde{e}^*_{k2}\right\} 1\left\{U_{i2} \leqslant \widetilde{p}^*_{k2}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)\right\} + \cdots\right) \\
&/ \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n_t} D_{it} 1\left\{X_{it} = x\right\}
\end{aligned}$$

We begin by considering the numerator. For the first term, since $\hat{e}_{k_1} = e'_{k_1} = \frac{1}{2}, \hat{p}_{k_1} = p'_{k_1} = \frac{1}{K}$, which are uncorrelated with the previous data, we can apply the LLN and CLT,

$$\begin{aligned}
& \frac{n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1} Y_{i1} 1\left\{U_{i1} \leqslant \widetilde{e}^*_{k1}\right\} 1\left\{U_{i1} \leqslant \widetilde{p}^*_{k1}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)\right\} \\
=& \kappa_1 e'_{k1} p'_{k1}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right) m_1(x) + O_p\left(\frac{1}{\sqrt{n_1}}\right)
\end{aligned}$$

For the second term, it's important to note that the set of functions

$$\left\{1\left(U_{i2} \leqslant \widetilde{e}^*_{k2}\right) 1\left(U_{i2} \leqslant \widetilde{p}^*_{k2}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)\right) Y_{i2}\right\}$$

is Euclidean and stochastic equicontinuous. Therefore, we have

$$\begin{aligned}
& \frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2} Y_{i2} 1\left\{U_{i2} \leqslant \widetilde{e}^*_{k2}\right\} 1\left\{U_{i2} \leqslant \widetilde{p}^*_{k2} p\left(X = x \mid X \in S_k, Z = 1\right)\right\} \\
=& \frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2} Y_{i2} 1\left\{U_{i2} \leqslant e'_{k2}\right\} 1\left\{U_{i2} \leqslant p'_{k2} p\left(X = x \mid X \in S_k, Z = 1\right)\right\} \\
&+ G_1\sqrt{n_2}\left(\widetilde{e}^*_{k2} - e'_{k2}\right) + G_2\sqrt{n_2}\left(\widetilde{p}^*_{k2} - p'_{k2}\right) + O_p(1)
\end{aligned}$$

where

$$\begin{aligned}G_1 &= \frac{\partial}{\partial e_{k2}}\mathbb{E}\left[1\left\{U_{i2} \leqslant e_{k2}\right\}1\left\{U_{i2} \leqslant p'_{k2}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)\right\}Y_{i2}\right] \mid e_{k2} = e'_{k2} \\ &= m_1(x)p'_{k2}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)\end{aligned}$$

Similarly,

$$G_2 = m_1(x)e'_{k2}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)$$

Given our assumptions that $\tilde{e}^*_{k2} - e'_{k2} = O_p\left(\frac{1}{\sqrt{n}}\right), \tilde{p}^*_{k2} - p'_{k2} = O_p\left(\frac{1}{\sqrt{n}}\right)$, by LLN and CLT the numerator can be written as

$$\kappa_2 e'_{k2}p'_{k2}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)m_1(x) + O_p\left(\frac{1}{\sqrt{n_2}}\right)$$

We handle the 3-T terms in the numerator and the 1-T terms in the denominator similarly. This allows us to express $\hat{m}_1(x)$ as

$$\begin{aligned}\hat{m}_1(x) = &\left(\sum_{t=1}^{T}\kappa_t e'_{kt}p'_{kt}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)m_1(x) + O_p\left(\frac{1}{\sqrt{n}}\right)\right) \\ &/\left(\sum_{t=1}^{T}\kappa_t e'_{kt}p'_{kt}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right) + O_p\left(\frac{1}{\sqrt{n}}\right)\right) \\ = &\, m_1(x) + O_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

The proof for $m_0(x) - \hat{m}_0(x) = O_p\left(n^{-1/2}\right)$ follows a similar but slightly different approach. This is because $\hat{m}_0(x)$ is estimated using data from all control units in the EMR data and trial data

$$\hat{m}_0(x) = \frac{\sum_{i=1}^{T}\sum_{i=1}^{n_t}\left(1 - D_{it}\right)Y_{it}1\left\{X_{it} = x\right\} + \sum_{i=1}^{N}\left(1 - D_{i0}\right)Y_{i0}1\left\{X_{i0} = x\right\}}{\sum_{t=1}^{T}\sum_{i=1}^{n_t}\left(1 - D_{it}\right)1\left\{X_{it} = x\right\} + \sum_{i=1}^{N}\left(1 - D_{i0}\right)1\left\{X_{i0} = x\right\}}$$

In this context, we use $t = 0$ to represent the EMR data for simplicity. As the EMR data exclusively consists of control units ($D_{i0} \equiv 0$), we can apply the CLT

