Contemporary Clinical Trials Communications 38 (2024) 101265

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

SECRETS: Subject-efficient clinical randomized controlled trials using synthetic intervention

Sayeri Lala [*], Niraj K. Jha

Department of Electrical and Computer Engineering, Princeton University, Princeton, 08544, NJ, USA

ARTICLE INFO

Dataset link: https://www.ninds.nih.gov/sites/default/files/migrate-documents/sig_form_revised_508c.pdf

Keywords:
Clinical randomized controlled trials
Counterfactual estimation
Hypothesis testing
Sample efficiency
Synthetic intervention

ABSTRACT

Background: The parallel-group randomized controlled trial (RCT) is commonly used in Phase-3 clinical trials to establish treatment effectiveness but requires hundreds-to-thousands of subjects, making it difficult to implement, which leads to high Phase-3 trial failure rates. One approach to increasing power of a trial is to augment data collected from an RCT with external data from prospective studies or prior RCTs. However, this requires that external data be comparable to data from the study of interest, a condition that does not hold for new interventions or populations being studied. Another approach is to lower sample size requirements by using the cross-over design, which measures individual treatment effects (ITEs) to remove inter-subject variability; however, this design is only suitable for chronic conditions and interventions with effects that wash out rapidly.

Method: We propose a novel and practical framework called SECRETS (Subject-Efficient Clinical Randomized Controlled Trials using Synthetic Intervention) to increase power of any parallel-group RCT by simulating the cross-over design using only data collected from the study. SECRETS first estimates ITEs across all subjects recruited to the RCT by using a state-of-the-art counterfactual estimation algorithm called synthetic intervention (SI). Since SI induces dependencies among the ITEs, we introduce a novel hypothesis testing strategy to test for treatment effectiveness.

Results: We show that SECRETS can increase the power of an RCT while maintaining comparable significance levels; in particular, on three real-world clinical RCTs (Phase-3 trials), SECRETS increases power over the baseline method by **6 – 54**% (average: 21.5%, standard deviation: 15.8%), thereby reducing the number of subjects needed to obtain a typically desired statistical operating point of 80% power and 5% significance level by **25 – 76**% (10-3,957 fewer subjects per arm). Our analyses show that SECRETS increases power by consistently reducing the variance of the average treatment effect, thereby mimicking the effects of a cross-over design.

Conclusion: SECRETS increases subject efficiency of an RCT by simulating the cross-over design using only data collected from the RCT; therefore, it is a feasible solution for increasing the trial's power, especially under settings where satisfying sample size requirements is difficult.

1. Introduction

The randomized controlled trial (RCT) is the gold-standard approach to estimating the population-level or average treatment effect (ATE) of a drug, therapy, or other medical interventions [1,2]. It is generally used in Phase-3 clinical studies to establish a treatment's effect [1,3]. The typical RCT, i.e., the two-arm, parallel-group, superiority trial [1], evaluates whether a treatment is superior to or worse than standard treatment or placebo by estimating the ATE and determining whether the estimation is statistically significant. To estimate the ATE, the RCT first recruits subjects representing the population of interest and randomly divides them between a control arm and a treatment arm, in which subjects receive the standard care/placebo and treatment of interest, respectively. Then, it monitors a pre-specified health metric across both groups in parallel till the end of the study, evaluates the average treatment outcome (defined by the health metric) under each arm, and reports their difference as the ATE. The random allocation

Abbreviations: RCT, Randomized Controlled Trial; ATE, Average Treatment Effect; ITE, Individual Treatment Effect; SECRETS, Subject-Efficient Clinical Randomized Controlled Trials using Synthetic Intervention; SI, Synthetic Intervention; VT, Virtual Twins; NINDS, National Institute of Neurologic Disease and Stroke

[*] Corresponding author.

E-mail addresses: slala@princeton.edu (S. Lala), jha@princeton.edu (N.K. Jha).

https://doi.org/10.1016/j.conctc.2024.101265
Received 16 September 2023; Received in revised form 15 December 2023; Accepted 28 January 2024
Available online 3 February 2024

is crucial to accurately estimating the ATE because it removes confounders by producing more comparable cohorts between the control and treatment arms and allows for standard statistical hypothesis testing to be used to detect target ATEs at desired accuracies [1]. However, as a consequence of between-subject variability in responses, RCTs typically require hundreds-to-thousands of subjects to detect a clinically significant ATE with high accuracy [3,4], making them expensive (a Phase-3 trial costed a median of $21.4M USD over Years 2010–2015 [5]) and difficult to implement, with recruitment generally being the "most difficult task" [1]. Consequently, a large fraction of Phase-3 trials fail because of inadequate sample size [6].

One approach to making RCTs more subject-efficient is to use large external datasets obtained from previous clinical trials, electronic health records, patient registries, etc. [7], to augment or replace existing RCT data, since they can be cheaper and easier to acquire [1,8]. However, ATE estimation from the integrated dataset is difficult because the data are no longer randomized. In particular, since external data can differ from the RCT data of interest in several ways, e.g., subject characteristics, medical protocol, etc., merging the two datasets can introduce confounders [1,7], which precludes accurate ATE estimation [2]. Consequently, external control data need to be carefully curated, preferably from historical RCTs with comparable data collection methods, study endpoints, and study populations, to reduce discrepancies from the RCT of interest [7]. In addition, to reduce the effect of confounders on ATE estimation, Bayesian algorithms or propensity-based matching can be used to construct comparable control and treatment groups [9]. However, the former approach requires concurrent control data while the latter approach requires that there are no hidden confounders and that there is overlap in the covariate distribution between the control and treatment groups [2], conditions that are unlikely to hold since verifying that there are no hidden confounders would require domain expertise and operating over high-dimensional datasets reduces overlap [10]. Consequently, attempts to estimate ATEs from nonrandomized data, whether combined with an existing RCT data or alone, require even larger sample sizes than RCTs [8,11].

Another approach to making RCTs more subject-efficient is to use the cross-over design, which increases power over a parallel RCT by reducing the variance of the ATE estimate [12]. It does this by measuring the individual treatment effect (ITE), defined as the difference in the individual's response under the treatment and control conditions, per subject, and averaging over the ITEs. The variance of the ATE estimate is lower than that under the parallel RCT because the number of observations is doubled (two per subject) and the variance of the outcome measure is expectedly lowered by removing between-patient variability [1,12]. While capable of lowering the sampling complexity relative to that of the parallel RCT [12,13], the cross-over design faces some constraints that preclude its widespread usage. To appreciate these limitations, we consider the simplest version, i.e., the two-period cross-over design. This design compares the two treatments (e.g., standard care/placebo and treatment of interest) by assigning each patient to a random sequence of treatments, in which the patient receives the standard care or treatment of interest in the first period and the opposite in the second period; the randomized sequence is used to rule out any temporal effects. Importantly, a washout phase is inserted between the two periods to remove any carryover effects from the treatment administered in the first period. The need for a washout restricts the types of condition-intervention pairs that can be studied under the cross-over, making it only suitable for chronic conditions for which treatments have effects that rapidly wash out [1,12].

To overcome the practical limitations of using external control datasets and implementing a cross-over design, we propose *simulating* the cross-over design using *only* data from the conducted RCT to increase the trial's power, especially under settings with insufficient sample size. Specifically, we implement our proposal in a new framework called SECRETS (**S**ubject-**E**fficient **C**linical **R**andomiz**E**d Controlled **T**rials using **S**ynthetic Intervention) that estimates ITEs across all subjects from concurrent control and treatment groups of a parallel-group RCT by using a state-of-the-art counterfactual estimation algorithm called synthetic intervention (SI) [14] and applies a novel hypothesis testing algorithm suitable for the estimated ITEs. We validate SECRETS on three real-world clinical RCT datasets, showing that it can boost power over the baseline approach by 6%–54% (average: 21.5%, standard deviation: 15.8%), thereby reduce the sample size needed to match a typical target statistical operating point (i.e., 5% significance level, 80% power) by 25%–76% or 10–3,957 for both the control and treatment groups. We empirically establish the premises underlying our framework; we show that SECRETS successfully simulates the cross-over design by analyzing its effect on the distribution of the ATE and also demonstrate the importance of each component underlying SECRETS through ablation studies.

The rest of the article is organized as follows. We describe our proposed framework in Section 2 and evaluation methodology in Section 3. We present and analyze experimental results in Section 4 and draw conclusions in Section 5.

2. Methods

In this section, we first provide an overview of the SECRETS framework and then describe each step of the framework in detail. We present it in the context of a conventional parallel-group RCT design, i.e., a two-arm, superiority trial.

