compute. collaborate. create.

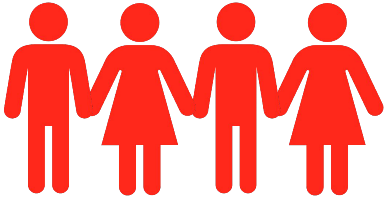
Falsification before Extrapolation in Causal Effect Estimation

Zeshan Hussain*, Michael Oberst*, Ming-Chieh Shih*, David Sontag

*equal contribution, by alphabetical order

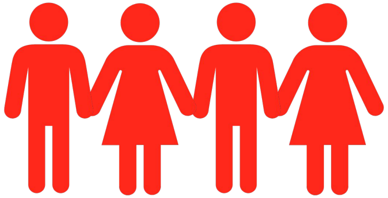
Motivating Example

Randomized Controlled Trial **(RCT)**



Motivating Example

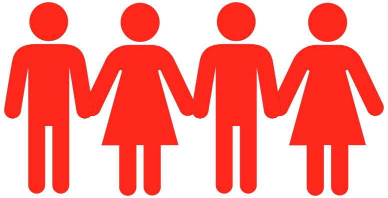
Randomized Controlled Trial **(RCT)**



RCTs often fail to include all types of patients (e.g.)

Motivating Example

Randomized Controlled Trial **(RCT)**



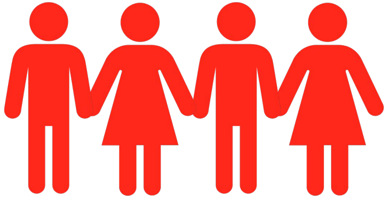
Real-world example: pregnant women were not included in initial COVID-19 trials[1]

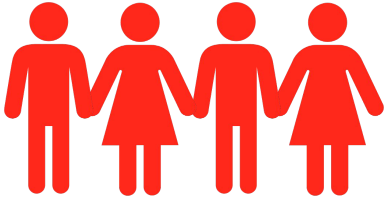
1] Dagan, Noa, et al. "Effectiveness of the BNT162b2 mRNA COVID-19 vaccine in pregnancy." *Nature medicine* 27.10 (2021): 1693-1695.

Motivating Example

Randomized Controlled Trial **(RCT)**

Observational Study **(OS)**





Observational studies contain a more **diverse cohort**, but may suffer from e.g. **unobserved confounding**.

Motivating Example

RCT

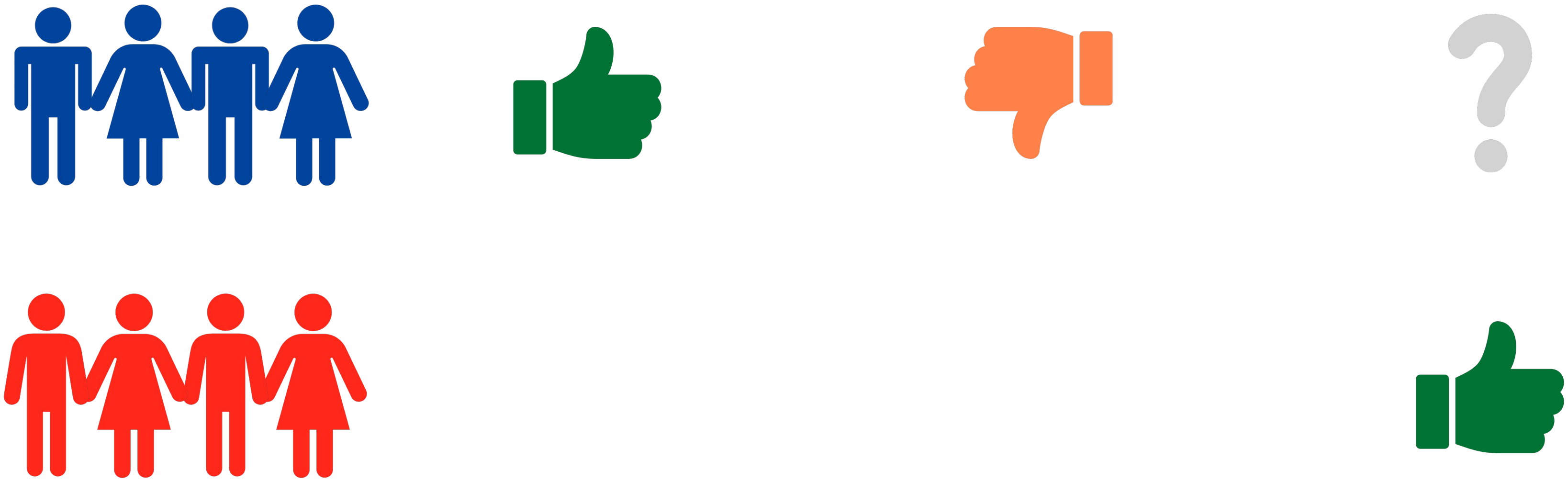
?