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left(1 - D_{i0}\right)Y_{i0}1\left\{X_{i0} = x\right\} &= \mathbb{E}(Y \mid X, D = 0, Z = 0)\mathbb{P}\left(X_i = x \mid Z_i = 0\right) + O_p\left(\frac{1}{\sqrt{N}}\right) \\ &= m_0(x)\mathbb{P}\left(X_i = x \mid Z_i = 0\right) + O_p\left(\frac{1}{\sqrt{N}}\right)\end{aligned}$$

The ultimate equation is a consequence of the transportability assumption. Similar to the proof of $\hat{m}_1(x) - m_1(x) = O_p\left(\frac{1}{\sqrt{n}}\right)$, we can express $\hat{m}_0(x)$ as

$$\begin{aligned}\hat{m}_0(x) = &\left(\sum_{t=1}^{T}\kappa_t\left(1 - e'_{kt}\right)p'_{kt}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right)m_0(x) + \kappa_0\mathbb{P}(X = x \mid Z = 0)m_0(x)\right) \\ &/\left(\sum_{t=1}^{T}\kappa_t\left(1 - e'_{kt}\right)p'_{kt}\mathbb{P}\left(X = x \mid X \in S_k, Z = 1\right) + \kappa_0\mathbb{P}(X = x \mid Z = 0)\right) \\ = &\, m_0(x) + O_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

The proof of the other equalities is similar. For $\mathbb{V}_1(x) - \hat{\mathbb{V}}_1(x) = O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\mathbb{V}_0(x) - \hat{\mathbb{V}}_0(x) = O_p\left(\frac{1}{\sqrt{n}}\right)$, we consider Y^2_{it} instead of Y_{it}. As $\hat{r}(x)$ is the ratio of two variance estimators that converge to zero at a rate of $\frac{1}{\sqrt{n}}$, we have $r(x) - \hat{r}(x) = O_p\left(\frac{1}{\sqrt{n}}\right)$. Regarding $\hat{p}_{tk} - \hat{p}^*_{tk} = O_p\left(\frac{1}{\sqrt{n}}\right)$, by employing similar empirical process arguments, we obtain

$$\begin{aligned}
\hat{p}_{tk} &= \frac{\sum_{s=1}^{t}\sum_{i=1}^{n_n} 1\left\{X_{it} \in S_k\right\}}{\sum_{s=1}^{t} n_s} \\
&= \frac{n_1 \tilde{p}_{1k}^{*} + O_p\left(\frac{1}{\sqrt{n}}\right)}{\sum_{s=1}^{t} n_s + O_p\left(\frac{1}{\sqrt{n}}\right)} + \frac{n_2 \tilde{p}_{2k}^{*} + O_p\left(\frac{1}{\sqrt{n}}\right)}{\sum_{s=1}^{t} n_s + O_p\left(\frac{1}{\sqrt{n}}\right)} + \dots \\
&= \hat{p}_{tk}^{*} + O_p\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

The proof of $\hat{e}_{tk} - \hat{e}_{tk}^{*} = O_p\left(\frac{1}{\sqrt{n}}\right)$ follows the same approach. □

Lemma 3 (Asymptotic normality I) *Under the regularity conditions described in Theorems 2.6 and 3.4 of Newey and McFadden (1994), and assumptions (i)-(iv) described in Lemma 2, we have*

$$\sqrt{N+n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

where

$$\begin{aligned}
\sigma^2 &= \sum_{k=1}^{K} p_{0k}^2 \left\{ \frac{\left(\tau_k - \tau\right)^2}{\kappa_0 p_{0k} + \left(1-\kappa_0\right) p_k'} + \frac{\sigma_{1k}^2}{\left(1-\kappa_0\right) p_k' e_k'} + \frac{\sigma_{0k}^2}{\left(1-\kappa_0\right) p_k' \left(1-e_k'\right) + \kappa_0 p_{0k} r_k} \right\} \\
\hat{\tau} &= \sum_{k=1}^{K} \frac{N_k}{N} \hat{\tau}_k \\
\hat{\tau}_k &= \left(\sum_{t=0}^{T}\sum_{i=1}^{n_t} 1\left\{X_{it} \in S_k\right\}\right)^{-1} \sum_{t=0}^{T}\sum_{i=1}^{n_t} 1\left\{X_{it} \in S_k\right\} \left[\frac{1-Z_{it}}{1-\hat{\pi}_k}\left(\hat{m}_1\left(X_{it}\right) - \hat{m}_0\left(X_{it}\right)\right)\right. \\
&\quad \left. + \frac{D_{it} Z_{it}}{\hat{\pi}_k \hat{e}_k}\left(Y_{it} - \hat{m}_1\left(X_{it}\right) - \frac{Z_{it}\left(1-D_{it}\right) + \left(1-Z_{it}\right)\hat{r}\left(X_{it}\right)}{\hat{\pi}_k\left(1-\hat{e}_k\right) + \left(1-\hat{\pi}_k\right)\hat{r}\left(X_{it}\right)}\left(Y_{it} - \hat{m}_0\left(X_{it}\right)\right)\right]
\end{aligned}$$