2.1. Overview

The SECRETS framework, illustrated in Fig. 1, improves power at a given sample size and thereby reduces the sample size required to reach a target power $1-\beta_{target}$ for a target ATE $\mu_{1,target}$ and significance level α_{target}, compared to the RCT. It simulates the cross-over design in order to reduce the variance of the ATE estimate. Given data collected from the RCT of interest, i.e., the observed control data $X_{ctrl} \in \mathbb{R}^{n_a \times n_t}$ and observed treatment data $X_{treat} \in \mathbb{R}^{n_a \times n_t}$, where n_a is the arm size and n_t is the duration of the trial with $t = 1$ indexing the pre-intervention timepoint (i.e., baseline visit) and $t > 1$ indexing the post-intervention timepoints, along with parameters for SI tuning and a function *get_outcome* that calculates the health outcome of interest from a patient's response trajectory under some intervention (i.e., *get_outcome*: $\mathbb{R}^{n_t} \rightarrow \mathbb{R}$), SECRETS first estimates the ITEs across all patients per arm, i.e., $Y_{ctrl} \in \mathbb{R}^{n_a}$ and $Y_{treat} \in \mathbb{R}^{n_a}$, using SI. Then, because SI induces dependencies among the estimated ITEs, the ITEs violate the independence assumption of conventional hypothesis testing strategies [15]; hence, SECRETS uses a novel bootstrapping procedure to implement a two-sided hypothesis test using the control data, the merged ITEs, the parameters for SI tuning, the function *get_outcome*, the number of samples to generate from the null distribution (T), and additional parameters for testing (see Appendix A, Alg. 1 for input descriptions). Next, we describe each step of SECRETS in detail.

2.2. Step 1: Estimate the individual treatment effects

To estimate the ITEs Y_{ctrl} and Y_{treat}, SECRETS uses the *estimate_ITEs* subroutine, shown in Fig. 2. Given data from an arm exposed to the target intervention (X_{target}) and data from an "unexposed" arm ($X_{\neg target}$), i.e., one exposed to a different intervention, *estimate_ITEs* first performs min–max normalization and calls *get_counterfactual* to calculate the counterfactual response for subjects from the unexposed arm under the target intervention ($X_{\neg target,c}$). Specifically, it sets the pre-intervention counterfactual response to the pre-intervention observed response, i.e., $X_{\neg target,pre,c} = X_{\neg target,pre}$, where $X_{\neg target,pre,c}, X_{\neg target,pre} \in \mathbb{R}^{n_a}$, and then uses SI [14] to estimate the counterfactual response over the post-intervention period; it does this by taking linear combinations over corresponding observed responses from subjects exposed to the

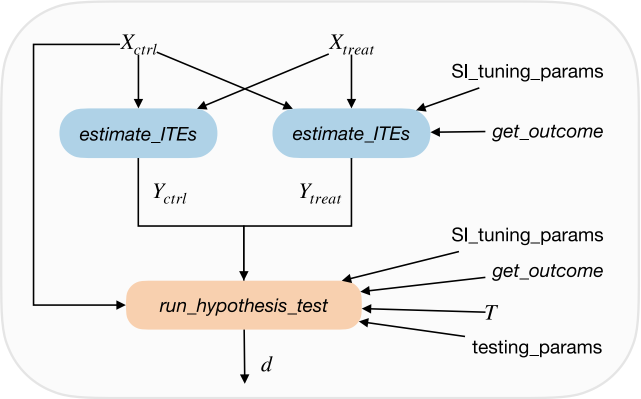

Fig. 1. Flowchart of the SECRETS framework. Note that both calls to *estimate_ITEs* take in SI_tuning_params and *get_outcome* but we have omitted the arguments for brevity.

target intervention (i.e., the “donor”” group) as shown in Eq. (1), where $X_{\neg target,post,c} \in \mathbb{R}^{n_a \times (n_t-1)}$ indexes the post-intervention counterfactual measurements, $W \in \mathbb{R}^{n_a \times n_a}$ contains weights over the donor subjects, and $X_{target,post} \in \mathbb{R}^{n_a \times (n_t-1)}$ indexes the post-intervention observed measurements of the donor subjects.

$$X_{\neg target,post,c} = W X_{target,post} \tag{1}$$

W is learned by performing linear regression using the pre-intervention data along with regularization. We implement the regularization scheme from [16], which first denoises the donor data matrix using low-rank approximation and also incorporates ridge regularization. Specifically, *estimate_ITEs* first tunes hyperparameters (λ_{svt}, the singular value threshold used for low-rank approximation and λ_{ridge}, the ridge regularization strength) used to learn W on X_{target} with the *tune_SI_hyperparams* subroutine and the SI tuning parameters (see Appendix A, Alg. 2) and then runs *SI* with X_{target} as the donor group, the tuned hyperparameters, and each unit from $X_{\neg target}$ as the target unit for which to calculate the counterfactual outcome (see Appendix A, Alg. 3). After unnormalizing $X_{\neg target,c}$, *estimate_ITEs* calls *get_ITE*, which takes the difference between the outcomes under the target and observed interventions, calculated with *get_outcome* (vectorized) on $X_{\neg target,c}$ and $X_{\neg target}$, respectively. In particular, SECRETS calculates Y_{ctrl} by setting $X_{\neg target}$ to X_{ctrl} and X_{target} to X_{treat} and *vice versa* to calculate Y_{treat}.

2.3. Step 2: Conduct a data-driven hypothesis test

After calculating the ITEs, SECRETS conducts a hypothesis test that uses the ITEs to determine whether to reject or not reject the null hypothesis. The standard two-sided one-sample tests are unsuitable because the estimated ITEs are not independent of each other, given that SI uses the same donor pool, i.e., the treatment (control) group, to estimate the counterfactuals per target unit in the control (treatment) group. In addition, existing methods to deal with dependent samples, e.g., testing based on the effective sample size [17], estimating confidence intervals from learned parametric models of sample dependencies [17], and blockwise bootstrapping [18], assume that samples reflect the full dependency structure of the distribution; however, this is not the case with SECRETS because we only have access to one sample of ITEs for a single RCT rather than the full distribution of ITEs generated from different realizations of the RCT (i.e., different samples for the control and treatment groups); specifically, the ITEs for a single RCT exhibit dependencies while ITEs across trials are independent. Therefore, SECRETS uses a new hypothesis testing procedure *run_hypothesis_test*, shown in Fig. 3, that can accurately measure the variance of the distribution from a single sample of ITEs.

In particular, *run_hypothesis_test* implements the critical-value test procedure used in the two-sided one-sample t-test but uses the data to tune the critical value t. To tune t, *run_hypothesis_test* first approximates the null distribution of the test statistic ($\hat{s}_{null}$) by sampling T times from it with $sample\_null$. To generate the ith sample, *sample_null* first simulates a random trial where the treatment effect is null by bootstrap sampling from the original control data (X_{ctrl}) using *random_sample* to construct control and treatment groups with an equal number of subjects as the original control and treatment groups and with comparable responses ($X_{ctrl,sample_i}$ and $X_{treat,sample_i}$). Then, as in Step 1, it runs *estimate_ITEs* (with the SI tuning parameters and function *get_outcome*) to calculate the ITEs for the constructed control and treatment groups. Since the constructed control and treatment groups were exposed to the same intervention (i.e., the control condition), the corresponding ITEs, $Y_{null,ctrl,i}$ and $Y_{null,treat,i}$, are samples from the null distribution. Given the merged set of ITEs, *get_test_stat* calculates the resulting test statistic, $\hat{s}_{null,i}$, using the test statistic formula from the one-sample t-test.

Afterwards, *run_hypothesis_test* tunes the test's critical value with *tune_critical_value*, which uses $\hat{s}_{null}$ and a set of testing parameters that includes the target significance level α_{target} and search range parameters to find the critical value t attaining the target significance level, in a fashion similar to binary search (see Appendix A, Alg. 4).

Finally, *run_hypothesis_test* runs the two-sided test by calculating the test statistic from the ITEs derived from the original control and treatment groups (Y) in Step 1 and comparing its magnitude against the tuned critical value t, rejecting the null hypothesis if $d = 1$ and otherwise failing to reject it.

3. Performance evaluation setup

In this section, we describe the experimental setup used to evaluate SECRETS. First, we present the performance metrics. Then, we describe the baseline method and ablations used to understand the framework. Finally, we describe the datasets used to conduct the experiments along with implementation details.

3.1. Metrics

To assess sample efficiency, we compare the powers obtained for a given sample size of n_a subjects per arm (i.e., control and treatment groups) and target significance level α_{target}. We also report the sample size required to obtain a desired power $1-\beta_{target}$. To measure power $1-\beta$ and significance level α, we follow the approach from [13], which simulates many trials under the alternative and null settings and calculates the percentage of trials where the test procedure returns a reject, respectively. Details of how we simulated trials under each setting are described in Appendix B.1.

