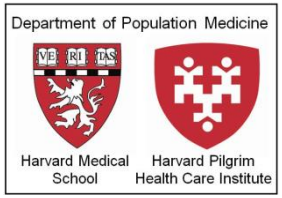
Practical Advice on Research Designs to Evaluate Natural Experiments

O. Kenrik Duru, MD, MSHS
Frank Wharam, MD, MPH



Natural Experiments for Translation in Diabetes: Funded by CDC/NIDDK

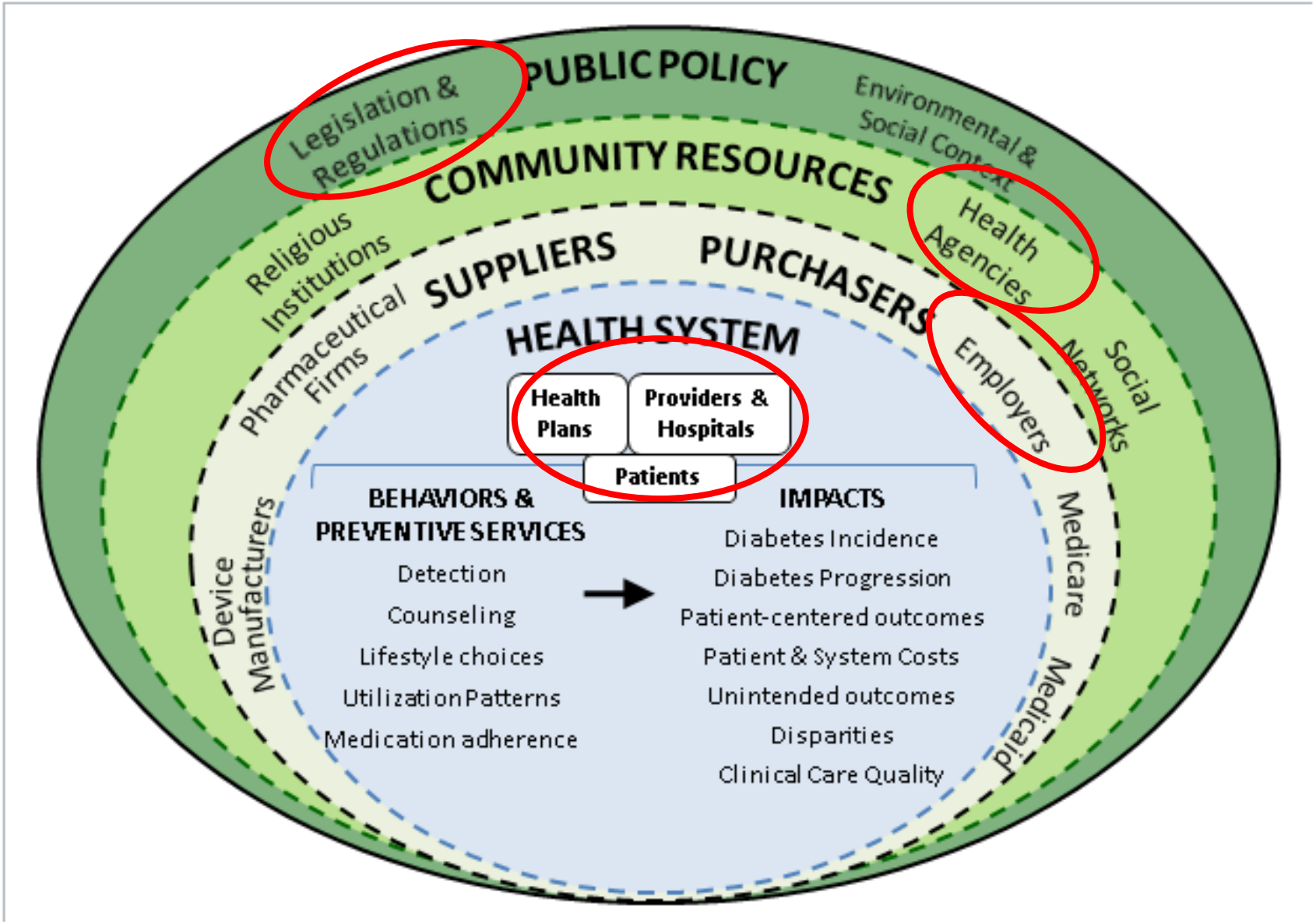
Overview of Seminar

- What is a “natural experiment?”
- What types of data can be used to study natural experiments?
- What rigorous study designs can/should be used to study natural experiments (examples given)?

What is a “Natural Experiment?”

- A policy or intervention which is implemented “naturally” (e.g., not controlled by researchers)
- This is a key difference from researcher-led RCTs and other experimental designs
- Unfortunately, these policies and interventions often go unevaluated, or are evaluated with designs that are vulnerable to considerable bias

Where Can You Find Natural Experiments?