Proof. By the Lemma 2 we have $\hat{p}_{tk} - \hat{p}_{tk}^{*} = O_p\left(\frac{1}{\sqrt{n}}\right)$, then

$$\hat{\pi}_{tk} = \frac{\hat{p}_{tk}\sum_{s=1}^{t} n_s}{N_k + \hat{p}_{tt}\sum_{s=1}^{t} n_s} = \frac{\hat{p}_{tk}^{*}\sum_{s=1}^{t} n_s + O_p\left(\frac{1}{\sqrt{n}}\right)}{N_k + \hat{p}_{tk}^{*}\sum_{s=1}^{t} n_s + O_p\left(\frac{1}{\sqrt{n}}\right)} = \hat{\pi}_{tk}^{*} + O_p\left(\frac{1}{\sqrt{n}}\right)$$

Let $t = T$, we have $\hat{\pi}_k = \hat{\pi}_k^{*} + O_p\left(\frac{1}{\sqrt{n}}\right) = \pi_k' + O_p\left(\frac{1}{\sqrt{n}}\right)$. Then we can rewrite the nominator of $\hat{\tau}_k$ as

$$\begin{aligned}
(I) = \frac{1}{N+n}\sum_{t=0}^{T}\sum_{i=1}^{n_t} 1\{U_{it} \leqslant \hat{p}_k^{*}\} &\left[\frac{1 - 1\{U_{it} \leqslant \hat{\pi}_k^{*}\}}{1-\pi_k'}(m_1(X_{it}) - m_0(X_{it}))\right. \\
&+ \frac{1\{U_{it} \leqslant \hat{e}_k^{*}\}1\{U_{it} \leqslant \hat{\pi}_k^{*}\}}{\pi_k' e_k'}(Y_{it} - m_1(X_{it})) \\
&\left. - \frac{1\{U_{it} \leqslant \hat{e}_k^{*}\}(1 - 1\{U_{it} \leqslant \hat{\pi}_k^{*}\})r(X_{it})}{\pi_k'(1-e_k') + (1-\pi_k')r(X_{it})}(Y_{it} - m_0(X_{it}))\right] + O_p\left(\frac{1}{\sqrt{N+n}}\right)
\end{aligned}$$

Here, we consider the T-stage trial data a unified dataset and utilize weighted average parameters such as $\hat{p}_k^{*}, \hat{e}_k^{*}$. By the empirical process arguments similar to Lemma 2, we have

$$\begin{aligned}
(I) = \frac{1}{N+n}\sum_{t=0}^{T}\sum_{i=1}^{n_t} 1\{U_{it} \leqslant p_k'\} &\left[\frac{1 - 1\{U_{it} \leqslant \pi_k'\}}{1-\pi_k'}(m_1(X_{it}) - m_0(X_{it}))\right. \\
&+ \frac{1\{U_{it} \leqslant e_k'\}1\{U_{it} \leqslant \pi_k'\}}{\pi_k' e_k'}(Y_{it} - m_1(X_{it})) \\
&\left. - \frac{1\{U_{it} \leqslant e_k'\}(1 - 1\{U_{it} \leqslant \pi_k'\})r(X_{it})}{\pi_k'(1-e_k') + (1-\pi_k')r(X_{it})}(Y_{it} - m_0(X_{it}))\right] + O_p\left(\frac{1}{\sqrt{N+n}}\right)
\end{aligned}$$

Here, p_k', π_k', and e_k' are uncorrelated with the previous data. Consequently, under the regularity conditions outlined in Theorems 2.6 and 3.4 of Newey and McFadden (1994), we can apply the classical CLT to establish the asymptotic normality of $\hat{\tau}$

$$\sqrt{N+n}(\hat{\tau}-\tau) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

□

Now that we have established Lemma 2 and Lemma 3, we can proceed to prove Theorem 1. We begin by demonstrating the consistency of our design strategies.

Theorem 3 (Consistency with oracle) *Under Assumption 1 and regularity conditions (i)-(iv) described in Lemma 2, we have*

$$\hat{p}_{kt}^* \xrightarrow{p} p_k^*$$
$$\hat{e}_{kt}^* \xrightarrow{p} e_k^*$$

where $t = 1, \ldots, T$, $k = 1, \ldots, K$.

Proof. For simplicity, define

$$\mathbb{V}\left(\boldsymbol{p}, \boldsymbol{e}\right) = \sum_{k=1}^{K} p_{0k}^2 \left\{ \frac{\left(\tau_k-\tau\right)^2}{\kappa_0 p_{0k} + (1-\kappa_0) p_k} + \frac{\sigma_{1k}^2}{(1-\kappa_0) p_k e_k} + \frac{\sigma_{0k}^2}{(1-\kappa_0) p_k \left(1-e_k\right) + \kappa_0 p_{0k} r_k} \right\}$$

Then the oracle problem can be written as

$$\min_{\boldsymbol{p}, \boldsymbol{e}} \mathbb{V}\left(\boldsymbol{p}, \boldsymbol{e}\right)$$
$$\text{s.t. } \left(\boldsymbol{p}, \boldsymbol{e}\right) \in \mathcal{E}$$