To understand how changes in the distribution (mean and variance) of the estimated ATE, i.e., the sample statistic, under the alternative and null settings contribute to changes in power, we consider the power equation for a two-sided z-test [15], i.e., Eq. (2), as a model.

$$1-\hat{\beta} \approx \Phi\left[-z_{1-\alpha_{target}/2} + \frac{|\hat{\mu}_0 - \hat{\mu}_1|}{\hat{\sigma}}\right] \tag{2}$$

The equation states that if the underlying distribution of the sample statistic is normal, then the power of the test, $1-\hat{\beta}$, is approximately given by the cumulative distribution function Φ, evaluated at the negative of the critical value $z_{1-\alpha_{target}/2}$, i.e., the $(1-\alpha_{target}/2)\%$ of a standard normal distribution, shifted by a constant, which we refer to as the *shift term*. The *shift term* is given by Eq. (3). The equation implies that for a given α_{target}, power increases as *shift term* increases, which occurs if the means get further separated and/or the variance of the distributions decreases. Hence, if the equation estimates power well, we can interpret differences in power in terms of differences in the means and variance of the distributions. Details of how $\hat{\mu}_0$, $\hat{\mu}_1$, and $\hat{\sigma}$ are calculated using results from simulated trials are given in Appendix B.2.

$$shift\ term = \frac{|\hat{\mu}_0 - \hat{\mu}_1|}{\hat{\sigma}} \tag{3}$$

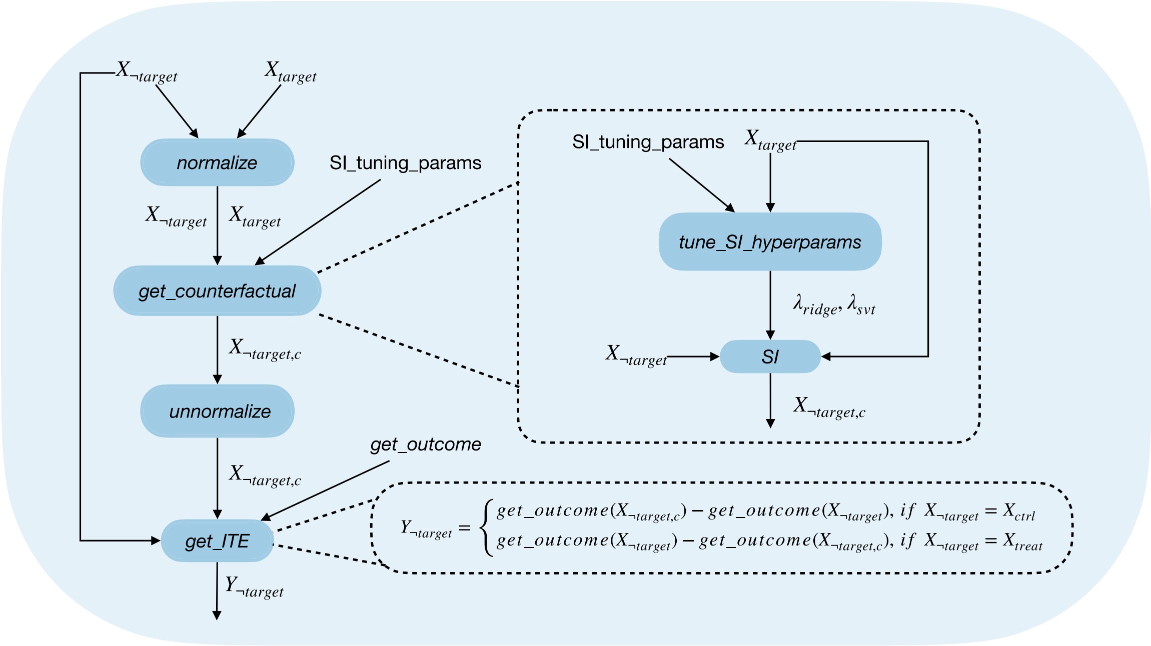
$X_{\neg target}$
X_{target}
normalize
SI_tuning_params
$X_{\neg target}$
X_{target}
get_counterfactual
$X_{\neg target,c}$
unnormalize
$X_{\neg target,c}$
get_outcome
get_ITE
$Y_{\neg target}$
SI_tuning_params
X_{target}
tune_SI_hyperparams
$\lambda_{ridge}, \lambda_{svt}$
$X_{\neg target}$
SI
$X_{\neg target,c}$
$Y_{\neg target} = \begin{cases} get\_outcome(X_{\neg target,c}) - get\_outcome(X_{\neg target}), \text{ if } X_{\neg target} = X_{ctrl} \\ get\_outcome(X_{\neg target}) - get\_outcome(X_{\neg target,c}), \text{ if } X_{\neg target} = X_{treat} \end{cases}$

Fig. 2. Flowchart of the *estimate_ITEs* subroutine.

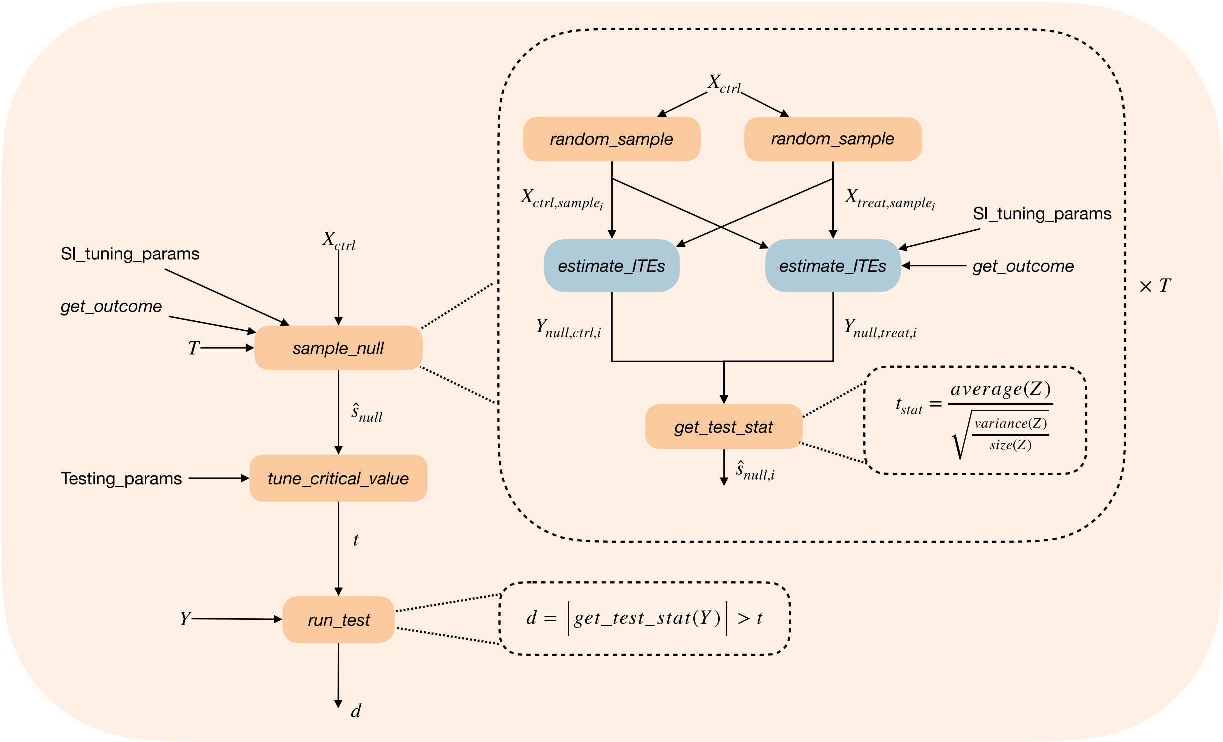

Fig. 3. Flowchart of the *run_hypothesis_test* routine. Note that both calls to *estimate_ITEs* in *sample_null* take in the SI_tuning_params and *get_outcome* function but we have omitted the arguments for brevity.

3.2. Baseline

Next, we describe the baseline method against which SECRETS is benchmarked. This method, which we refer to as *Standard*, is the approach used by two-arm, parallel-group, superiority RCT studies to determine a treatment's effect [1]. Its flowchart is shown in Fig. 4. First, given the control and treatment data, it calculates the corresponding outcome data, O_{ctrl} and O_{treat}, using *get_outcome* (vectorized). Then it conducts a two-sample t-test for independent samples with unequal variances [15] using the outcome data and desired significance level, α_{target}.

3.3. Ablation studies

In this section, we describe the ablation studies used to evaluate the efficacy of each step underlying SECRETS.