Motivating Example

Observational Study #1	Observational Study #2	RCT

Motivating Example

Observational Study #1	Observational Study #2	RCT

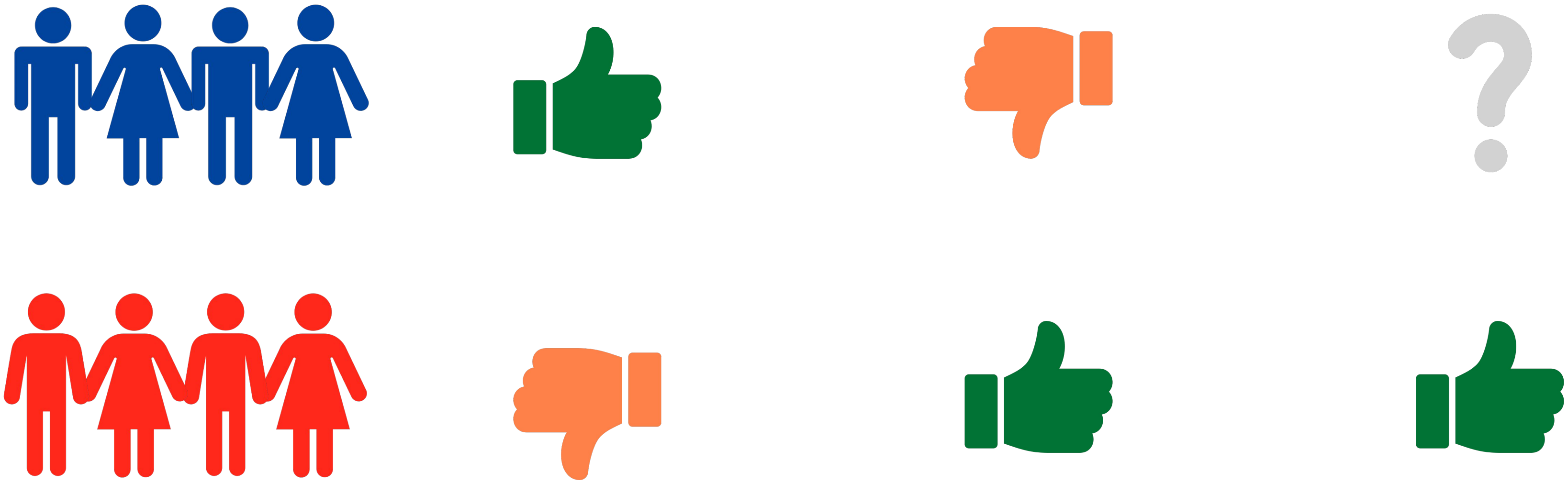


Motivating Example

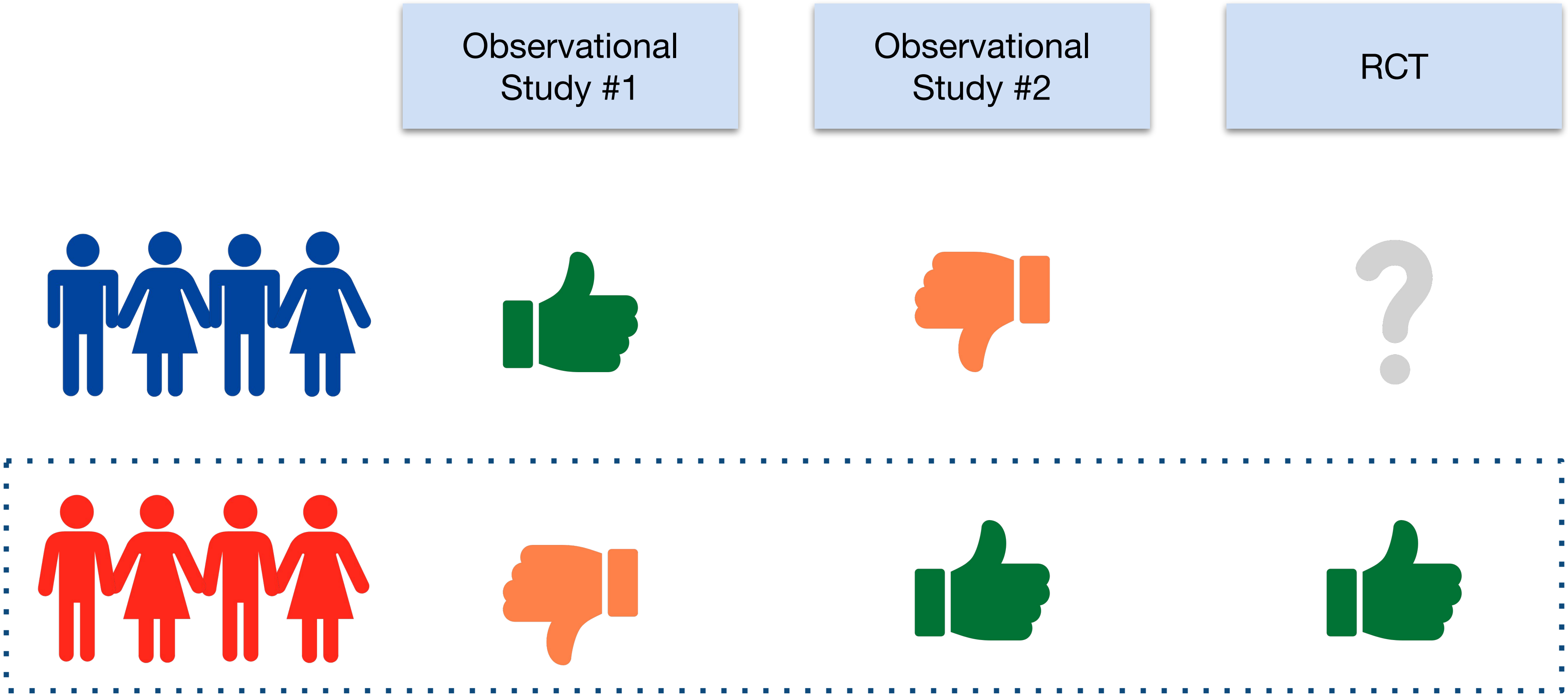
Observational Study #1

Observational Study #2

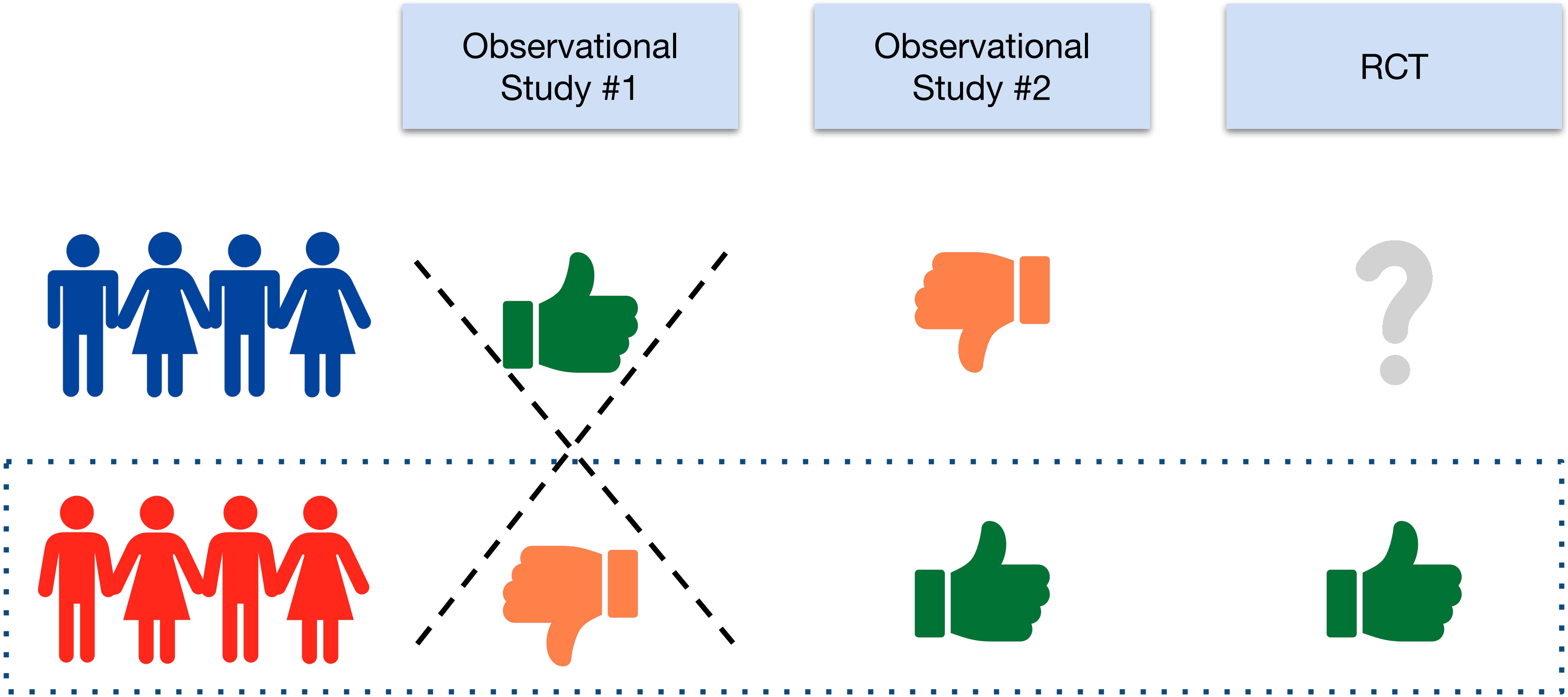
RCT



Motivating Example

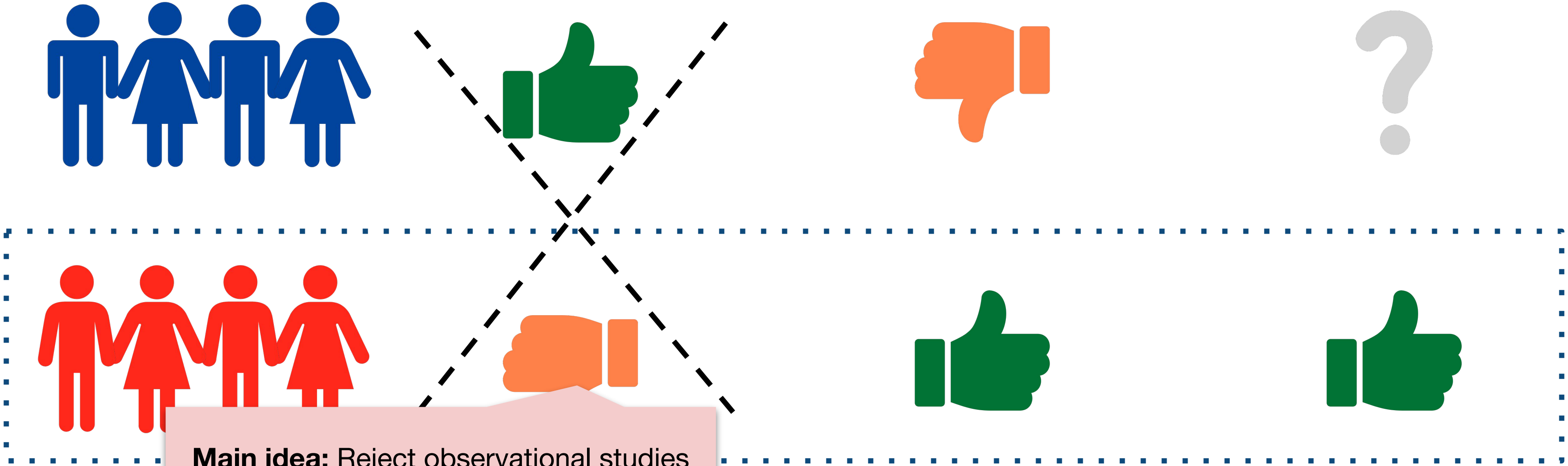


Motivating Example



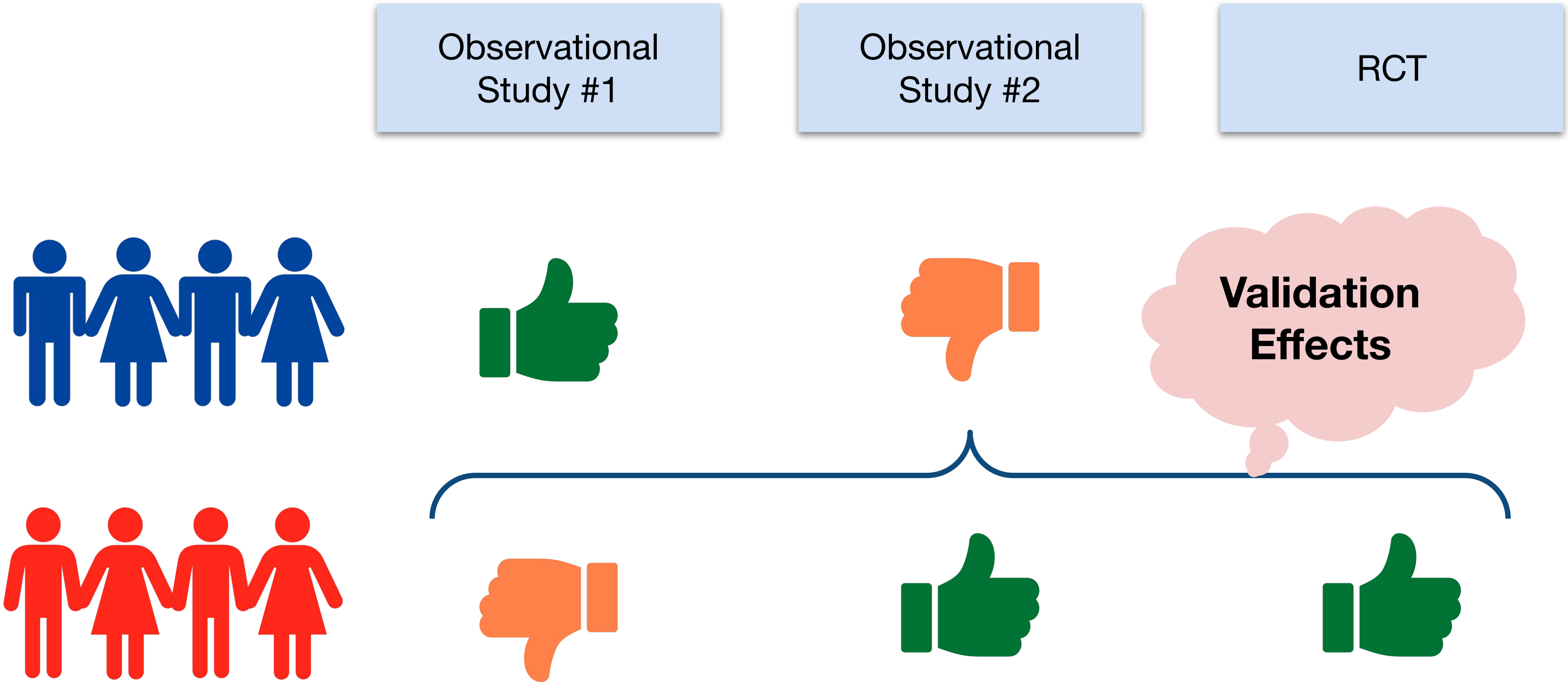
Motivating Example

Observational Study #1	Observational Study #2	RCT

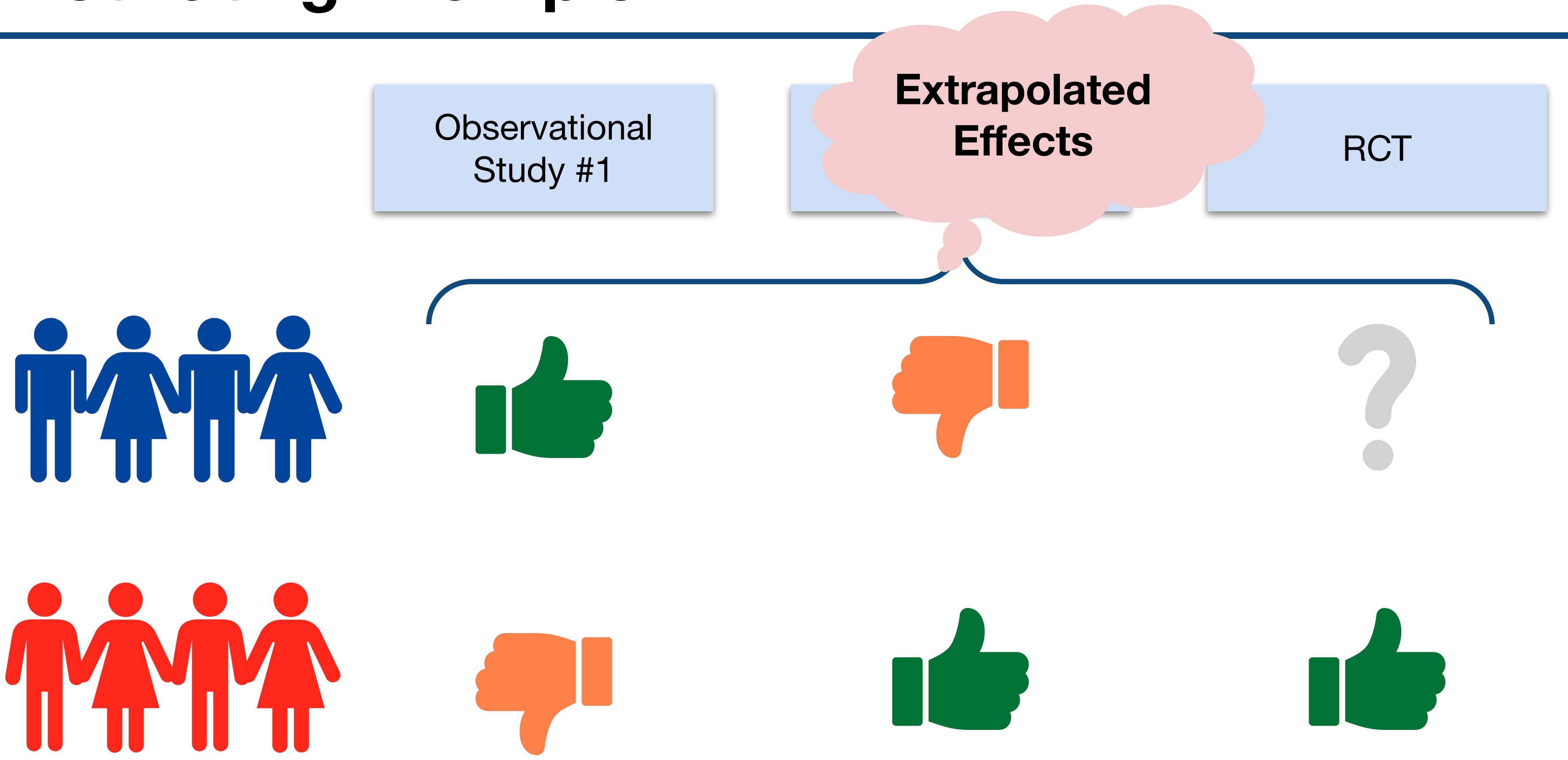


Main idea: Reject observational studies that fail to replicate RCT results

Motivating Example