We have proved that the oracle problem has unique solution $\mathcal{E}^* = (\boldsymbol{p}^*, \boldsymbol{e}^*)$. We define

$$\hat{\mathbb{V}}_t\left(\boldsymbol{p}, \boldsymbol{e}\right) = \sum_{k=1}^{K} \frac{N_k^2}{N^2} \left\{ \frac{\left(\hat{\tau}_k^{(t-1)} - \hat{\tau}^{(t-1)}\right)^2}{N_k + n p_k} + \frac{\left(\hat{\sigma}_{ik}^{(t-1)}\right)^2}{n p_k e_k} + \frac{\left(\hat{\sigma}_{0k}^{(t-1)}\right)^2}{n p_k \left(1-e_k\right) + N_k \hat{r}_k^{(t-1)}} \right\}$$

and the t-stage optimization problem can be written as

$$\min_{\boldsymbol{p}, \boldsymbol{e}} \hat{\mathbb{V}}_t\left(\boldsymbol{p}, \boldsymbol{e}\right)$$
$$\text{s.t. } \left(\boldsymbol{p}, \boldsymbol{e}\right) \in \mathcal{E}$$

The unique solution is $\hat{\mathcal{E}}_t^* = (\hat{\boldsymbol{p}}_t^*, \hat{\boldsymbol{e}}_t^*)$. We further define "approximate minimizes",

$$\mathcal{E}^*(\varepsilon) = \left\{ (\boldsymbol{p}, \boldsymbol{e}) : (\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}, \mathbb{V}\left(\boldsymbol{p}, \boldsymbol{e}\right) \geqslant \max_{\boldsymbol{p}, \boldsymbol{e}} \mathbb{V}\left(\boldsymbol{p}, \boldsymbol{e}\right) - \varepsilon \right\}$$
$$\hat{\mathcal{E}}_t^*(\varepsilon_t) = \left\{ (\boldsymbol{p}, \boldsymbol{e}) : (\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}, \hat{\mathbb{V}}_t\left(\boldsymbol{p}, \boldsymbol{e}\right) \geqslant \max_{\boldsymbol{p}, \boldsymbol{e}} \hat{\mathbb{V}}_t\left(\boldsymbol{p}, \boldsymbol{e}\right) - \varepsilon_t \right\}$$

where ε and ε_t are adapted to optimization slackness, and $\mathcal{E}^*(\varepsilon)$ and $\hat{\mathcal{E}}_t^*(\varepsilon_t)$ are the set of solutions. We assume that $\varepsilon_t \to 0$ as $n_t \to \infty, t \geqslant 2$. We further define the δ-enlargement of $\mathcal{E}^*$ as $\mathcal{E}^* + B_\delta$:

$$\mathcal{E}^* + B_\delta = \left\{ (p_k^* + u, e_k^* + v) : \|u\| + \|v\| = \delta \right\}.$$

First, for $\delta > 0$, we can define

$$\varepsilon = \sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}} \mathbb{V}(\boldsymbol{p}, \boldsymbol{e}) - \sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E} : \|\boldsymbol{p} - \boldsymbol{p}^*\| + \|\boldsymbol{e} - \boldsymbol{e}^*\| \geqslant \delta} \mathbb{V}(\boldsymbol{p}, \boldsymbol{e})$$

Therefore, for any $\delta > 0$, there exists $\varepsilon > 0$, such that $\mathcal{E}^*(\varepsilon) \subseteq \mathcal{E}^* + B_\delta$. Consider the event

$$A_t(\varepsilon/3) : \sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}} \left| \widehat{\mathbb{V}}_t(\boldsymbol{p}, \boldsymbol{e}) - \mathbb{V}(\boldsymbol{p}, \boldsymbol{e}) \right| \leqslant \varepsilon/3$$

Second, for an arbitrary $(\boldsymbol{p}, \boldsymbol{e}) \notin \mathcal{E}^* + B_\delta$, we have

$$\mathbb{V}(\boldsymbol{p}, \boldsymbol{e}) < \sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}} \mathbb{V}(\boldsymbol{p}, \boldsymbol{e}) - \varepsilon$$

Under the event $A_t(\varepsilon/3)$,

$$\widehat{\mathbb{V}}_t(\boldsymbol{p}, \boldsymbol{e}) \leqslant \mathbb{V}(\boldsymbol{p}, \boldsymbol{e}) + \frac{\varepsilon}{3} < \sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}} \mathbb{V}(\boldsymbol{p}, \boldsymbol{e}) - \varepsilon + \frac{\varepsilon}{3}.$$