First, we run ablations on Estimate_Ites by swapping the counterfactual estimation algorithm of SECRETS, i.e., SI, with the Virtual

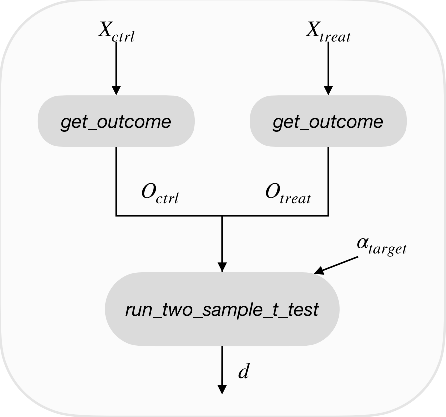

Fig. 4. Flowchart of the *Standard* baseline.

Twins (VT) algorithm [19], which estimates counterfactual outcomes using a regression model fit on covariates. We conduct this ablation to evaluate the advantage of SI, which learns patient-specific models rather than shared models. In particular, we fit two shared models, i.e., one model to predict responses under the control condition and one to predict responses under the treatment condition, using all the patient data (i.e., pre-intervention/baseline outcomes) from control and treatment groups, respectively, and use the models to predict the post-intervention counterfactual response of each patient under the treatment and control groups, respectively. As with SI, we use the ridge regression estimator and tune the regularization strength using the validation set performance. We refer to this version of the framework as SECRETS-VT.

Then, we run ablations on Run_Data_Driven_Hypothesis_Test by swapping the hypothesis testing algorithm of SECRETS with standard ones. We conduct this ablation to evaluate the need for a suitable hypothesis testing strategy, given the dependencies among the estimated ITEs. First, we try the one-sample t-test for the mean of a normal distribution [15], which we denote as SECRETS-T; this test assumes that the underlying samples (i.e., the estimate ITEs) are independent and identically distributed from a Gaussian distribution. Then, to assess if non-Gaussianity of the samples is an issue, we try the bootstrap hypothesis test, which we denote as SECRETS-B; this test still assumes that the samples are independent but relaxes the normality assumption, only requiring that samples be drawn from populations where the parameter being estimated is the same under the null hypothesis [20]. We specifically use the bias-corrected and accelerated version [21] since it yields better estimates of the parameter of interest [20]. Then, to assess whether data dependence across samples is an issue, we try a variant of SECRETS-T, called SECRETS-T-P, in which we first permute samples across the trials to remove dependencies and then run the one-sample t-test. After comparing against these standard test strategies, we assess how well the proposed hypothesis testing strategy performs by comparing against an “oracle”, which we denote as SECRETS-O. While SECRETS constructs the null distribution *per* trial (via the *sample_null* subroutine), SECRETS-O constructs the null distribution using the estimated ITEs across *all* the trials under the null setting, allowing it to identify the correct critical value that attains the significance level of α_{target}.

3.4. Clinical RCT datasets

We evaluate the framework on Phase-3 parallel-group RCTs as they typically require hundreds-to-thousands of subjects to establish a treatment's effect. We focus on the two-arm, superiority trial design typically adopted in clinical RCTs [1] to assess the general utility of our method. We obtained datasets for three such RCTs, i.e., CHAMP (NCT01581281) [22,23], ICARE (NCT00871715) [24,25], and MGTX (NCT00294658) [26], from the National Institute of Neurologic Disease and Stroke (NINDS) [27], which approved the use of these datasets; the datasets are described in Appendix C.

3.5. Implementation details

In this section, we specify the evaluation parameters used to perform our experiments. Algorithmic parameters are detailed in Appendix D.1 and computing configurations are specified in Appendix D.2.

For evaluation, we set α_{target} to 5%, and, for sample size requirements, we additionally set $1-\beta_{target}$ to 80%, since this statistical operating point is commonly used in clinical RCTs [1]. We set the number of trials L used to measure power $1-\beta$ and significance level α to 1000, since we empirically found this sufficient for the methods to converge close to α_{target} and $1-\beta_{target}$ within reasonable computation time. For the null setting, we set $\mu=\mu_0=0$, and for the alternative settings, we set $\mu=\mu_{1,target}$, with $\mu_{1,target}$ given by the ATE measured on each dataset, i.e., −3.17 for CHAMP, −3.00 for ICARE, and −2.70 for MGTX.

4. Results and discussion

This section presents the results from our experiments. First, we show that SECRETS outperforms the baseline method, *Standard*, by consistently reducing the variance of the ATE. Then, we demonstrate the importance of each component underlying SECRETS through ablation studies.

4.1. SECRETS vs. Baseline

Across all datasets, SECRETS obtains better power and comparable significance level compared to *Standard*, and the gains in power translate into reductions in the number of subjects required to obtain a desired power $1-\beta_{target}$ and significance level α_{target}. For example, on the CHAMP dataset (Table 1), for an arm size (n_a) of 550, *Standard* has a power $(1-\beta)$ of 53% and a significance level (α) of 5.0% while SECRETS has a power of 78% and a significance level of 5.1%. By increasing the power at a given arm size, SECRETS is able to converge close to the desired statistical operating point of 80% power $(1-\beta_{target})$ and 5% significance level (α_{target}) at an arm size of 550 while *Standard* does so at an arm size of 1130. Hence, SECRETS reduces the number of required samples by almost 51% or 580 subjects per arm.

On some datasets, *Standard* appears to have significance levels below the target 5% while SECRETS appears to have significance levels slightly above it (e.g., CHAMP with n_a of 1130 in Table 1), a performance gap which might explain why SECRETS has higher power than *Standard*. However, this gap in the significance levels is explained away by experimental limitations and is expected to vanish upon removing them while maintaining the power gains under SECRETS (see Appendix E).

To assess the factors contributing to the increased power under SECRETS, we analyze the distributions of the ATE estimates using the power estimation model (Eq. (2)), given that it estimates the measured power well, e.g., under an arm size of 550, the model estimates a power $(1-\hat{\beta})$ of 52% and 73% compared to the measured powers $(1-\beta)$ of 53% and 78% for *Standard* and SECRETS, respectively. First, we note that the means are further separated under *Standard*, i.e., the distance between the means ($|\hat{\mu}_1-\hat{\mu}_0|$) is 3.27 and 1.85 under *Standard* and SECRETS, respectively. Compared to SECRETS, *Standard* better separates the means because it has lower error in the ATE estimate under the alternative setting ($\hat{\mu}_1-\mu_{1,target}$) while both have comparably low error in the ATE estimate under the null setting ($\hat{\mu}_0-\mu_0$). Despite their enhanced mean separation under *Standard*, the distributions have lower variance under SECRETS than under *Standard*, i.e., the average

Table 1
Standard vs. SECRETS and SECRETS vs. SECRETS-VT.

Dataset	n_a	Method	$1-\beta$ (%)	α (%)	$1-\hat{\beta}$ (%)	$\lvert\hat{\mu}_1-\hat{\mu}_0\rvert$	$\hat{\mu}_1-\mu_{1,target}$	$\hat{\mu}_0-\mu_0$	$\hat{\sigma}$ (std.)	*Shift term*
CHAMP	550	*Standard*	53	5.0	52	3.27	−0.06	0.03	1.63 (0.01)	2.01
		SECRETS	78	5.1	73	1.85	1.31	−0.01	0.72 (0.04)	2.58
		SECRETS-VT	58	5.4	57	3.08	0.07	−0.02	1.43 (0.06)	2.15
	1130	*Standard*	81	3.8	84	3.20	−0.06	−0.03	1.09 (0.04)	2.93
		SECRETS	97	6.3	95	2.03	1.15	0.01	0.56 (0.04)	3.62
		SECRETS-VT	91	5.8	90	3.10	0.05	−0.02	0.95 (0.03)	3.25
ICARE	1250	*Standard*	27	4.1	30	3.02	−0.01	0.01	2.12 (0.03)	1.43
		SECRETS	81	5.5	81	3.37	−0.33	0.04	1.19 (0.03)	2.83
		SECRETS-VT	73	6.5	72	4.96	−1.93	0.04	1.94 (0.02)	2.56
	5207	*Standard*	81	5.8	81	2.98	−0.04	−0.06	1.05 (0.01)	2.83
		SECRETS	100	5.7	100	3.74	−0.77	−0.02	0.67 (0.07)	5.56
		SECRETS-VT	100	6.2	100	4.87	−1.91	−0.04	0.94 (0.01)	5.16
MGTX	30	*Standard*	70	5.1	63	2.69	−0.09	−0.10	1.17 (0.06)	2.30
		SECRETS	79	5.7	69	1.33	1.33	−0.04	0.54 (0.03)	2.45
		SECRETS-VT	71	5.1	77	2.49	0.17	−0.04	0.92 (0.01)	2.70
	40	*Standard*	81	4.8	75	2.67	−0.06	−0.09	1.01 (0.02)	2.63
		SECRETS	87	5.2	80	1.32	1.34	−0.04	0.47 (0.01)	2.80
		SECRETS-VT	85	5.5	87	2.45	0.20	−0.04	0.79 (0.02)	3.09

standard deviation ($\hat{\sigma}$) is 1.63 and 0.72 under *Standard* and SECRETS, respectively. The lower variance under SECRETS outweighs the advantage of better mean separation under *Standard*, giving SECRETS a higher *shift term* of 2.58 compared to 2.01 under *Standard*, which translates into higher power. Similar trends hold for an arm size of 1130, as well as on the ICARE and MGTX datasets (Table 1).