Motivating Example



Contributions

Our Approach

1

Falsification of observational estimates

Contributions

Our Approach

1

Falsification of observational estimates

Use framework of **hypothesis testing**

Contributions

Our Approach

1

Falsification of observational estimates

Use framework of **hypothesis testing**

Reject estimators that do not replicate **RCT estimates**

Contributions

Our Approach

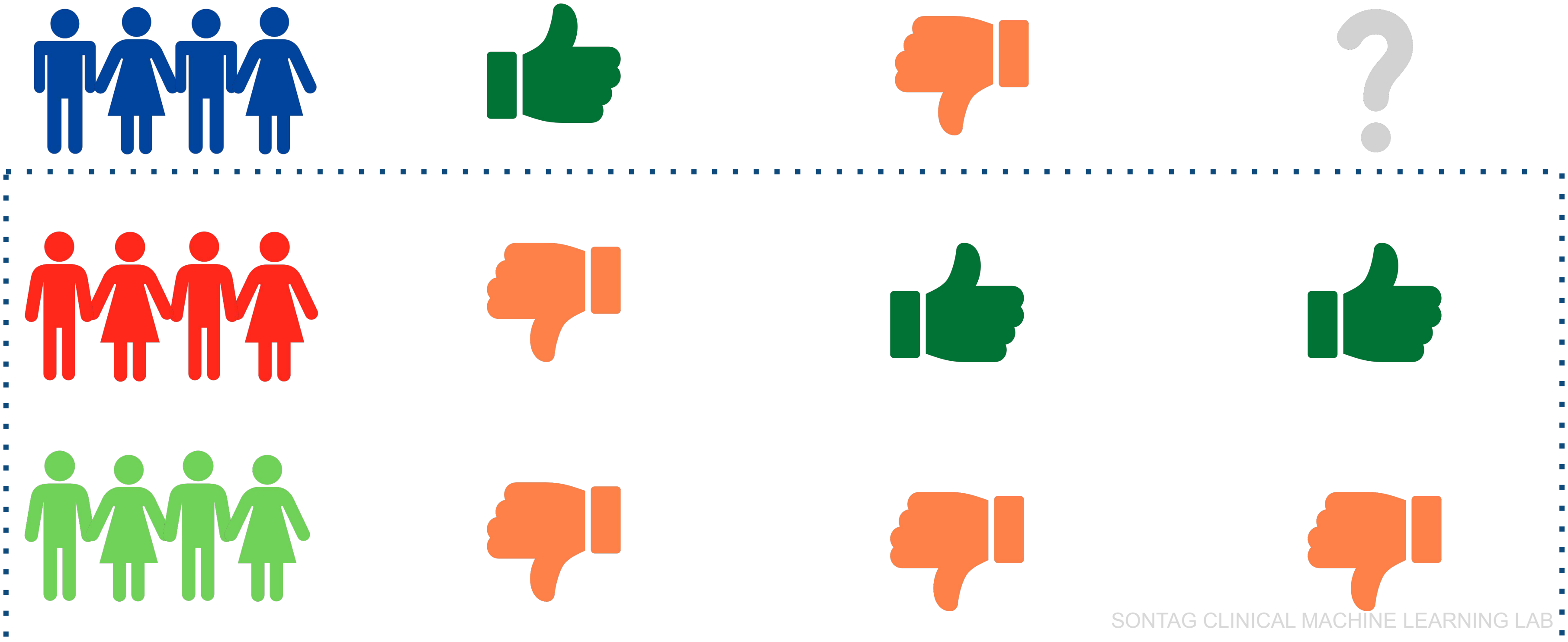
1 Falsification of observational estimates

2 Pessimistic Combination of Confidence Intervals

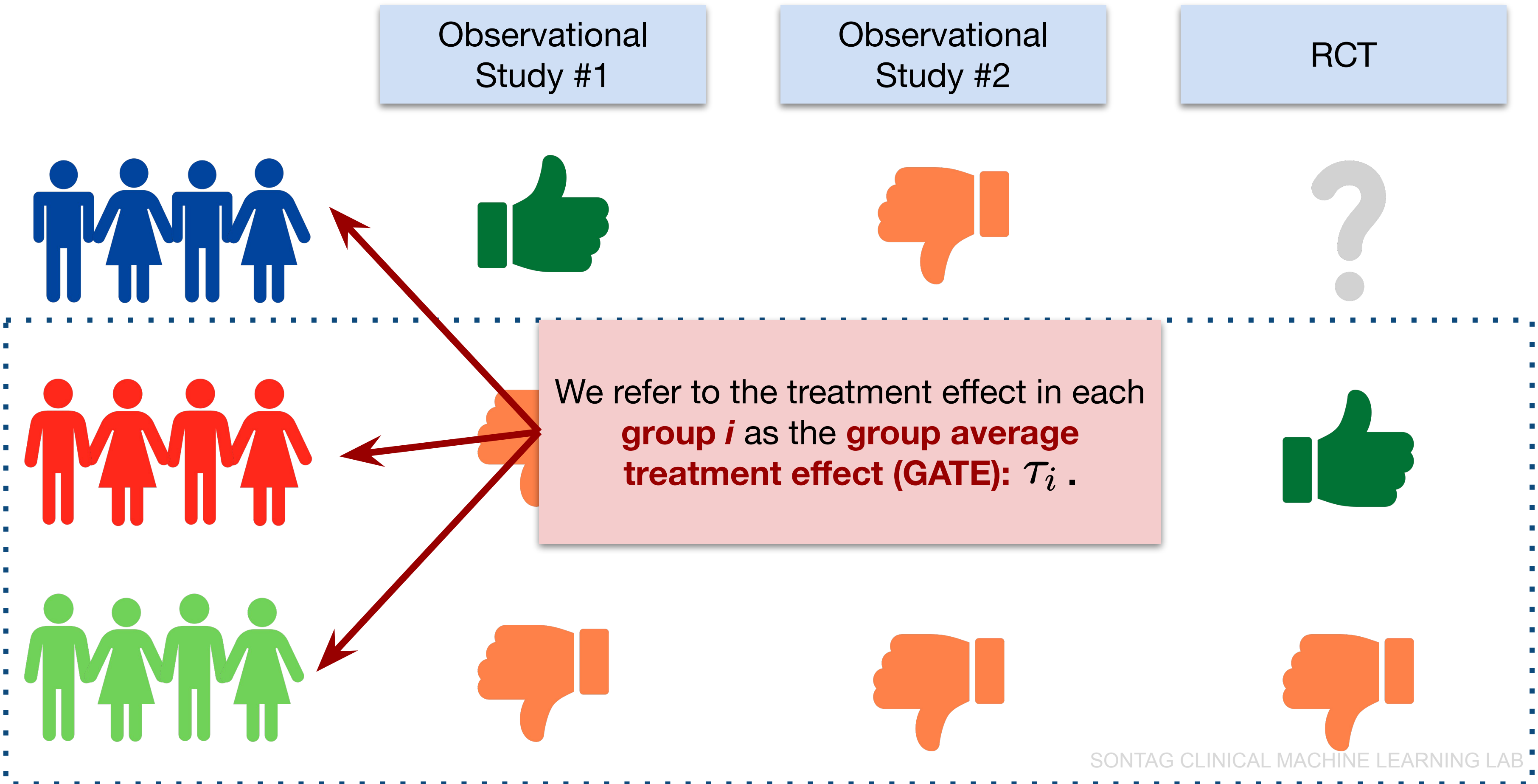
Take the **union** over all the intervals of the **accepted estimators**.