In addition, since $\left|\sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}} \mathbb{V}(\boldsymbol{p}, \boldsymbol{e}) - \sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}} \widehat{\mathbb{V}}_t(\boldsymbol{p}, \boldsymbol{e})\right| \leqslant \varepsilon/3$,

$$\widehat{\mathbb{V}}_t(\boldsymbol{p}, \boldsymbol{e}) \leqslant \mathbb{V}(\boldsymbol{p}, \boldsymbol{e}) + \frac{\varepsilon}{3} < \sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}} \mathbb{V}(\boldsymbol{p}, \boldsymbol{e}) - \frac{2}{3}\varepsilon \leqslant \sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}} \widehat{\mathbb{V}}_t(\boldsymbol{p}, \boldsymbol{e}) - \frac{1}{3}\varepsilon$$

which gives that

$$\widehat{\mathcal{E}}_t^*(\varepsilon/3) \subseteq \mathcal{E}^* + B_\delta$$

Since we assume that $\varepsilon_t \to 0$ when $n_t \to \infty$, then

$$\widehat{\mathcal{E}}_t^*(\varepsilon_t) \subseteq \widehat{\mathcal{E}}^*(\varepsilon/3) \subseteq \mathcal{E}^* + B_\delta$$

Therefore, as $n_t \to \infty$,

$$\mathbb{P}\left(A_t(\varepsilon/3)\right) \leqslant \mathbb{P}\left(\widehat{\mathcal{E}}_t^*(\varepsilon_t) \subseteq \mathcal{E}^* + B_\delta\right)$$

Now we simply need to show $\mathbb{P}\left(A_t(\varepsilon/3)\right) \to 1$ as $n_t \to \infty$. We have previously established in Lemma 2 and 3 that $\hat{\tau}_k^{(t-1)} \to \tau_k, \hat{\tau}^{(t-1)} \to \tau, \left(\hat{\sigma}_{1k}^{(t-1)}\right)^2 \to \sigma_{1k}^2, \left(\hat{\sigma}_{0k}^{(t-1)}\right)^2 \to \sigma_{0k}^2$ for $k = 1, \ldots, K, 2 \leqslant t \leqslant T$. Based on the results, we can derive

$$\sup_{(\boldsymbol{p}, \boldsymbol{e}) \in \mathcal{E}} \left|\widehat{\mathbb{V}}_t(\boldsymbol{p}, \boldsymbol{e}) - \mathbb{V}(\boldsymbol{p}, \boldsymbol{e})\right| \to 0$$

when $n_t \to \infty$. As a result, $\mathbb{P}\left(A_t(\varepsilon/3)\right) \to 1$ and $\mathbb{P}\left(\widehat{\mathcal{E}}_t^*(\varepsilon_t) \subseteq \mathcal{E}^* + B_\delta\right) \to 1$. Consequently,

$$\mathbb{P}\left(\|\hat{\boldsymbol{p}}_t^* - \boldsymbol{p}^*\| + \|\hat{\boldsymbol{e}}_t^* - \boldsymbol{e}^*\| \leqslant \delta\right) \to 1$$

and

$$\hat{p}_{kt}^* \xrightarrow{p} p_k^*$$
$$\hat{e}_{kt}^* \xrightarrow{p} e_k^*$$

where $t = 2, \ldots, T$, $k = 1, \ldots, K$. □

After proving Theorem 3, we can then move on to demonstrate the latter part of Theorem 1.

Theorem 4 (Asymptotic normality II) *Under assumptions described in Theorem 3 and regularity conditions described in theorems 2.6 and 3.4 of Newey and McFadden (1994), we have*

$$\sqrt{N+n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}^*\right)$$

where

$$\mathbb{V}^* = \sum_{k=1}^{K} p_{0k}^2 \left\{\frac{\left(\tau_k - \tau\right)^2}{\kappa_0 p_{0k} + \left(1 - \kappa_0\right) p_k^*} + \frac{\sigma_{1k}^2}{\left(1 - \kappa_0\right) p_k^* e_k^*} + \frac{\sigma_{0k}^2}{\left(1 - \kappa_0\right) p_k^* \left(1 - e_k^*\right) + \kappa_0 p_{0k} r_k}\right\}$$

Proof. We have demonstrated in Lemma 2 that $p_k' = plim \hat{p}_k^*, e_k' = plim \hat{e}_k^*$. Moreover, as established in Theorem 3, the oracle design strategies $p_k^* = plim \hat{p}_k^*$ and $e_k^* = plim \hat{e}_k^*$. Consequently, we can conclude that $p_k' = p_k^*$ and $e_k' = e_k^*$, allowing us to replace the probability limits p_k' and e_k' in the lemma with the oracle values p_k^* and e_k^*. □

We now show the asymptotic efficiency gain of our EMR data-assisted digital clinical trial for estimating $\hat{\tau}$ with the benchmark clinical trial without using EMR data.