4.2. Ablation studies

Having demonstrated the ability of SECRETS to increase power and thereby reduce sampling complexity, we assess the importance of each design choice in the framework. First, we show that ITE estimation with SI outperforms ITE estimation with VT, by comparing the performance of SECRETS against SECRETS-VT. Then, we show that our data-driven hypothesis testing strategy is essential to the performance of SECRETS by comparing against alternative standard test strategies. We also show that our test strategy performs well, converging close to the performance obtained by an oracle version of the strategy.

4.2.1. Synthetic intervention vs. Virtual twins

Compared to SECRETS, SECRETS-VT has higher variance in the ATE estimate, giving it lower power but comparable significance levels, in general. For example, on the CHAMP dataset (Table 1), for an arm size of 550, SECRETS and SECRETS-VT both have significance levels near 5% but score 78% and 58% power, respectively. Applying the power equation model shows that, despite being better at separating the means, SECRETS-VT increases the variance of the ATE estimates, thereby lowering its *shift term* compared to that under SECRETS. Similar trends hold for the ICARE dataset (Table 1). Although the model is unsuitable for the MGTX dataset given its small sample size, SECRETS-VT still has higher variance than SECRETS, likely explaining the reduced power (Table 1). Our results attest that SI is better able to reduce the variance of the distributions compared to standard regression (i.e., the VT algorithm). SI's advantage stems from its nonparametric framing of the problem, in which it learns how to weight responses of donor units to predict the target unit's response, unlike the VT algorithm that predicts responses using parametric models based only on the baseline measurement.

4.2.2. Data-driven hypothesis testing vs. Alternative standard hypothesis testing strategies

Next, we show that our hypothesis testing strategy is essential for SECRETS to keep significance levels close to α_{target} by comparing against conventional testing strategies, with results reported in Table 2. SECRETS-T, which uses the one-sample t-test, achieves an average α of 1.8% across the datasets and arm sizes (n_a), which suggests that the test's assumptions, i.e., observations are drawn independently and identically from a Gaussian distribution, fail to hold. First, we check if Gaussianity is violated by swapping the one-sample t-test with the bootstrap hypothesis test, i.e., SECRETS-B. SECRETS-B obtains a slightly higher average α of 1.9%, implying that non-Gaussianity is not the problem. To show that the samples (ITEs) are dependent, we run a version of the one-sample t-test, i.e., SECRETS-T-P, in which the estimated ITEs are shuffled across the trials to remove the sample dependencies existing within a single trial. SECRETS-T-P obtains an average α of 5.5%, which is significantly closer to α_{target}; this result affirms that the sample dependency induced under SI is the violation incurred under conventional hypothesis tests. To address the sample dependency problem, SECRETS implements a hypothesis test that uses the data to construct a null distribution that captures the dependencies in the samples and thereby obtains an average α of 5.6%. Increasing T is expected to help significance levels of SECRETS converge to 5% by fitting more accurate null distributions. For example, on the CHAMP dataset with n_a of 1130, increasing T from 100 to 500 samples reduced α from 6.3% to 5.8%, and on the MGTX dataset with n_a of 40, increasing T from 100 to 1 K samples reduced α from 20.7% to 5.2%. In addition, the oracle version of SECRETS, i.e., SECRETS-O, which constructs the null distribution using ITEs across *all* the trials, obtains nearly 5% significance level across all datasets and arm sizes, suggesting that SECRETS can benefit from more diverse samples of ITEs under the null hypothesis, which may be obtained by increasing T.

5. Conclusions

In conclusion, we have developed SECRETS, a novel framework that, for the first time, *simulates* the cross-over design using only data from the RCT of interest in order to boost its power (i.e., subject efficiency). It does this by using SI to estimate ITEs per patient across both control and treatment groups in order to reduce the variance of the ATE estimate. Then, it implements a novel data-driven hypothesis testing strategy suitable for the estimated ITEs since their properties violate the independence assumption under conventional hypothesis testing schemes. Evaluated on three real-world Phase-3 clinical RCTs, i.e., the CHAMP, ICARE, and MGTX studies, SECRETS improves power over the baseline approach by 6%–54% with an average improvement of 21.5% (standard deviation of 15.8%) while maintaining comparable significance levels. In addition, the gains in power reduce the number of subjects required to obtain a typically desired statistical operating point of 80% power and 5% significance level by 51% or 580 subjects per arm on the CHAMP dataset, 76% or 3957 subjects per arm on

Table 2
Significance level, α (%), under different hypothesis testing strategies.

Dataset	n_a	SECRETS-T	SECRETS-B	SECRETS-T-P	SECRETS	SECRETS-O
CHAMP	550	2.2	2.5	6.8	5.1	5.0
	1130	5.3	5.6	5.4	6.3	5.0
ICARE	1250	0.0	0.0	4.3	5.5	5.0
	5207	3.5	3.5	5.8	5.7	5.0
MGTX	30	0.0	0.0	5.3	5.7	5.1
	40	0.0	0.0	5.3	5.2	5.1
Average	–	1.8	1.9	5.5	5.6	5.0

the ICARE dataset, and 25% or 10 subjects per arm on the MGTX dataset. Our empirical results demonstrate that SECRETS can feasibly increase success rates for any completed parallel-group Phase-3 RCT, in contrast to prior approaches that either require external data or impose conditions on the intervention-condition pairs, thereby saving millions of US dollars [5]. In addition, while we presented SECRETS in the context of a typical setup for a Phase-3 parallel-group RCT (i.e., a two-arm, superiority design [1,3]), SECRETS extends to multi-arm settings since they reduce to pairwise comparisons between select arms [28] and extends to equivalence or non-inferiority studies since they just require appropriate modification of the test design [29]. To further increase trial success rates, we plan to develop a scheme for sample size estimation under SECRETS as a study planning aide.

CRediT authorship contribution statement

Sayeri Lala: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Niraj K. Jha:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The RCT datasets used for evaluation can be obtained from NINDS [27] by filling out the “NINDS Data Request Form" (https://www.ninds.nih.gov/sites/default/files/migrate-documents/sig_form_revised_508c.pdf) per dataset and emailing it to NINDS (CRLiaison@ninds.nih.gov). Under the “Dataset Being Requested” section, provide the information per field as shown in Table F.1. Pseudocodes for algorithms referenced in Section 2 have been provided in Appendix A.

Acknowledgments

This work was performed using computing resources from Princeton Research Computing.

This research is based on data from NINDS obtained from its Archived Clinical Research Dataset website. The CHAMP dataset was obtained from the Childhood and Adolescent Migraine Prevention Study, conducted under principal investigators (PIs) Drs. Powers, Hershey, and Coffey, under Grant #1U01NS076788-01. The ICARE dataset was obtained from the Arm Rehabilitation Study After Stroke (ICARE), conducted under PIs Drs. Winstein, Dromerick, and Wolf, under Grant #U01NS056256. The MGTX dataset was obtained from the Thymectomy Trial in Non-Thymomatous Myasthenia Gravis Patients Receiving Prednisone Therapy, conducted under PIs Drs. Cutter, Wolfe, and Kaminski under Grant #1U01NS042685-01A2.

The authors thank Dr. Wolfe for his guidance on RCTs and the MGTX study. All authors approved the version of the manuscript to be published.

Funding source

This work was supported by the National Science Foundation under Grant No. CNS-1907381.

Appendix A. Secrets pseudocode

TUNE_SI_HYPERPARAMS (Alg. 2) first initializes the set of best hyperparameters (lines 3–5) and then sweeps over a wide range of candidate values for λ_{ridge} and λ_{svt}, tracking the validation performance under each hyperparameter configuration (lines 6–19), where the validation performance is the R^2 score between the true validation data and corresponding SI-derived predictions. To obtain the SI-derived predictions per unit, TUNE_SI_HYPERPARAMS runs SI (Alg. 3) with the given candidate hyperparameters and X_{train} as the donor data, and $X_{val}[i]$ as the target unit data (lines 9–11). After estimating the trajectories across the validation units, TUNE_SI_HYPERPARAMS calculates the validation performance; if the validation performance improves, it updates the best hyperparameter configuration (lines 12–17). It concludes by returning the best hyperparameter configuration over the search range.