Formalizing Falsification

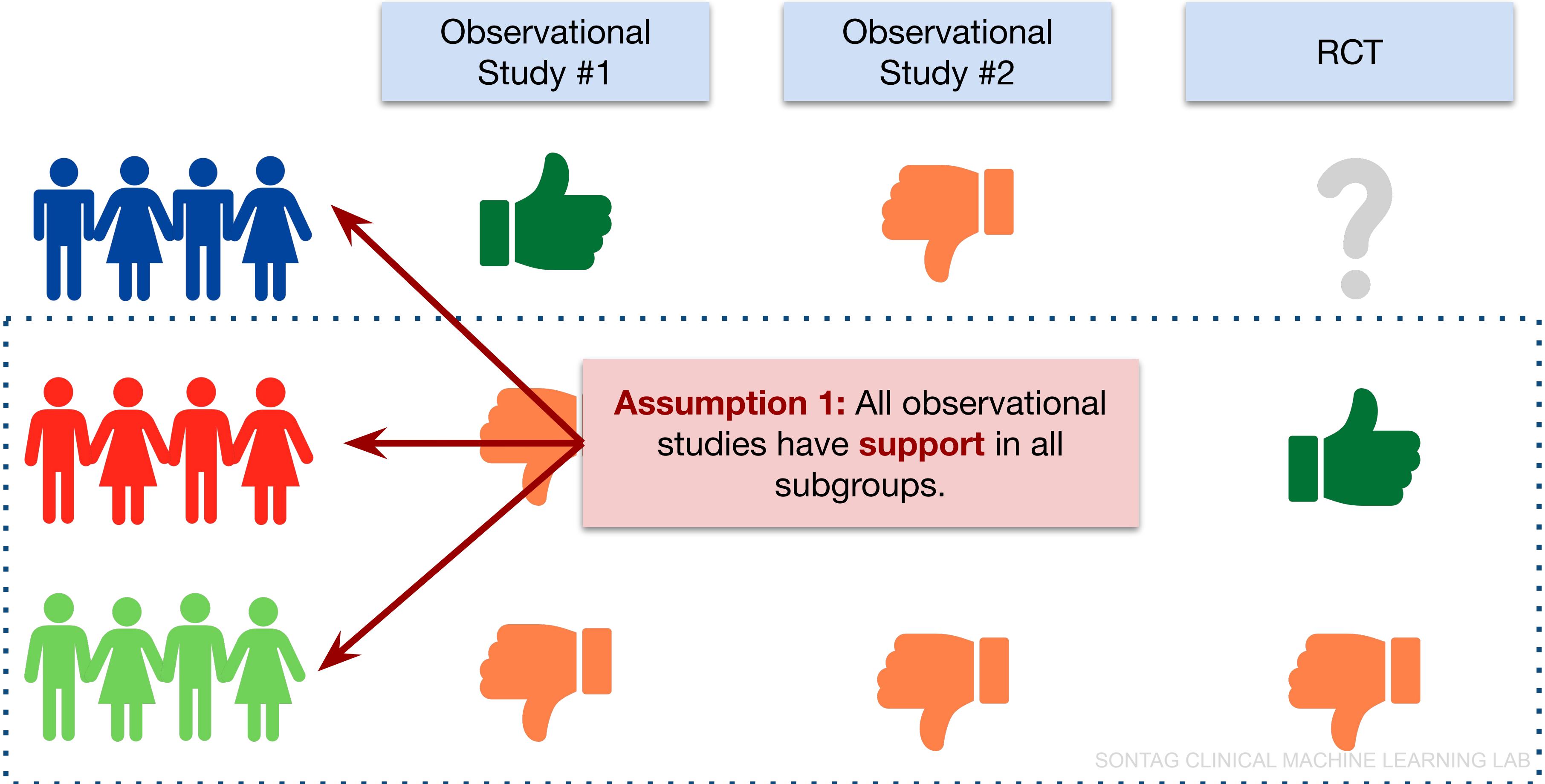
Observational Study #1	Observational Study #2	RCT