Theorem 5 (Efficiency gain) *Efficiency gain by utilizing EMR data:*

$$(N+n)^{-1}\mathbb{V}^{*}<n^{-1}\mathbb{V}^{*\prime}$$

Proof. Since $(p_k^{*\prime}, e_k^{*\prime})$ is suboptimal for **Problem 1** in comparison to (p_k^{*}, e_k^{*}), we can deduce that:

$$\begin{aligned}
&\sum_{k=1}^{K} p_{0k}^{2}\left\{\frac{(\tau_k-\tau)^2}{Np_{0k}+np_k^{*}}+\frac{\sigma_{1k}^{2}}{np_k^{*}e_k^{*}}\right\}+\frac{\sigma_{0k}^{2}}{np_k^{*}\left(1-e_k^{*}\right)+Np_{0k}r_k}\Bigg\}\\
\leqslant &\sum_{k=1}^{K} p_{0k}^{2}\left\{\frac{(\tau_k-\tau)^2}{Np_{0k}+np_k^{*\prime}}+\frac{\sigma_{1k}^{2}}{np_k^{*\prime}e_k^{*\prime}}\right\}+\frac{\sigma_{0k}^{2}}{np_k^{*\prime}\left(1-e_k^{*\prime}\right)+Np_{0k}r_k}\Bigg\}
\end{aligned}$$

Given that $r_k \geqslant 0$, and at least one $p_{0k}>0$ for $k=1,\cdots,K$, it is evident that:

$$\begin{aligned}
&\sum_{k=1}^{K} p_{0k}^{2}\left\{\frac{(\tau_k-\tau)^2}{Np_{0k}+np_k^{*\prime}}+\frac{\sigma_{1k}^{2}}{np_k^{*\prime}e_k^{*\prime}}\right\}+\frac{\sigma_{0k}^{2}}{np_k^{*\prime}\left(1-e_k^{*\prime}\right)+Np_{0k}r_k}\Bigg\}\\
< &\sum_{k=1}^{K} p_{0k}^{2}\left\{\frac{(\tau_k-\tau)^2}{np_k^{*\prime}}+\frac{\sigma_{1k}^{2}}{np_k^{*\prime}e_k^{*\prime}}\right\}+\frac{\sigma_{0k}^{2}}{np_k^{*\prime}\left(1-e_k^{*\prime}\right)}\Bigg\}
\end{aligned}$$

Therefore, we conclude that:

$$(N+n)^{-1}\mathbb{V}^{*}<n^{-1}\mathbb{V}^{*\prime}$$

□

E Proof of validity of percentile bootstrap (Theorem 2)

For simplicity, we consider a scenario where the working models $m(X;\alpha)$ and $r(X;\gamma)$ are estimated using the actual sample prior to the bootstrap procedure. Consequently, $m(X;\alpha)$ and $r(X;\gamma)$ are treated as known functions when the bootstrap is applied. We generate an i.i.d bootstrap sample of size N_k+n_k from the empirical distribution, where N_k samples are drawn from the EMR data and n_k samples from the trial data. This yields a bootstrap sample denoted as $\{Y_i^*, X_i^*, D_i^*, Z_i^*=0\}_{i=1}^{N_k}$ and $\{Y_i^*, X_i^*, D_i^*, Z_i^*=1\}_{i=N_k+1}^{N_k+n_k}$, where we use an asterisk (*) to signify the bootstrapped data. According to Zhao *et al.* (2019), we can establish the validity of the percentile bootstrap method by proving the following lemma.

Lemma 4 *Under the assumption that $r(X;\gamma)$ is correctly specified and uniformly bounded in x, and regularity conditions described in Theorems 2.6 and 3.4 of Newey and McFadden (1994), for fixed sensitivity parameters $(u(0),\delta)$, we have*

$$\sqrt{N_k+n_k}\left(\hat{\tau}_k^{(u(0),\delta)}-\tau_k^{(u(0),\delta)}\right)\xrightarrow{d}\mathcal{N}\left(0,\left(\sigma_k^{(u(0),\delta)}\right)^2\right) \tag{L.1}$$

$$\sqrt{N_k+n_k}\left(\hat{\tau}_k^{*(u(0),\delta)}-\hat{\tau}_k^{(u(0),\delta)}\right)\xrightarrow{d}\mathcal{N}\left(0,\left(\sigma_k^{(u(0),\delta)}\right)^2\right) \tag{L.2}$$

where $\hat{\tau}_k^*$ is computed from the bootstrap sample.

Proof. When the transportability assumption holds, we have proved in Theorem 4 that $\hat{\tau}_k$ is consistent and asymptotically normal for τ_k under certain assumptions. When there exists imperfect transportability, $\hat{\tau}_k$ remains consistent and asymptotically normal, but now for a different estimand:

$$\widetilde{\tau}_k=\mathbb{E}\left[\frac{1-Z}{1-\pi_k}\left(m_1(x)-\widetilde{m_0}(x)\right)+\frac{DZ}{\pi_k e_k}\left(Y-m_1(x)\right)-\frac{Z(1-D)+(1-Z)r(X)}{\pi_k\left(1-e_k\right)+\left(1-\pi_k\right)r(X)}\left(Y-\widetilde{m_0}(X)\right)\right]$$