SI (Alg. 3) estimates the counterfactual post-intervention trajectory for the target unit by denoising the donor data with singular value thresholding (lines 3–5), learning the donor weights with ridge regression on the pre-intervention (i.e., baseline) data (line 6), and using the learned donor weights to estimate the target unit’s post-intervention trajectory from the donors’ post-intervention data (line 9). Since SI is used to predict the post-intervention trajectory of the target unit, the unit’s “counterfactual” pre-intervention data is initialized to its original pre-intervention data (line 8).

TUNE_CRITICAL_VALUE (Alg. 4) searches over critical values to find the one yielding significance level params.α_{target}. First, it sweeps over a range of candidate critical values, $t_{candidates}$, defined by params.n_s, params.t_{lower}, and params.t_{upper}, and checks if any value obtains the desired significance level params.α_{target} (lines 9–18). For each t, it estimates the resulting significance level $\hat{\alpha}$ by calling GET_ALPHA (line 13), which calculates $\hat{\alpha}$ by evaluating the two-sided test with the candidate critical value t. GET_ALPHA (lines 1–8) runs the test on each sample from the null distribution s_{null} (line 4) and calculates the fraction of samples on which the test returns a reject (line 7). If the critical value yields $\hat{\alpha}$ close to params.α_{target} (i.e., based on the error tolerance params.$\delta_{\alpha_{target}}$), TUNE_CRITICAL_VALUE returns the critical value (lines 14–16); otherwise, it stores the corresponding significance level and resumes the sweep (line 17).

If none of the candidate values yield significance levels close to params.α_{target}, TUNE_CRITICAL_VALUE updates its search range in a fashion similar to binary search. First, it identifies the significance level of the

Algorithm 1 SECRETS

```
    Input:
    $X_{ctrl} \in \mathbb{R}^{n_a \times n_t}$: control group data, where $n_a$ is the arm size, $n_t$ is the duration of the study, where $t = 1$ is the pre-intervention timepoint and $t > 1$ is the post-intervention period
    $X_{treat} \in \mathbb{R}^{n_a \times n_t}$: treatment group data, where $n_a$ is the arm size, $n_t$ is the duration of the study, where $t = 1$ is the pre-intervention timepoint and $t > 1$ is the post-intervention period
    SI_tuning_params: dictionary containing the following arguments for SI tuning
        $r_{train,val}$: ratio of training to validation set size for tuning SI's hyperparameters
    $get\_outcome : \mathbb{R}^{n_t} \to \mathbb{R}$: function that calculates the outcome of interest from a patient's response trajectory under some intervention over the study duration $n_t$
    $T$: number of samples to generate from the null distribution
    Testing_params: dictionary containing the following arguments for tuning the test's critical value
        $\alpha_{target}$: target significance level
        $t_{lower}$: critical value search lowerbound; $\geq 0$ (because of two-sided testing)
        $t_{upper}$: critical value search upperbound; $\geq t_{lower}$
        $t_{limit,exp}$: critical value search limit expansion term
        $n_s$: number of candidate critical values to search over
        $\delta_{\alpha_{target}}$: significance level error tolerance
    Output:
    $d$: test outcome, 1 means reject and 0 means do not reject the null hypothesis
 1: $Y_{ctrl}$ = ESTIMATE_ITES($X_{ctrl}, X_{treat}$, SI_tuning_params, $get\_outcome$)
 2: $Y_{treat}$ = ESTIMATE_ITES($X_{treat}, X_{ctrl}$, SI_tuning_params, $get\_outcome$)
 3: $Y = concatenate(Y_{ctrl}, Y_{treat})$
 4: $d$ = RUN_HYPOTHESIS_TEST($X_{ctrl}, Y$, SI_tuning_params, $get\_outcome, T$,
                                   Testing_params)
 5: return $d$
```

Algorithm 2 TUNE_SI_HYPERPARAMS

```
    Input:
    $X \in \mathbb{R}^{n_a \times n_t}$: data from a group exposed to a single intervention, where $n_a$ is the number of subjects in the group, $n_t$ is the duration of the study, where $t = 1$ is the pre-intervention timepoint and $t > 1$ is the post-intervention period
    params: dictionary containing the following arguments for SI tuning
        $r_{train,val}$: ratio of the training to validation set size for tuning SI's hyperparameters
    Output:
    $\lambda_{ridge,best}$: tuned ridge regularization strength
    $\lambda_{svt,best}$: tuned truncation threshold for singular value thresholding
 1: $X_{train}, X_{val} = split\_train\_val(X, \text{params}.r_{train,val})$
 2: $size, n_t = shape(X_{val})$
 3: $best\_score = -\infty$
 4: $\lambda_{ridge,best}$ = NIL
 5: $\lambda_{svt,best}$ = NIL
 6: for $\lambda_{ridge} \in logspace(-3, 3, 7)$ do
 7:     for $\lambda_{svt} \in linspace(0.1, 1, 10)$ do
 8:         $X_{val,pred} = \mathbf{0}^{size \times n_t}$
 9:         for $i = 1, size$ do
10:             $X_{val,pred}[i]$=SI($\lambda_{ridge}$,$\lambda_{svt}$,$X_{train}$,$X_{val}[i]$)
11:         end for
12:         $score = get\_rsquared(X_{val}, X_{val,pred})$
13:         if $score > best\_score$ then
14:             $best\_score = score$
15:             $\lambda_{ridge,best} = \lambda_{ridge}$
16:             $\lambda_{svt,best} = \lambda_{svt}$
17:         end if
18:     end for
19: end for
20: return $\lambda_{ridge,best}, \lambda_{svt,best}$
```

middle candidate critical value α_m (lines 19–20). If params.α_{target} is less than α_m, TUNE_CRITICAL_VALUE updates its search range to the right half of the original critical value search space since increasing the critical value would decrease the significance level (lines 21–22). In addition, if params.α_{target} is less than the $\hat{\alpha}$ associated with params.t_{upper}, params.t_{upper} needs to be sufficiently increased (i.e., by params.$t_{limit,exp}$) to ensure the updated search range contains a critical value yielding α_{target} (lines 23–24). Analogously, if params.α_{target} is greater than α_m, TUNE_CRITICAL_VALUE updates its search range to the left half of the original critical value search space since decreasing the critical value would increase the significance level (lines 26–27). In addition, if params.α_{target} is greater than $\hat{\alpha}$ associated with params.t_{lower}, params.t_{lower} needs to

Algorithm 3 SI

```
Input:
$\lambda_{ridge}$: ridge regularization strength
$\lambda_{svt}$: truncation threshold for singular value thresholding
$X_{donor} \in \mathbb{R}^{n_d \times n_t}$: donor data, where $n_d$ is the number of donor subjects, $n_t$ is the duration of the study, where $t = 1$ is the pre-intervention timepoint and $t > 1$ is the post-intervention period
$X_{unit} \in \mathbb{R}^{n_t}$: target unit data, where $n_t$ is the duration of the study, where $t = 1$ is the pre-intervention timepoint and $t > 1$ is the post-intervention period
Output:
$X_{unit,c} \in \mathbb{R}^{n_t}$: target unit's counterfactual data (under the intervention assigned to the donor group) over the study duration
$n_d, n_t = shape(X_{donor})$
$X_{unit} = reshape(X_{unit}, (1, n_t))$
$U, s, V^T = \text{SVD}(X_{donor})$
$S = \{i : s[i] \geq \lambda_{svt}\}$
$X_{donor,trunc} = \sum_{i \in S} s[i]U[i]V[i]^T$
$w = \text{argmin}_w \|X_{unit}[:,1] - w^T X_{donor,trunc}[:,1]\|_2^2 + \lambda_{ridge}\|w\|_2^2$
$X_{unit,c} = \mathbf{0}^{n_t}$
$X_{unit,c}[1] = squeeze(X_{unit})[1]$
$X_{unit,c}[2:] = squeeze(w^T X_{donor,trunc}[:, 2:])$
return $X_{unit,c}$
```

be sufficiently decreased (i.e., by params.$t_{limit,exp}$ but lowerbounded by 0 because of two-sided testing) to ensure the updated search range contains a critical value yielding α_{target} (lines 28–29). After updating the search range, TUNE_CRITICAL_VALUE continues searching by recursing with the updated search parameters and returning the identified critical value (line 32).