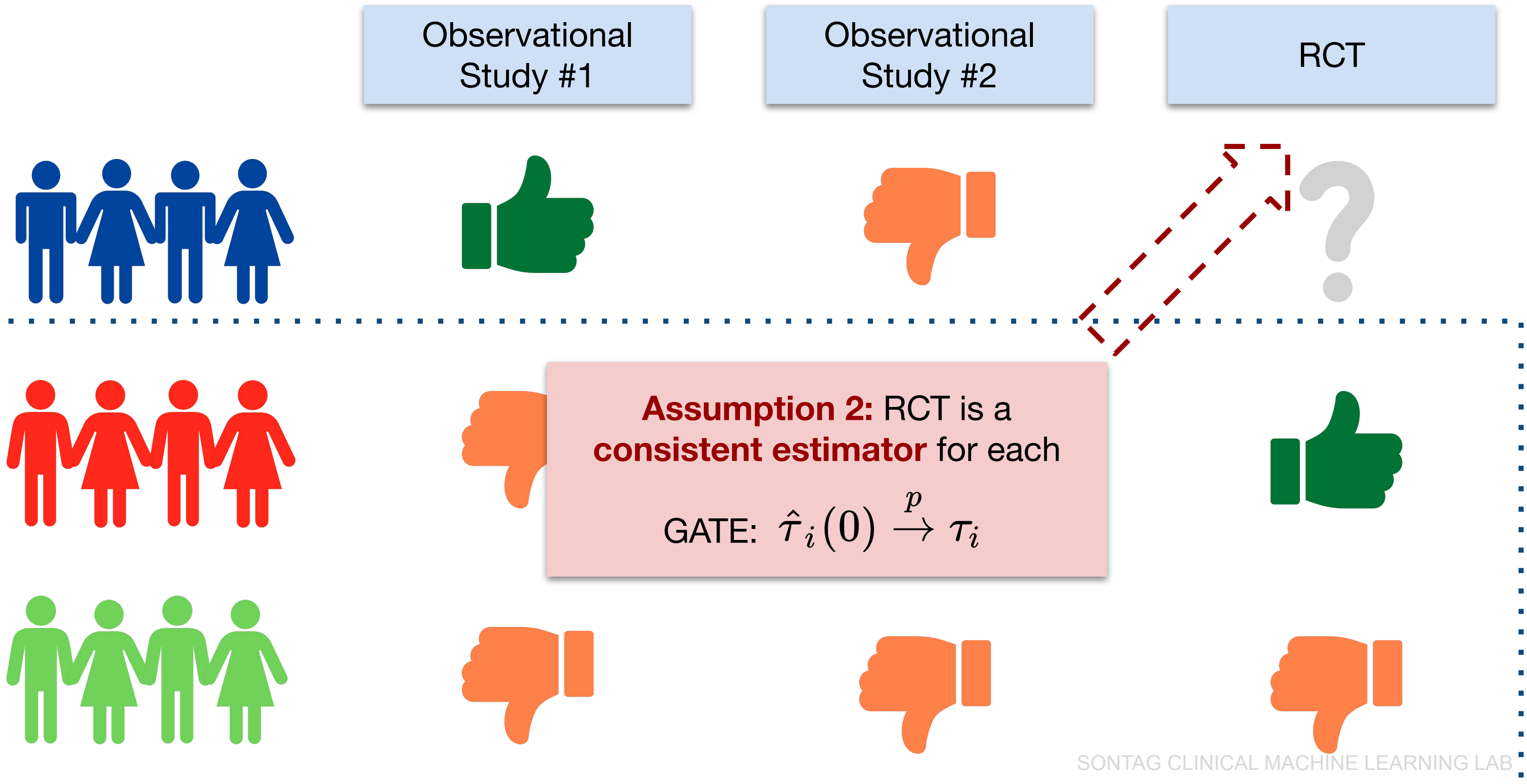
Formalizing Falsification



Formalizing Falsification



Formalizing Falsification

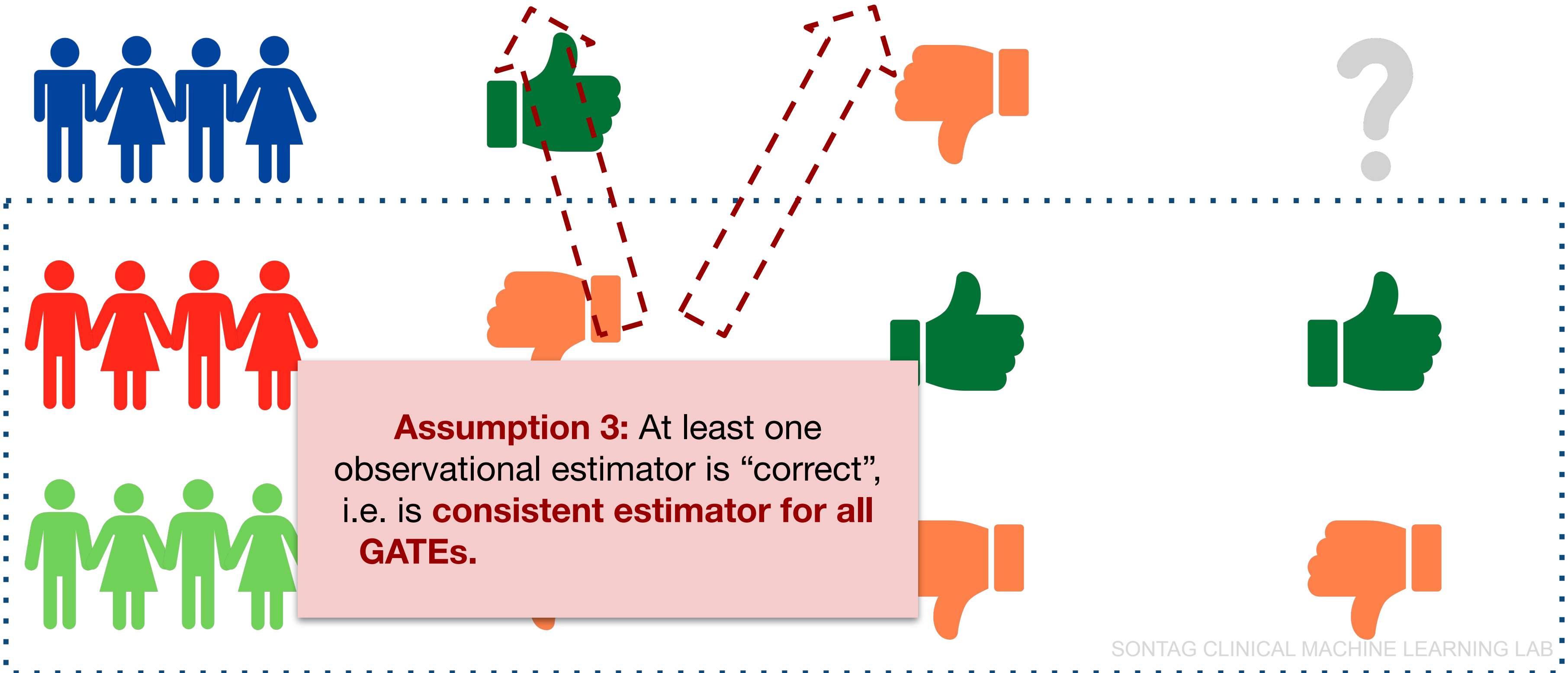


Assumption 2: RCT is a **consistent estimator** for each

GATE: $\hat{\tau}_i(0) \xrightarrow{p} \tau_i$

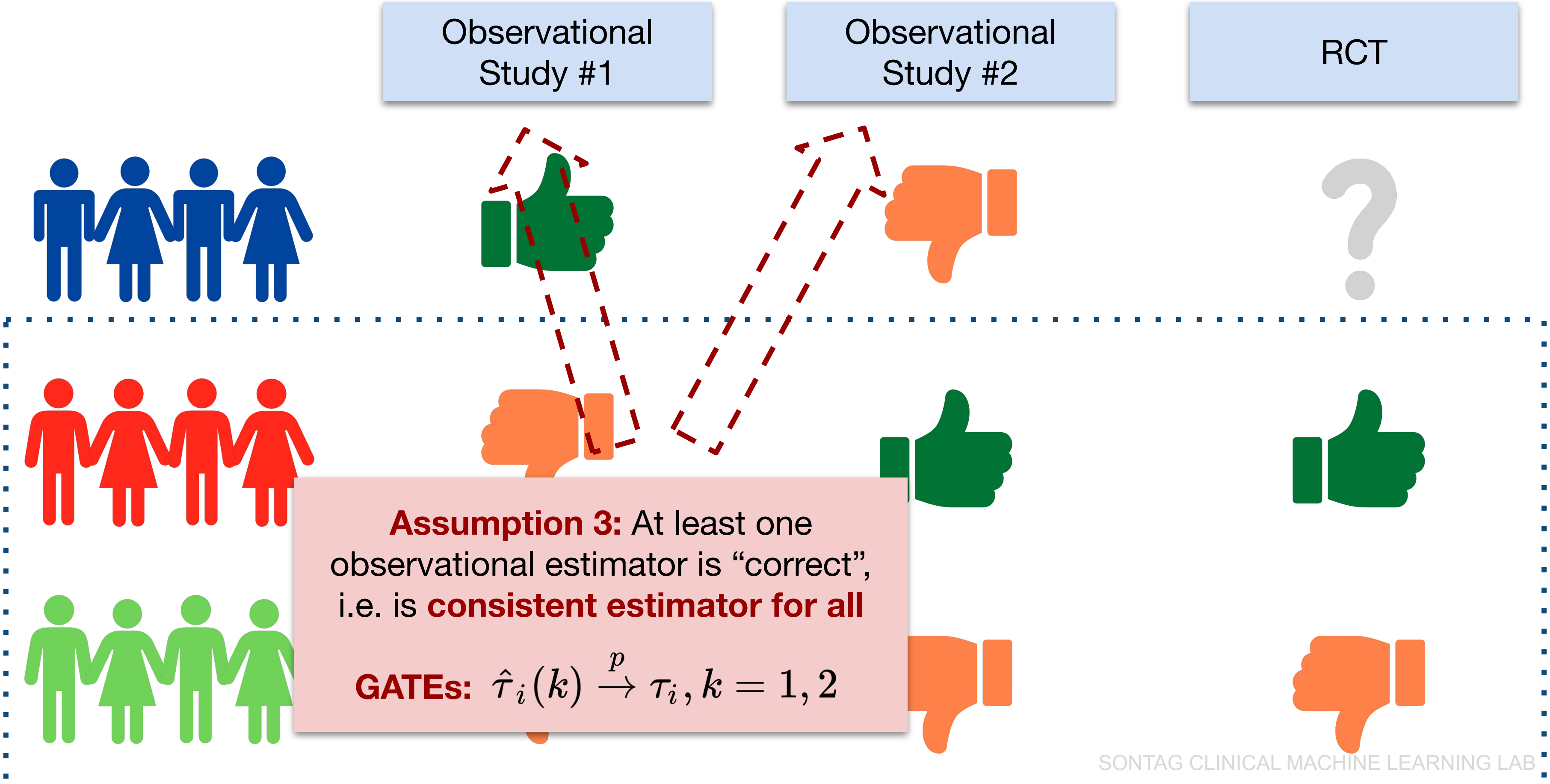
Formalizing Falsification

	Observational Study #1	Observational Study #2	RCT



Assumption 3: At least one observational estimator is “correct”, i.e. is **consistent estimator for all GATEs.**

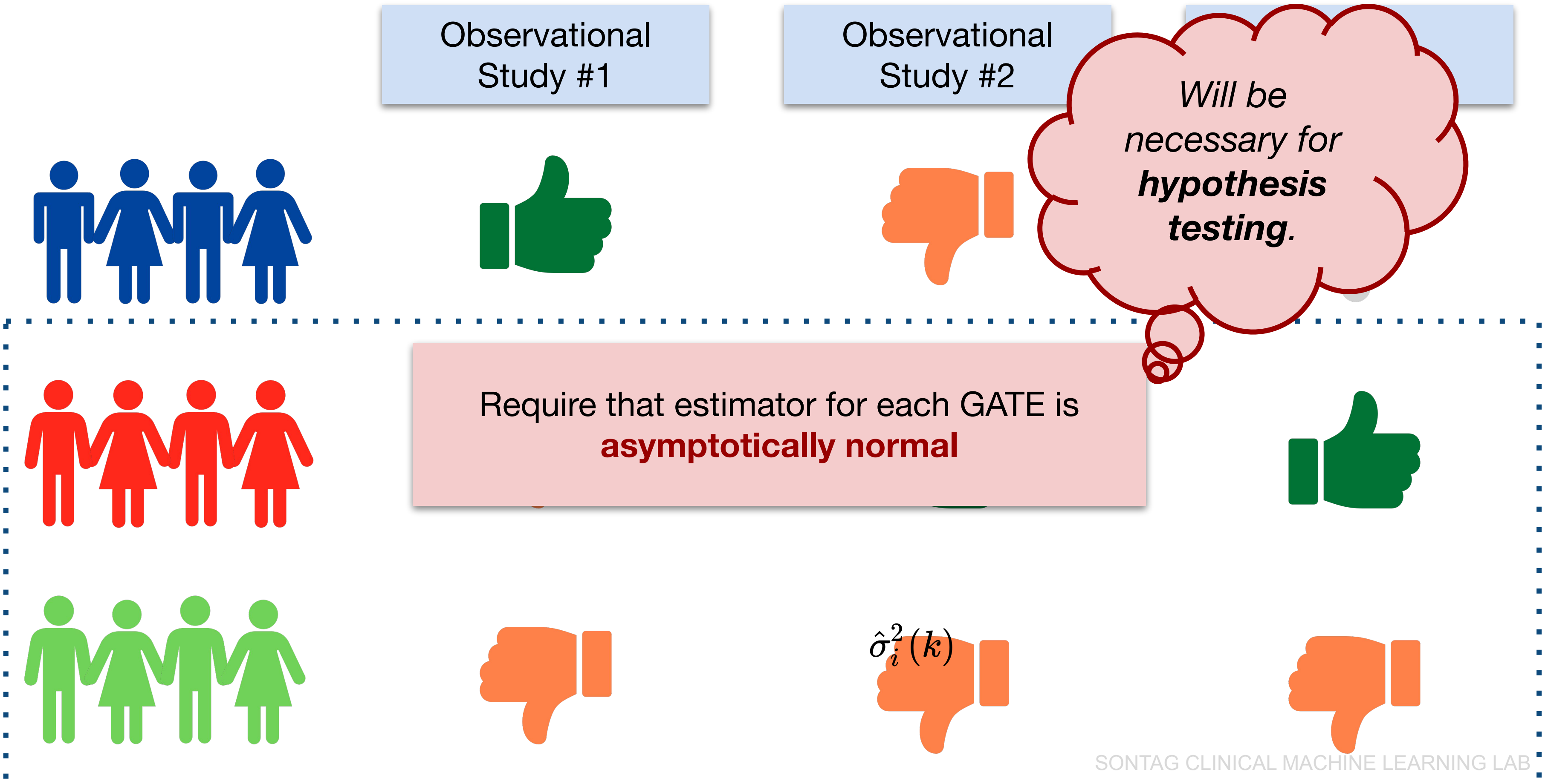
Formalizing Falsification



Assumption 3: At least one observational estimator is "correct", i.e. is **consistent estimator for all**

GATEs: $\hat{\tau}_i(k) \xrightarrow{p} \tau_i, k = 1, 2$

Formalizing Falsification



Formalizing Falsification

Observational Study #1

Obser…

Will be necessary for ***hypothesis testing,*** *and we give examples where this is reasonable.*