In other words, we can express $\hat{\tau}_k-\widetilde{\tau}_k=O_p\left(\left(N_k+n_k\right)^{-1/2}\right)$. Consequently, to establish L.1, our goal is to demonstrate that the bias estimator $(\widetilde{\tau}_k,\tau_k)=\widetilde{\tau}_k-\tau_k$ root-n converges to zero. Similar to the proof presented in Lemma 2, we can establish that $\hat{p}_{tk}\xrightarrow{p}\hat{p}_{tk}^{*\prime}, \hat{e}_{tk}\xrightarrow{p}\hat{e}_{tk}^{*\prime}$. It's worth noting that for **Problem 3**, the true design strategies are $\hat{p}^{*\prime}_{tk}$ and $\hat{e}^{*\prime}_{tk}$, rather than $\hat{p}^{*}_{tk}$ and $\hat{e}^{*}_{tk}$. Given that $r(x;\hat{\gamma}) - r(x) = O_p(1)$, $N_k/N - p_{0k} = O_p(1)$, for fixed sensitivity parameters $(u(0), \delta)$, we have

$$
\begin{aligned}
&\delta + \frac{N_k r(x;\hat{\gamma})}{n\hat{p}_{tk}(1-\hat{e}_{tk}) + N_k r(x;\hat{\gamma})} u(0) \\
&= \delta + \frac{(p_{0k} + O_p(1))(r(x) + O_p(1))}{n(\hat{p}^{*\prime}_{tk} + O_p(1))(1 - \hat{e}^{*\prime}_{tk} - O_p(1))/N + (p_{0k} + O_p(1))(r(x) + O_p(1))} u(0) \\
&= \delta + \frac{p_{0k} r(x)}{n\hat{p}^{*\prime}_{tk}(1-\hat{e}^{*\prime}_{tk})/N + p_{0k} r(x)} u(0) + O_p(1)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\widehat{Bias}\left(\widetilde{\tau}_k, \tau_k\right) &= \text{Bias}\left(\widetilde{\tau}_k, \tau_k\right) + O_p\left((N_k + n_k)^{-\frac{1}{2}}\right) \\
\hat{\tau}_k^{(u(0),\delta)} &= \hat{\tau}_k - \widehat{\text{Bias}}\left(\widetilde{\tau}_k, \tau_k\right) \\
&= \widetilde{\tau}_k - \text{Bias}\left(\widetilde{\tau}_k, \tau_k\right) + O_p\left((N_k + n_k)^{-\frac{1}{2}}\right) \\
&= \tau_k^{(u(0),\delta)} + O_p\left((N_k + n_k)^{-\frac{1}{2}}\right)
\end{aligned}
$$

In order to establish L.2, we start by re-expressing $\hat{\tau}_k^{(u(0),\delta)}$ as $(N_k + n_k)^{-1} \sum_{i=1}^{N_k+n_k} \mu_i$, where μ_i is a function of $\{Y_i, X_i, D_i, Z_i\}$ and $(u(0), \delta)$ for simplicity. Given that the working model is estimated prior to the bootstrap, we can represent $\hat{\tau}_k^{*(u(0),\delta)}$ as $(N_k + n_k)^{-1} \sum_{i=1}^{N_k+n_k} \mu_i^*$. The proof for L.2 can be established by applying Theorem 2.1 from Bickel and Freedman (1981).

Both the bootstrap pivot $(N_k + n_k)^{-1} \sum_{i=1}^{N_k+n_k} (\mu_i^* - \mu_i)$ and the classic one $(N_k + n_k)^{-1} \sum_{i=1}^{N_k+n_k} \left(\mu_i - \tau_k^{(u(0),\delta}\right)$ share the same asymptotic distribution. Hence, we observe that

$$
\sqrt{N_k + n_k}\left(\hat{\tau}_k^{*(u(0),\delta)} - \hat{\tau}_k^{(u(0),\delta)}\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\sigma_k^{(u(0),\delta)}\right)^2\right).
$$

If the working models are estimated based on bootstrap samples, we should introduce more rigorous assumptions regarding $r(X, \hat{\gamma})$.

Once we have successfully established both L.1 and L.2, the subsequent steps of the proof closely follow the approach outlined in Corollary C.3 of Zhao *et al.* (2019). □

F Detailed simulation setups

We provide a detailed description of the simulation setups in Section 6. When the transportability assumption holds, we consider the following DGP. We assume that there are two covariates, X_1 and X_2, generated from a multivariate Gaussian distribution

$$
\begin{aligned}
\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \mid (Z = 0) &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}\right) \\
\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \mid (Z = 1) &\sim \mathcal{N}\left(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}\right)
\end{aligned}
$$