Appendix B. Metrics

B.1. Power and significance level

First, we define the target ATE under the alternative setting, i.e., $\mu = \mu_{1,target}$, as the ATE measured on the full RCT dataset, and define $\mu = \mu_0 = 0$ as the ATE under the null setting. Then, to measure power, $1-\beta$, given a sample of size n_a subjects per arm, we run L trials; we simulate a trial under the alternative setting by constructing new control and treatment groups by sampling n_a subjects with replacement from the original RCT's control and treatment groups, respectively. Likewise, we measure the significance level α with L trials by simulating the null setting, in which we construct both the control and treatment groups by sampling n_a subjects with replacement from the original RCT's control group.

B.2. Power formula evaluation

To evaluate the power formula (Eq. (2)), we set $\hat{\mu}_0$ and $\hat{\mu}_1$ as the mean of the distribution of the ATE estimate (derived from the trials) under the null and alternative settings, respectively. We estimate $\hat{\sigma}$ by averaging over the standard deviations of the distribution of the ATE estimate under the null and alternative settings and also report the standard deviation in this estimate (std.) to show that the standard deviations of the distributions are comparable. To assess whether the power formula is a good model of performance, we report the model's estimate of power, $1 - \hat{\beta}$, given α_{target}, $\hat{\mu}_0$, $\hat{\mu}_1$, and $\hat{\sigma}$. While the equation may not predict measured power, $1 - \beta$, exactly because of non-normality of the ATE estimate under a small sample size and differences in the test procedure, we can still use it to compare methods and qualitatively assess how differences in distributions explain differences in power.

We also assess the quality of the estimated ATEs since they may be of interest [1]. We calculate the errors in the ATE estimates under the null and alternative settings by measuring the difference between the means of the distribution of the estimated ATEs from their corresponding true values, given by $\hat{\mu}_0 - \mu_0$ and $\hat{\mu}_1 - \mu_{1,target}$, respectively.

Appendix C. Clinical RCT datasets

C.1. CHAMP

The CHAMP study [22] evaluated whether different medications could reduce headache frequency and heachache effects among children and adolescents suffering from migraines. To assess this, the study implemented an RCT containing a placebo group and two treatment groups receiving amitriptyline and topiramate, respectively. At the end of the trial, treatment effects were measured on various outcomes, including change in headache frequency, number of headache days, and headache-related disability scores, all relative to baseline measurements. Based on data from 328 subjects collected over 24 weeks, the study did not find any clinically significant between-group differences across the health outcomes.

For our experiments, we use the change in the Pediatric Migraine Disability Assessment (PedMIDAS) score as the outcome metric defining the ATE and the amitriptyline and topiramate as the control and treatment groups, respectively, because the corresponding ATE was more statistically significant among the ATEs defined by other continuous metrics and arm pairs [22,23]. To calculate this ATE, the study analyzed a subset of subjects monitored over two visits, i.e., one at baseline and another near the 24-week endpoint, which comprised 211 subjects with 107 and 104 subjects in the amitriptyline and topiramate groups, respectively. After applying the same criterion, we extracted 204 subjects from the original dataset, with 106 in the amitriptyline group and 98 in the topiramate group. From our extracted dataset, we calculated the ATE to be −3.17 units, somewhat comparable to the −4.3 units reported in the study [23].

C.2. ICARE

The ICARE study [24] evaluated whether a new motor training program (Accelerated Skill Acquisition Program or ASAP) could reduce upper extremity disability among patients with motor stroke more effectively than usual customary care (UCC). To assess this, the study implemented an RCT containing a treatment group exposed to ASAP, a control group exposed to dose-equivalent usual customary care (DEUCC), and another control group exposed to UCC with no constraint on the dose. At the end of the trial, treatment effects were assessed over various outcomes, including changes in Wolf Motor Function Test (WMFT) time, Stroke Impact Scale (SIS) scores, and arm muscle torque. Based on data from 304 patients collected over one year, the study did

Algorithm 4 TUNE_CRITICAL_VALUE

```
Input:
$s_{null} \in \mathbb{R}^T$: $T$ samples from the null distribution of the test statistic
params: dictionary containing the following arguments for tuning the test's critical value
    $\alpha_{target}$: target significance level
    $t_{lower}$: critical value search lowerbound; $\geq 0$ (because of two-sided testing)
    $t_{upper}$: critical value search upperbound; $\geq t_{lower}$
    $t_{limit,exp}$: critical value search limit expansion term
    $n_s$: number of candidate critical values to search over
    $\delta_{\alpha_{target}}$: significance level error tolerance
Output:
$d$: test outcome, 1 means reject and 0 means do not reject the null hypothesis
procedure GET_ALPHA($s_{null}, t$)
    $\alpha = 0$
    for $i = 1, length(s_{null})$ do
        $r = \left|s_{null}[i]\right| > t$
        $\alpha = \alpha + r$
    end for
    return $\alpha / length(s_{null})$
end procedure
$t_{candidates} = linspace(\text{params}.t_{lower}, \text{params}.t_{upper}, \text{params}.n_s)$
$\hat{\alpha}\_by\_t = []$
for $i = 1, \text{params}.n_s$ do
    $t = t_{candidates}[i]$
    $\hat{\alpha}$=GET_ALPHA($s_{null}, t$)
    if $|\hat{\alpha} - \text{params}.\alpha_{target}| < \text{params}.\delta_{\alpha_{target}}$ then
        return $t$
    end if
    $\hat{\alpha}\_by\_t[i] = \hat{\alpha}$
end for
$m = \text{params}.n_s/2$
$\alpha_m = \hat{\alpha}\_by\_t[m]$
if params.$\alpha_{target} < \alpha_m$ then
    params.$t_{lower} = t_{candidates}[m+1]$
    if params.$\alpha_{target} < \hat{\alpha}\_by\_t[-1]$ then
        params.$t_{upper}$ = params.$t_{upper}$ + params.$t_{limit,exp}$
    end if
else
    params.$t_{upper} = t_{candidates}[m+1]$
    if params.$\alpha_{target} > \hat{\alpha}\_by\_t[0]$ then
        params.$t_{lower} = max(0, \text{params}.t_{lower} - \text{params}.t_{limit,exp})$
    end if
end if
return TUNE_CRITICAL_VALUE($s_{null}$, $params$)
```

Table F.1
Effect of number of trials (L) on *Standard*'s performance.

Dataset	L = 1K		L = 10K	
	$1-\beta$ (%)	α (%)	$1-\beta$ (%)	α (%)
CHAMP, $n_a = 1130$	81	3.8	81	4.5
ICARE, $n_a = 1250$	27	4.1	28	4.9
ICARE, $n_a = 5207$	81	5.8	80	5.5

not find any clinically significant between-group differences in a subset of these scores, i.e., WMFT and SIS hand function score [24,25].

For our experiments, we use the change in arm muscle torque based on shoulder flexors as the outcome metric defining the ATE, and the DEUCC and ASAP as the control and treatment groups, respectively, to speed up experiment time since detecting this ATE with high power and low significance level required a relatively small sample size for the baseline method. After applying the study's data processing protocol to the original dataset, we extracted data from 183 subjects, with 93 in the control group and 90 in the treatment group, from visits at the baseline and one-year endpoints (the number of subjects analyzed per group is equal to that reported in the study's analysis). From our extracted dataset, we calculated the ATE to be −3.00 units, close to the −2.99 reported in the study [25].

C.3. MGTX

The MGTX study [26] investigated whether thymectomy combined with standard prednisone therapy could treat Myasthenia Gravis more effectively than prednisone therapy alone. To assess this, the study implemented an RCT, in which the control arm received prednisone therapy over three years and the treatment arm underwent thymectomy and received the same prednisone therapy; at the end of the trial, treatment effects were measured with respect to the Quantitative Myasthenia Gravis (QMG) total score and required prednisone dose, both averaged over the study period. Based on data from 126 subjects collected over three years, the study found that thymectomy improved health outcomes with clinical significance; the time-weighted average

Table F.2
Information to provide in the "NINDS Data Request Form" under the "Dataset Being Requested" section.

Dataset/Fields	Trial acronym	ClinicalTrials.gov NCT #	Trial title	Trial PI name
CHAMP	CHAMP	01581281	The Childhood and Adolescent Migraine Prevention Study (CHAMP)	Scott W. Powers, PhD
ICARE	ICARE	00871715	Arm Rehabilitation Study After Stroke (ICARE)	Carolee J. Winstein, PhD
MGTX	MGTX	00294658	Thymectomy Trial in Non-Thymomatous Myasthenia Gravis Patients Receiving Prednisone Therapy	Dr. Gary Cutter

QMG scores decreased by an average of 2.85 units and the time-weighted average prednisone dose also decreased by an average of 22 mg.