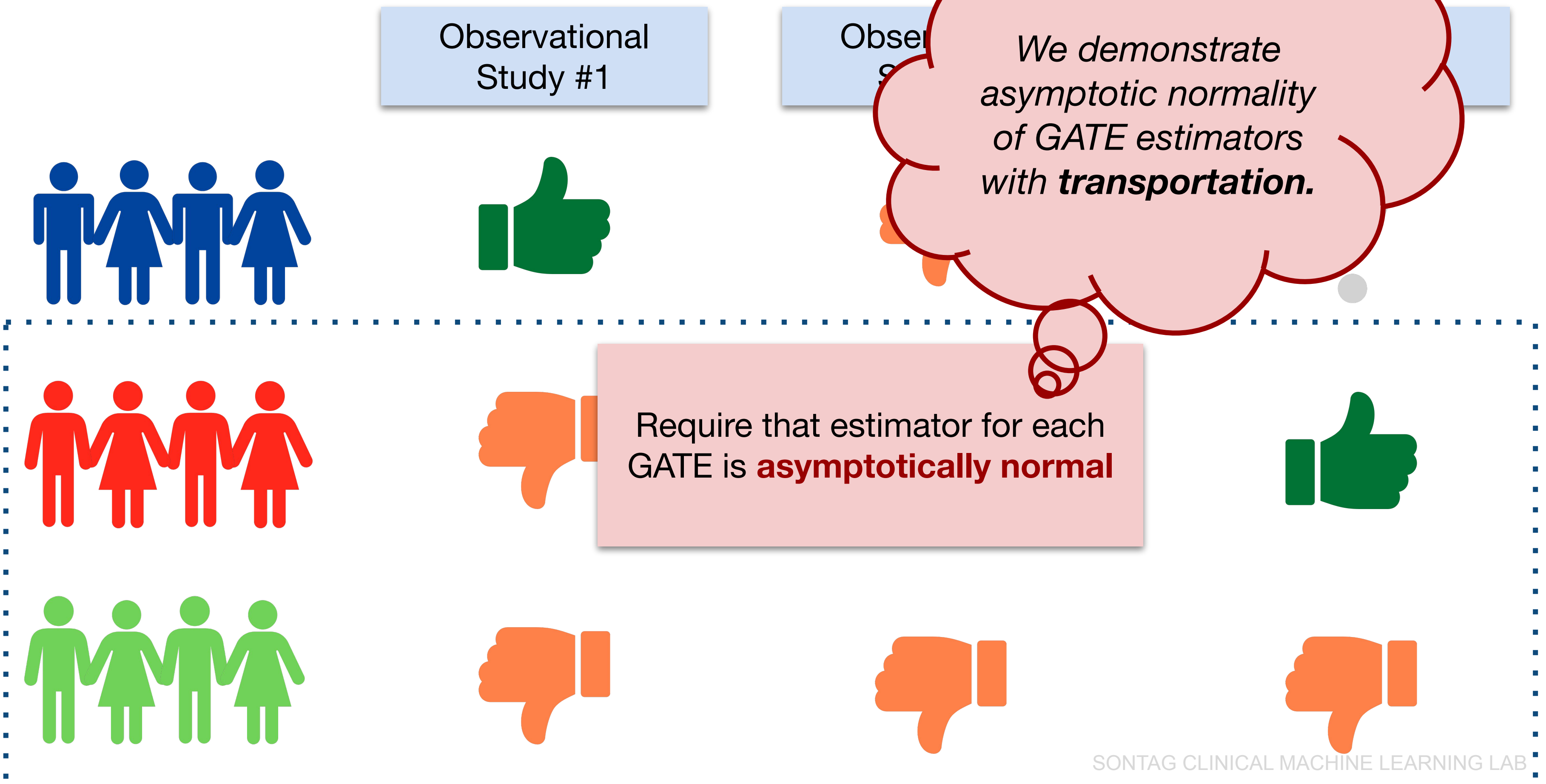
Require that estimator for each GATE is **asymptotically normal**

$$\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))/\hat{\sigma}_i(k) \xrightarrow{d} \mathcal{N}(0,1)$$

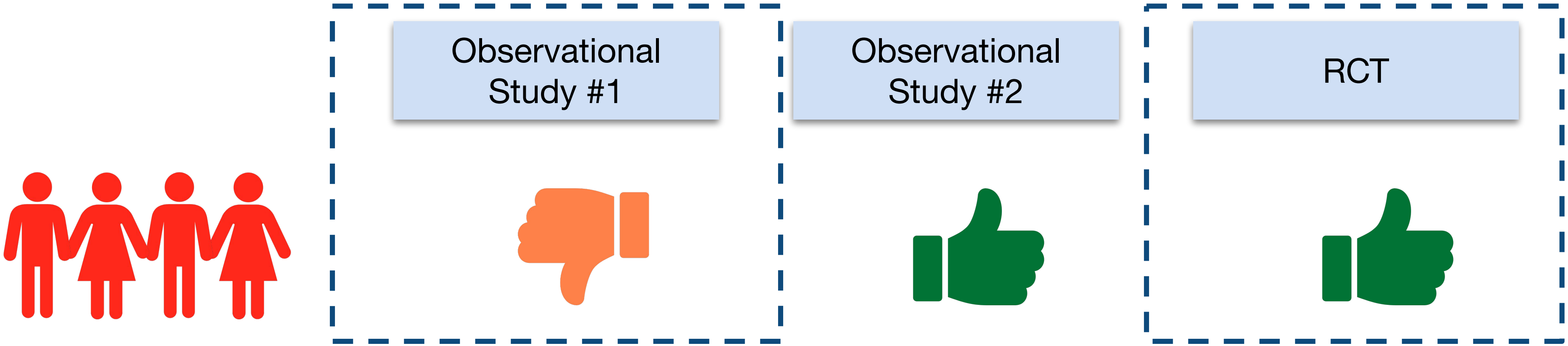
Sample size of observational study (k=1,2) or RCT (k=0)

$\hat{\sigma}_i^2(k)$ is estimate of variance, converges in probability to asymptotic variance

Formalizing Falsification



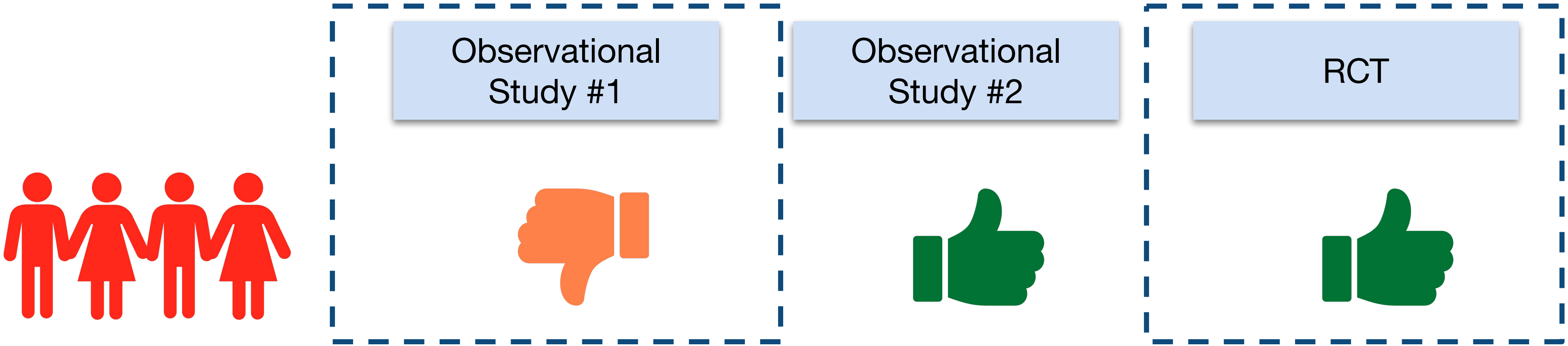
Hypothesis Test Construction



$$H_0 : \tau_{\text{red}}(1) = \tau_{\text{red}}$$

Want to perform above **hypothesis test** with **asymptotic level**, α

Hypothesis Test Construction



$$H_0 : \tau_{\text{red}}(1) = \tau_{\text{red}}$$

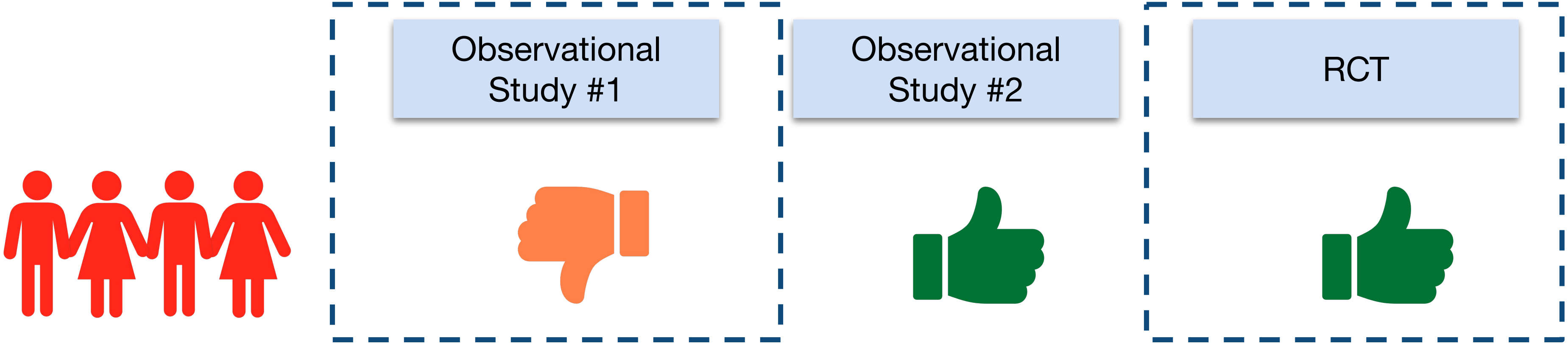
We can use the following test statistic, which we show **converges in distribution** to a standard **normal distribution**

$$\hat{T}_N(k = 1, i = \text{red people}) := \frac{(\hat{\tau}_i(1) - \hat{\tau}_i(0)) - (\tau_i(1) - \tau_i)}{\frac{\hat{\sigma}_i^2(1)}{N_1} + \frac{\hat{\sigma}_i^2(0)}{N_0}}$$

Set equal to 0

Estimated variance

Hypothesis Test Construction



$$H_0 : \tau_{\text{red}}(1) = \tau_{\text{red}}$$

Set equal to 0

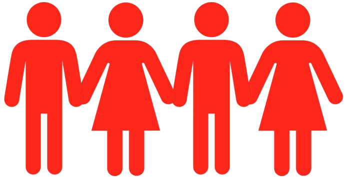
Reject the observational study if

$$|\hat{T}_N(k=1, i=\text{red people})| > z_{\alpha/2}$$

$$\hat{T}_N(k=1, i=\text{red people}) := \frac{(\hat{\tau}_i(1) - \hat{\tau}_i(0)) - (\tau_i(1) - \tau_i)}{\frac{\hat{\sigma}_i^2(1)}{N_1} + \frac{\hat{\sigma}_i^2(0)}{N_0}}$$

Estimated variance

Hypothesis Test Construction



Observational Study #1

Observational Study #2

RCT

$$H_0 : \tau_{\text{red}}(1) = \tau_{\text{red}}$$

Set equal to 0

$$\hat{T}_N(k = 1, i = \text{red people}) := \frac{(\hat{\tau}_i(1) - \hat{\tau}_i(0)) - (\tau_i(1) - \tau_i)}{\frac{\hat{\sigma}_i^2(1)}{N_1} + \frac{\hat{\sigma}_i^2(0)}{N_0}}$$

Estimated variance

Note that we use **Bonferroni correction to control FPR of test**, since we test many subgroups (e.g. red people, blue people, etc.)