The stratification is based on the quantiles of X_1, and the number of strata is $K = 4$. For convenience, we assume that the outcomes have linear expectations and log-linear variance:

$$
\begin{aligned}
Y(1) \mid (X, Z = 1) \sim& \mathcal{N}(3 + X_1 + 0.5X_2, \exp(0.5 + 0.5X_1 + 0.5X_2)), \\
Y(0) \mid (X, Z = 1) \sim& \mathcal{N}(1 + 1.5X_1 + X_2, \exp(2 + 0.5X_1 + 0.5X_2)), \\
Y(0) \mid (X, Z = 0) \sim& \mathcal{N}(1 + 1.5X_1 + X_2, \exp(0.5 + 0.5X_1 + 0.5X_2))
\end{aligned}
$$

The true ATE, denoted as $\mathbb{E}(Y(1) - Y(0) \mid Z = 0)$, is set at 2. For the digital clinical trial, we consider a multi-stage adaptive experiment with $T = 4$ experiment stages. We have a total of $N = 500$ units in the EMR data, and the number of units in each stage of the experiment is defined as $\{n_t\}_{t=1}^{T} = \{250, 500, 750, 1000\}$.

In cases where imperfect transportability exists, the outcomes in the trial population are generated using a different DGP

$$\begin{aligned}Y(1) \mid (X, Z=1) \sim & \mathcal{N}(3 + u(0) + \delta + X_1 + 0.5X_2, \exp(0.5 + 0.5X_1 + 0.5X_2)),\\ Y(0) \mid (X, Z=1) \sim & \mathcal{N}(1 + u(0) + 1.5X_1 + X_2, \exp(2 + 0.5X_1 + 0.5X_2)).\end{aligned}$$

Here, $(u(0), \delta)$ represents the sensitivity parameters (bias functions) discussed in Section 4.2. While the true ATE remains at 2, it's important to note that our proposed estimator will exhibit bias in this scenario.

For the comparison between designs B[1] and C[2], our motivation is that, while design C offers a statistical efficiency advantage, it may also introduce more bias when the transportability assumption is violated. Specifically, the bias in the k'th stratum for design (C) is given by $[\delta + \frac{\kappa_0 p_{0k} r_k}{(1-\kappa_0)p_k(1-e_k)+\kappa_0 p_{0k} r_k} \cdot u(0)]$. In contrast, the bias of design (B) is simply represented by δ. Intuitively, when $u(0)$ and δ have opposite signs, utilizing EMR data can potentially reduce bias. However, this is not necessarily the case when $u(0)$ and δ share the same signs. Drawing inspiration from Section C of Li *et al.* (2023), we maintained a common $\delta = 0$, and varied $u(0)$ within the range of $\pm(0, 0.05, 0.1, 0.15, 0.2)$. For each case, we conducted 1000 replicates and assessed the efficiency gain of incorporating EMR data, measured by mean squared errors (MSE).

G Detailed synthetic data-generating process and assessment of transportability

We provide a detailed description of the synthetic data-generating process in Section 7 of our paper. For the EMR dataset, we generate the outcome in each subgroup k as

$$Y_i(0)|(X_i \in \mathcal{S}_k, Z_i = 0) \sim \text{Bernoulli}(\mu_{0k}), k = 1, 2,$$

where $\boldsymbol{\mu}_0 = (0.90,\ 0.91)'$, and subgroup proportions are $\boldsymbol{p}_0 = (0.38,\ 0.62)'$.

For the RCT data, for $k = 1, 2$, we generate the outcome in subgroup k under treatment or control as

$$\begin{aligned}Y_i(0)|(X_i \in \mathcal{S}_k, Z_i = 1) \sim \text{Bernoulli}(\nu_{0k}),\\ Y_i(1)|(X_i \in \mathcal{S}_k, Z_i = 1) \sim \text{Bernoulli}(\nu_{1k}),\end{aligned}$$

where $\boldsymbol{\nu}_0 = (0.80,\ 0.92)'$, $\boldsymbol{\nu}_1 = (0.96,\ 0.95)'$. The parameter values are obtained through regression analysis using raw data from both the EMR database and the cash transfer RCT conducted in Tanzania.

Before considering which design strategy to employ, it is crucial to assess the transportability assumption. For continuous covariates, a nonparametric omnibus test, as described in Luedtke *et al.* (2019), can be used. However, in our HIV cash transfer study described in Section 7, the stratification is based on a single discrete covariate, biological sex (0/1). In this case, we simply use a t-test for the following hypothesis:

$$H_0 : \mathbb{E}(Y|D = 0, X = k, Z = 1) = \mathbb{E}(Y|D = 0, X = k, Z = 0),$$

This test is applied to the EMR data and stage-1 trial data for each stratum $k = 1, 2$. The resulting p-values indicate a significant violation of the transportability assumption for the $k = 1$ subgroup but not for the $k = 2$ subgroup. Consequently, it is reasonable to consider the design strategy outlined in **Problem 3** and set covariate-dependent sensitivity bounds $(\Gamma_{0k}, \Gamma_{1k})_{k=1}^2$. In this context, we set Γ_{01} and Γ_{11} to be greater than zero, while Γ_{02} and Γ_{12} are both set to zero.

[1] A design that does not integrate EMR data.

[2] Our proposed design under Problem 3.