For our experiments, we use the time-weighted average QMG total score as the outcome metric defining the ATE, because it was easier to reproduce [26]. The study calculated the ATE by analyzing a subset of the population monitored over the three-year window (14 patient visits or timepoints), which comprised 62 subjects and 56 subjects in the treatment and control groups, respectively. We followed the study's data processing protocol and from the original dataset, we extracted a dataset with 49 subjects in the treatment group and 47 subjects in the control group. From our extracted dataset, we calculated the ATE to be −2.70 units, comparable to the −2.85 units reported in the study.

Appendix D. Implementation details

D.1. Algorithmic parameters

We define the *get_outcome* function, used by *Standard* and SECRETS, for each RCT dataset according to the corresponding outcome metric specified in Appendix C. For *Standard*, we set the target significance level $\alpha_{target} = 5\%$. For SECRETS, we set the parameters contained in SI_tuning_params as follows: $r_{train,val}$ to 7/3. We set T to 100 for CHAMP and ICARE and 1000 for MGTX. We set the parameters contained in Testing_params as follows: $\alpha_{target} = 5\%$, $t_{lower} = 3$, $t_{upper} = 5$, $t_{limit,exp} = 2$, $n_s = 10$, and $\delta_{\alpha_{target}} = 1e\text{–}3$. We found this parameter configuration enabled SECRETS to achieve significance level close to α_{target}. See Appendix A, Alg. 1 for descriptions of these parameters.

D.2. Computing setup

To run our experiments, we used 28–32 CPU cores, 4 GB of memory per CPU, and 2.4 GHz Intel Broadwell and 2.6 GHz Intel Skylake processors. We implemented the framework and experiments with Python using standard numerical packages.

Appendix E. Results

We explain how experimental limitations prevented significance levels under *Standard* and SECRETS from reaching the target value (5%) on some datasets. Specifically, increasing the number of trials L enables *Standard* to converge to the 5% significance level while maintaining comparable power (see Table F.1). In addition, increasing T or the number of samples generated under the null distribution under SECRETS enables it to converge to 5% significance level while maintaining comparable power. For example, on the CHAMP dataset with n_a of 1130, increasing T from 100 to 500 samples reduced α from 6.3% to 5.8% while preserving power, and on the MGTX dataset with n_a of 40, increasing T from 100 to 1 K samples reduced α from 20.7% to 5.2% while power decreased from 89.7% to 87.4%. Therefore, since the significance levels of *Standard* are actually near 5% (based on more trials) and those of SECRETS converge to 5% while maintaining power (by increasing T), SECRETS is expected to outperform *Standard* in these cases.

Appendix F. Tables

See Tables F.1 and F.2.

References

[1] L.M. Friedman, C.D. Furberg, D.L. DeMets, D.M. Reboussin, C.B. Granger, Fundamentals of Clinical Trials, fifth ed., Springer International Publishing, Switzerland, 2015.

[2] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, A. Zhang, A survey on causal inference, ACM Trans. Knowl. Discov. Data 15 (2021) 1–46, http://dx.doi.org/10.1145/3444944.

[3] K. Stanley, Design of randomized controlled trials, Circulation 115 (2007) 1164–1169, http://dx.doi.org/10.1161/CIRCULATIONAHA.105.594945.

[4] FDA drug approval process, 2023, https://www.fda.gov/media/82381/download, Accessed: 2023-03-01.

[5] L. Martin, M. Hutchens, C. Hawkins, A. Radnov, How much do clinical trials cost, Nat. Rev. Drug Discov. 16 (2017) 381–382, http://dx.doi.org/10.1038/nrd.2017.70.

[6] T.J. Hwang, D. Carpenter, J.C. Lauffenburger, B. Wang, J.M. Franklin, A.S. Kesselheim, Failure of investigational drugs in late-stage clinical development and publication of trial results, JAMA Intern. Med. 176 (2016) 1826–1833, http://dx.doi.org/10.1001/jamainternmed.2016.6008.

[7] K. Thorlund, L. Dron, J.J. Park, E.J. Mills, Synthetic and external controls in clinical trials–a primer for researchers, Clin. Epidemiol. 12 (2020) 457–467, http://dx.doi.org/10.2147/CLEP.S242097.

[8] V. Prasad, Reliable, cheap, fast and few: What is the best study for assessing medical practices? Randomized controlled trials or synthetic control arms? Eur. J. Clin. Investig. 51 (2021) http://dx.doi.org/10.1111/eci.13580.

[9] J. Lim, R. Walley, J. Yuan, J. Liu, A. Dabral, N. Best, et al., Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities, Ther. Innov. Regul. Sci. 52 (2018) 546–559, http://dx.doi.org/10.1177/2168479018778282.

[10] I. Bica, A.M. Alaa, C. Lambert, M. Van Der Schaar, From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges, Clin. Pharmacol. Ther. 109 (2021) 87–100, http://dx.doi.org/10.1002/cpt.1907.

[11] Z. Qian, Y. Zhang, I. Bica, A. Wood, M. van der Schaar, SyncTwin: Treatment effect estimation with longitudinal outcomes, in: Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 3178–3190.

[12] S.S. Senn, Cross-over Trials in Clinical Research, second ed., John Wiley & Sons, England, 2002.

[13] J.W. Blackston, A.G. Chapple, J.M. McGree, S. McDonald, J. Nikles, Comparison of aggregated N-of-1 trials with parallel and crossover randomized controlled trials using simulation studies, Healthcare 7 (2019) 137, http://dx.doi.org/10.3390/healthcare7040137.

[14] A. Agarwal, D. Shah, D. Shen, Synthetic A/B testing using synthetic interventions, 2023, Preprint at arXiv:2006.07691v5.

[15] B. Rosner, Fundamentals of Biostatistics, eighth ed., Cengage Learning, Boston, MA, 2015.

[16] M. Amjad, D. Shah, D. Shen, Robust synthetic control, J. Mach. Learn. Res. 19 (2018) 802–852.

[17] M.R. Dale, M.-J. Fortin, Spatial autocorrelation and statistical tests in ecology, Ecoscience 9 (2002) 162–167, http://dx.doi.org/10.1080/11956860.2002.11682702.

[18] Z. Liu, F. Peng, Statistical testing on ASR performance via blockwise bootstrap, in: INTERSPEECH, Vol. 34, ISCA, 2020, pp. 596–600.

[19] J.C. Foster, J.M. Taylor, S.J. Ruberg, Subgroup identification from randomized clinical trial data, Stat. Med. 30 (2011) 2867–2880, http://dx.doi.org/10.1002/sim.4322.

[20] P. Good, Permutation, Parametric, and Bootstrap Tests of Hypotheses, third ed., Springer Science + Business Media, Inc., New York, NY, 2005.

[21] B. Efron, Better bootstrap confidence intervals, J. Amer. Statist. Assoc. 82 (1987) 171–185, http://dx.doi.org/10.2307/2289144.

[22] S.W. Powers, C.S. Coffey, L.A. Chamberlin, D.J. Ecklund, E.A. Klingner, J.W. Yankey, et al., Trial of amitriptyline, topiramate, and placebo for pediatric migraine, N. Engl. J. of Med. 376 (2017) 115–124, http://dx.doi.org/10.1056/NEJMoa1610384.

[23] The childhood and adolescent migraine prevention study (CHAMP), 2023, https://clinicaltrials.gov/ct2/show/results/NCT01581281, Accessed: 2023-01-01.

[24] C.J. Winstein, S.L. Wolf, A.W. Dromerick, C.J. Lane, M.A. Nelsen, R. Lewthwaite, et al., Effect of a task-oriented rehabilitation program on upper extremity recovery following motor stroke: The ICARE randomized clinical trial, JAMA 315 (2016) 571–581, http://dx.doi.org/10.1001/jama.2016.0276.

[25] Arm rehabilitation study after stroke ICARE, 2023, https://clinicaltrials.gov/ct2/show/results/NCT00871715, Accessed: 2023-01-01.

[26] G.I. Wolfe, H.J. Kaminski, I.B. Aban, G. Minisman, H.-C. Kuo, A. Marx, et al., Randomized trial of thymectomy in myasthenia gravis, N. Engl. J. Med. 375 (2016) 511–522, http://dx.doi.org/10.1056/NEJMoa1602489.

[27] Archived clinical research datasets, 2023, https://www.ninds.nih.gov/current-research/research-funded-ninds/clinical-research/archived-clinical-research-datasets, Accessed: 2023-10-26.

[28] E. Juszczak, D.G. Altman, S. Hopewell, K. Schulz, Reporting of multi-arm parallel-group randomized trials: Extension of the CONSORT 2010 statement, JAMA 321 (2019) 1610–1620, http://dx.doi.org/10.1001/jama.2019.3087.

[29] S.-C. Chow, J. Shao, H. Wang, A note on sample size calculation for mean comparisons based on noncentral t-statistics, J. Biopharm. Stat. 12 (2002) 441–456, http://dx.doi.org/10.1081/BIP-120016229.