Our Approach

1

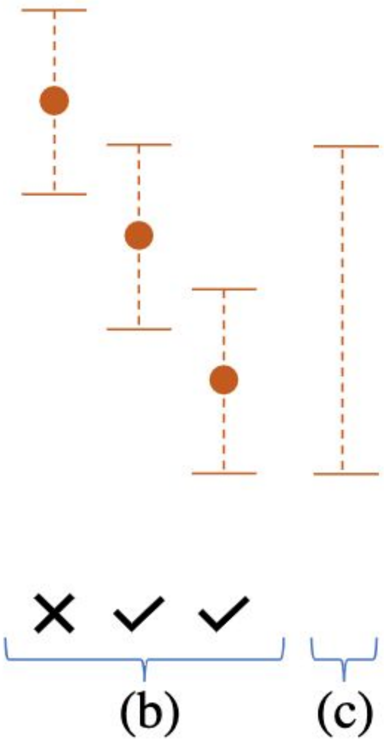
Falsification of observational estimates

2

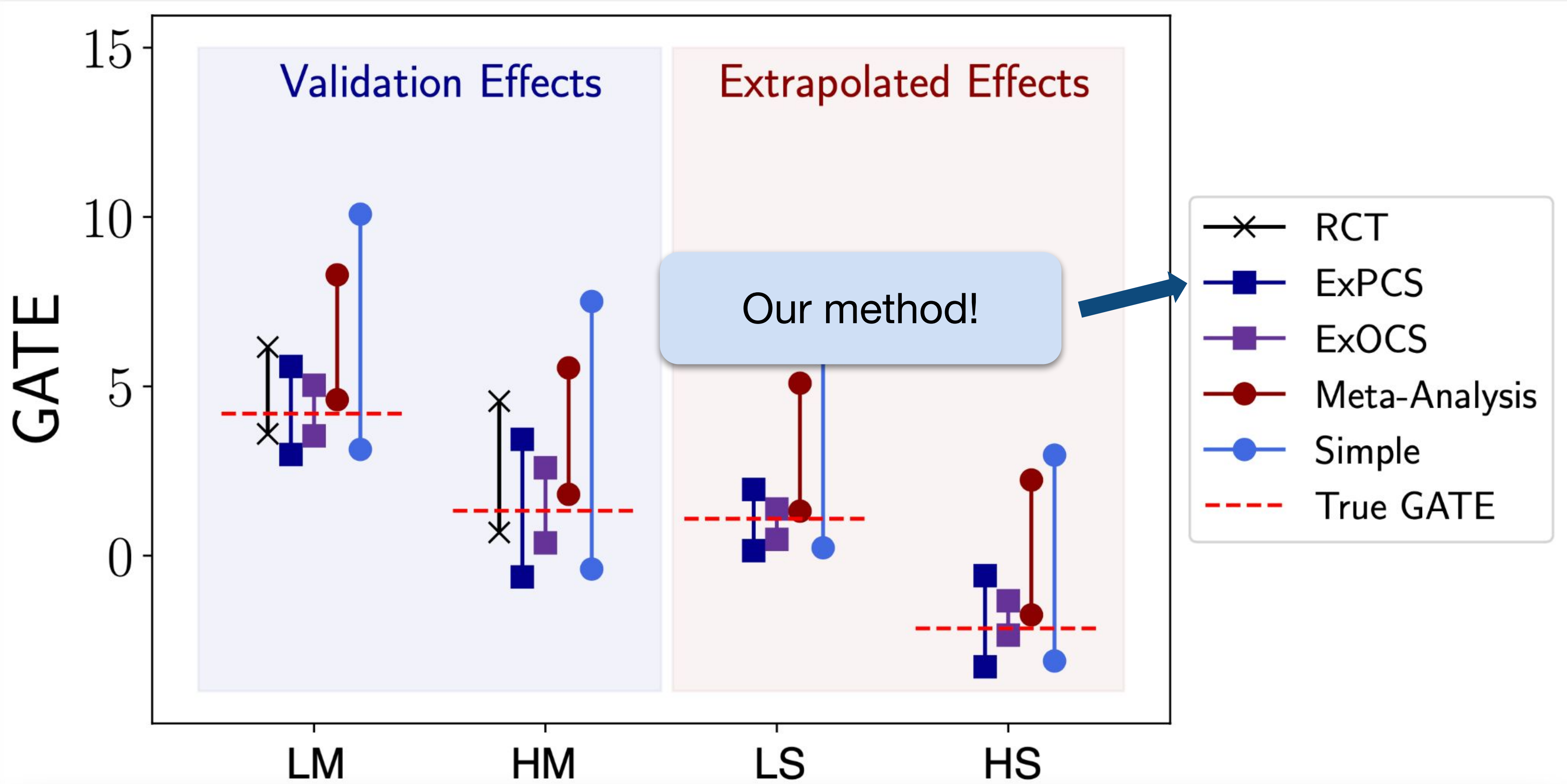
Pessimistic Combination of Confidence Intervals

2

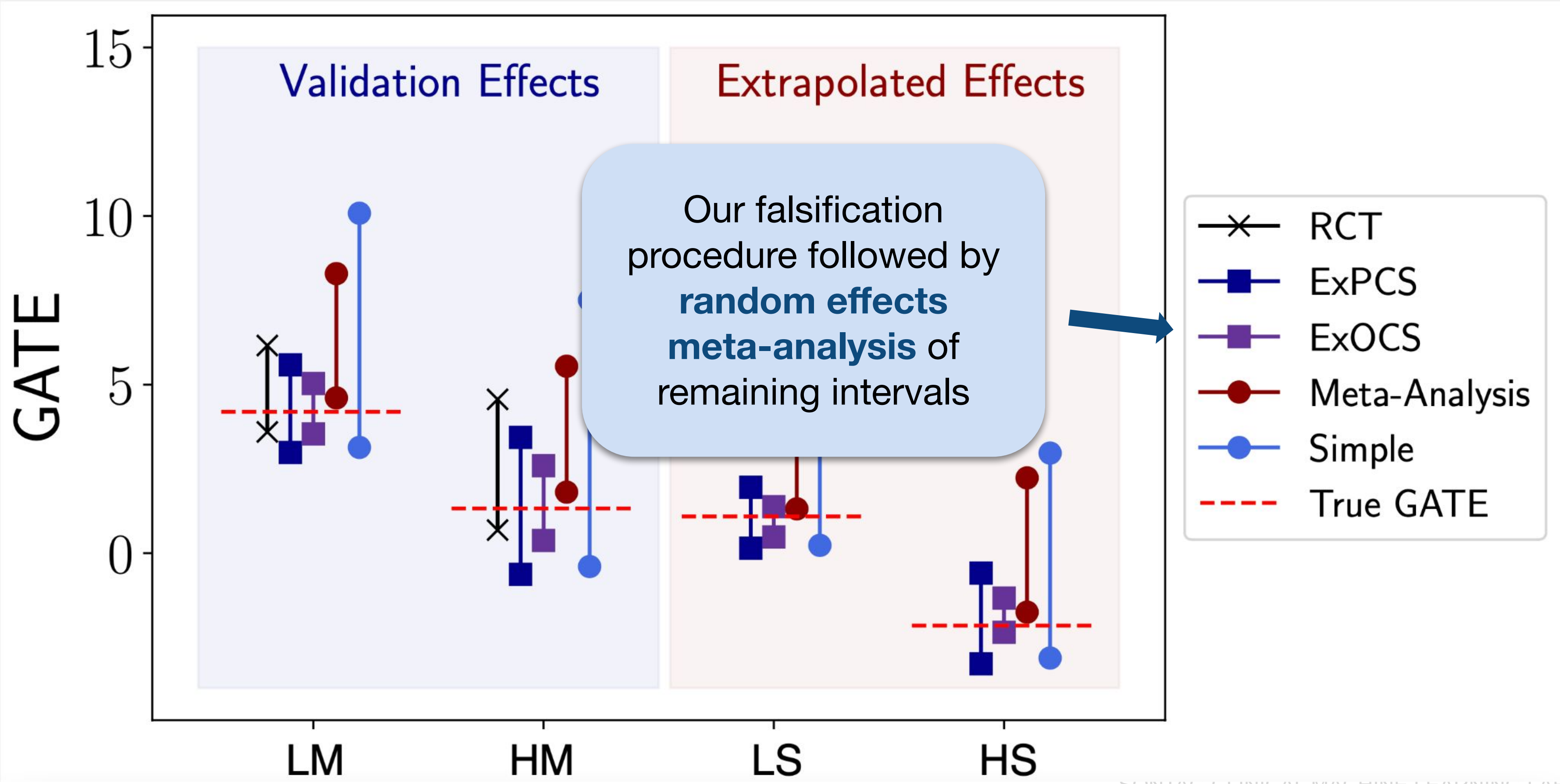
Pessimistic Combination of Confidence Intervals



Results on Semi-Synthetic

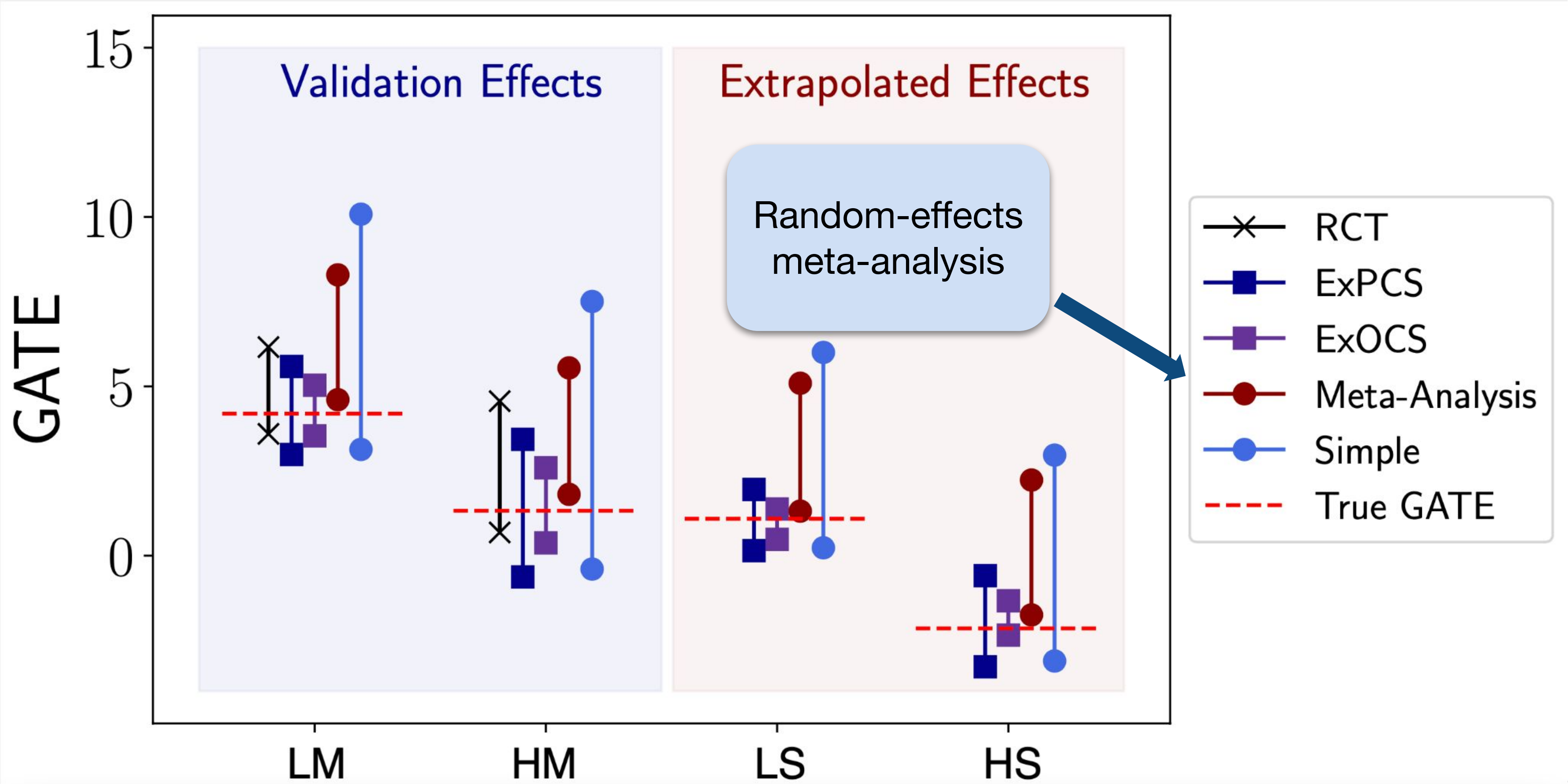


Results on Semi-Synthetic

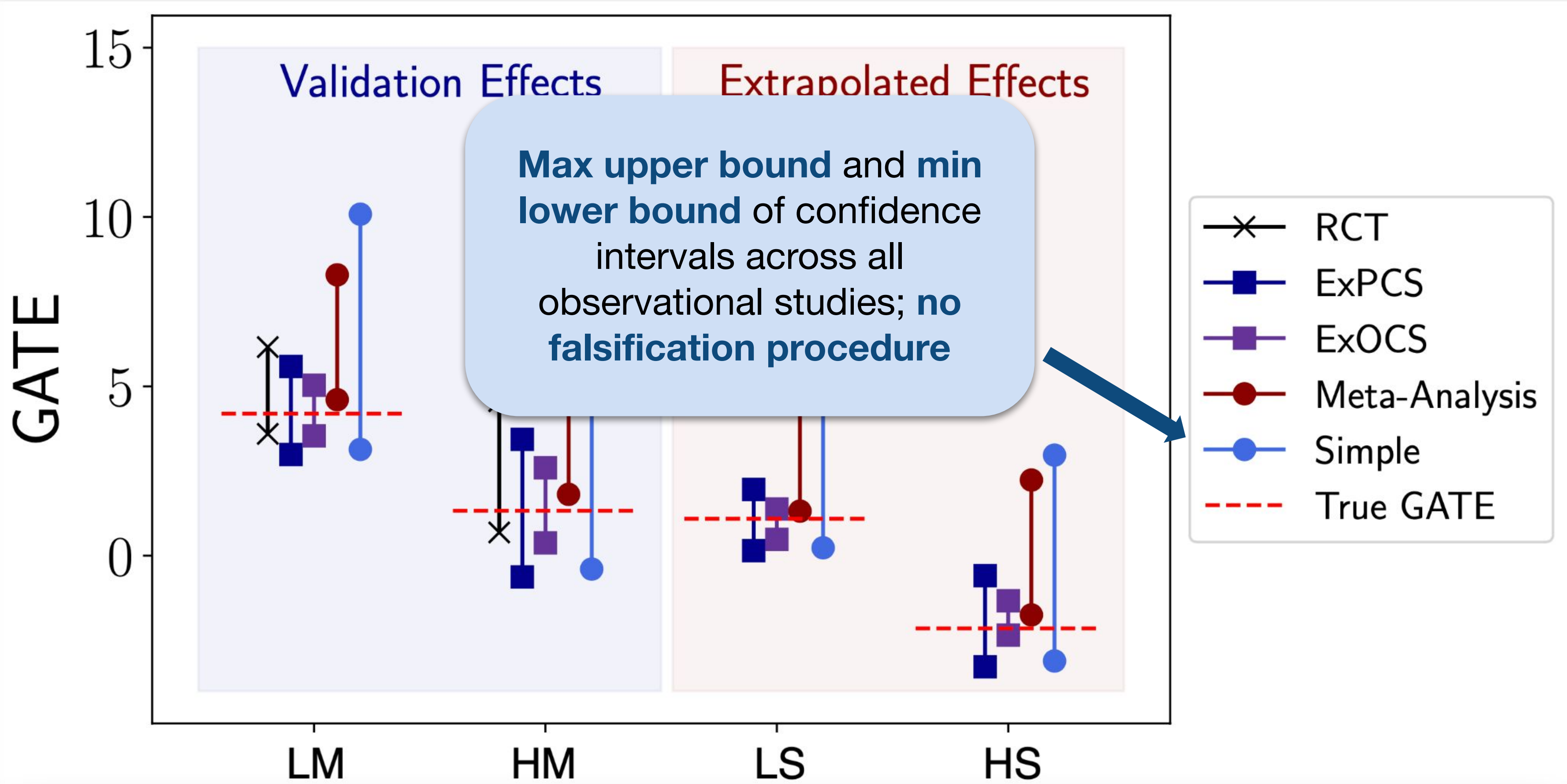


Our falsification procedure followed by **random effects meta-analysis** of remaining intervals

Results on Semi-Synthetic



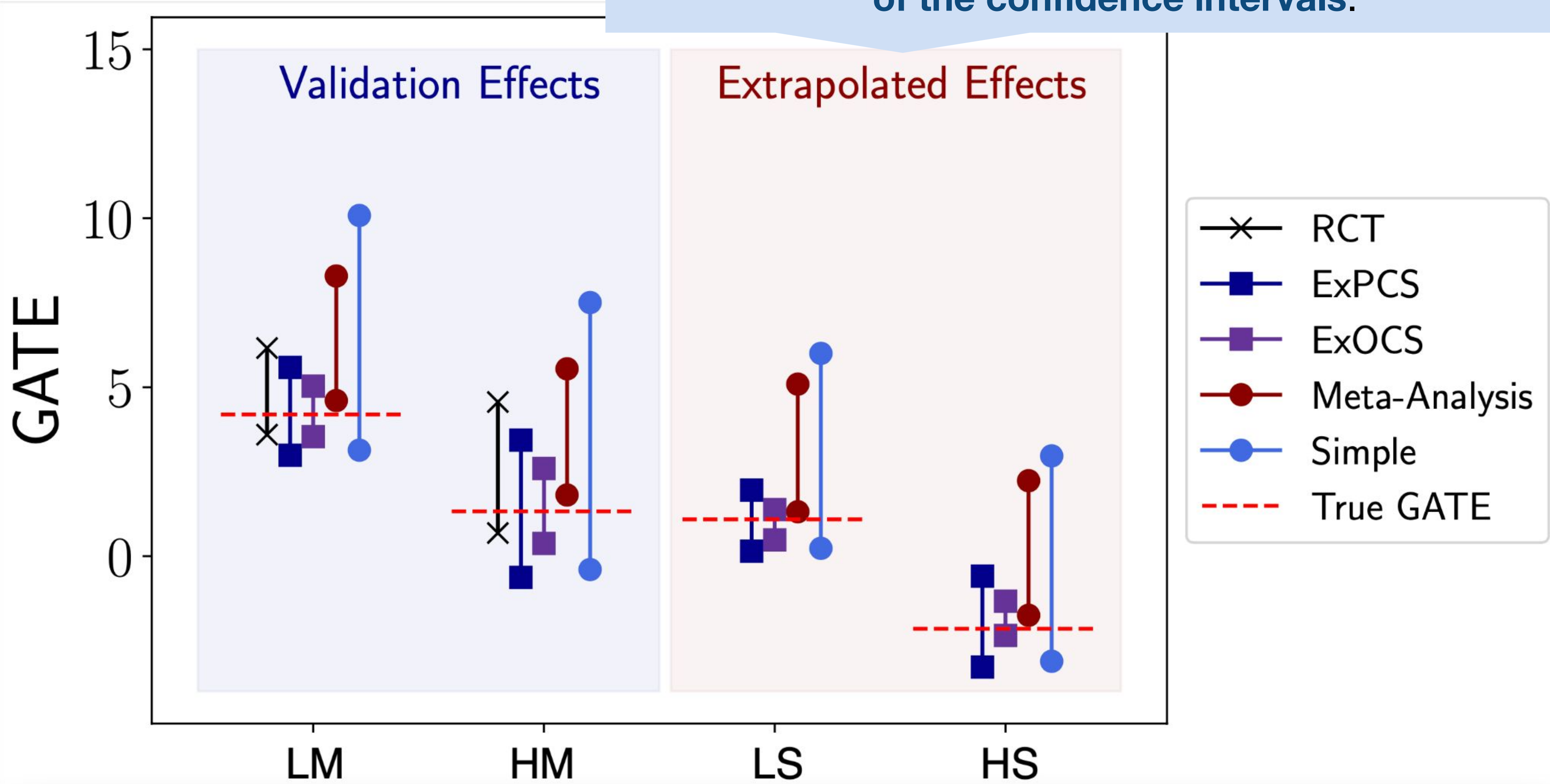
Results on Semi-Synthetic



Max upper bound and **min lower bound** of confidence intervals across all observational studies; **no falsification procedure**

Results on Semi-S

Compared to baselines, our approach has the best balance between **coverage of the true GATE** and **width of the confidence intervals**.



For more results and discussion, visit us at poster ID 54677!

Thank you!

Results on Women's Health Initiative Data

	Coverage	Length	OS %
Simple	0.39	0.416	–
Meta-Analysis	0.03	0.260	–
ExOCS	0.28	0.058	–
ExPCS (ours)	0.45	0.081	0.99
Oracle	0.44	0.068	–

Table 1: Coverage, length, and unbiased OS % of ExPCS and baselines. ExPCS achieves comparable coverage to the oracle method with highly efficient intervals. Additionally, we do not reject the unbiased OS in 99% of the tasks.