Why Evaluate Natural Experiments?

- RCTs are the gold standard but are often impractical or inappropriate to evaluate policies or “real-world” interventions
- Conversely, natural experiments use pragmatic designs and readily available data to evaluate and compare a new or existing policy to other policy alternatives of what may have happened in the absence of any intervention
- Natural experiments can be designed to yield robust causal inference

Types of Routine Health System Data

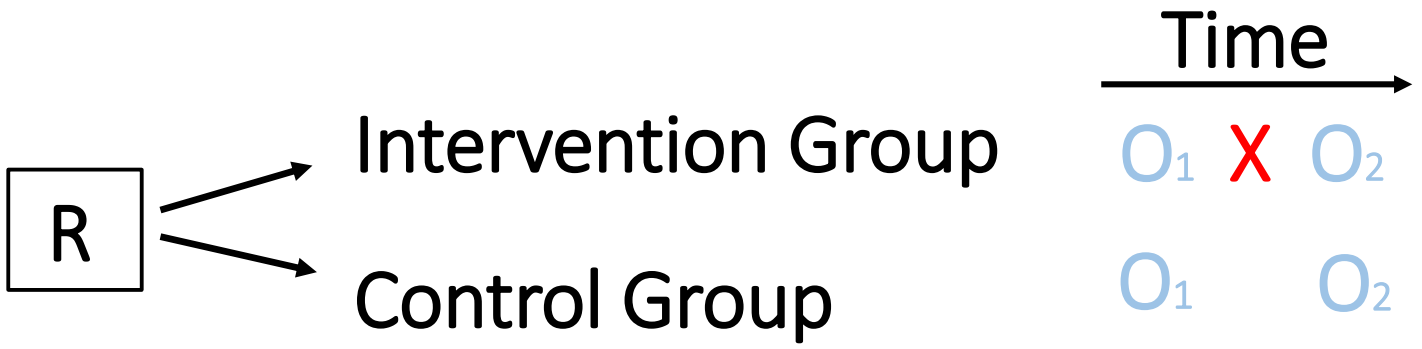
- Clinical data (e.g., EHR data)
  - Generated during process of care
  - Rich clinical detail but difficult to standardize
- Administrative data (e.g., claims)
  - Derived from payment or supply systems
  - More standardized but less detail
- Medical technologies
  - Collected when drugs or devices are approved, sold to providers, or used

Secondary Uses of Routine Data

- Disease surveillance
- Safety surveillance
- Longitudinal large-scale epidemiology
- Clinical quality & effectiveness
- Efficiency & cost-effectiveness
- Evaluating natural experiments

RCTs: Gold Standard in Study Design but Rare in Natural Experiments

Randomized Controlled Trial



X=policy intervention O_t=Measurement at time t

- Typically uses a Difference in Difference (DiD) analysis approach
- Baseline comparability maximized by randomization

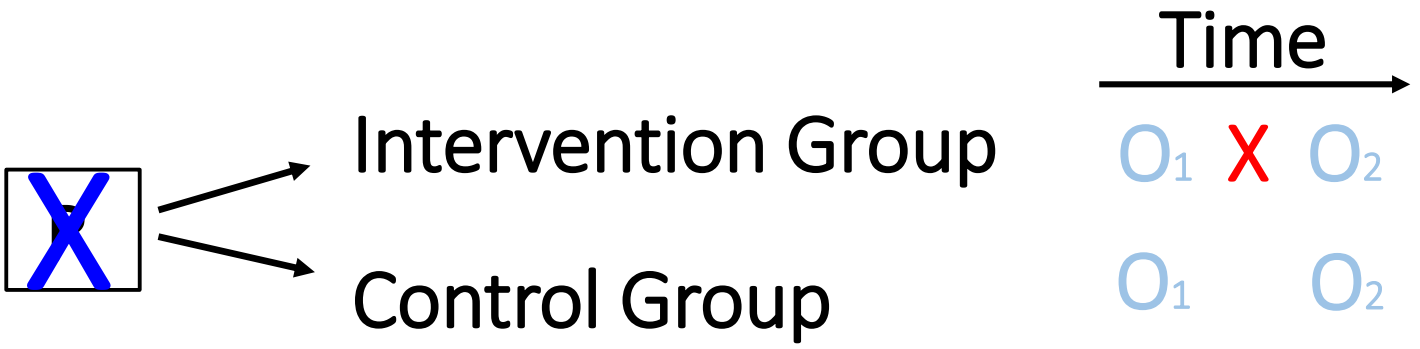
Feasibility of Randomization in Natural Experiments

- Structural constraints
  - Impossible to prevent exposure
- Political constraints
  - Assignment to controls unacceptable
  - Desire to roll out quickly
- Ethical constraints
  - Withholding beneficial program

Best practice: Retain as many elements of randomization as possible in design and test for group differences

"Quasi-Experimental" Study Design Is Typical in Natural Experiments

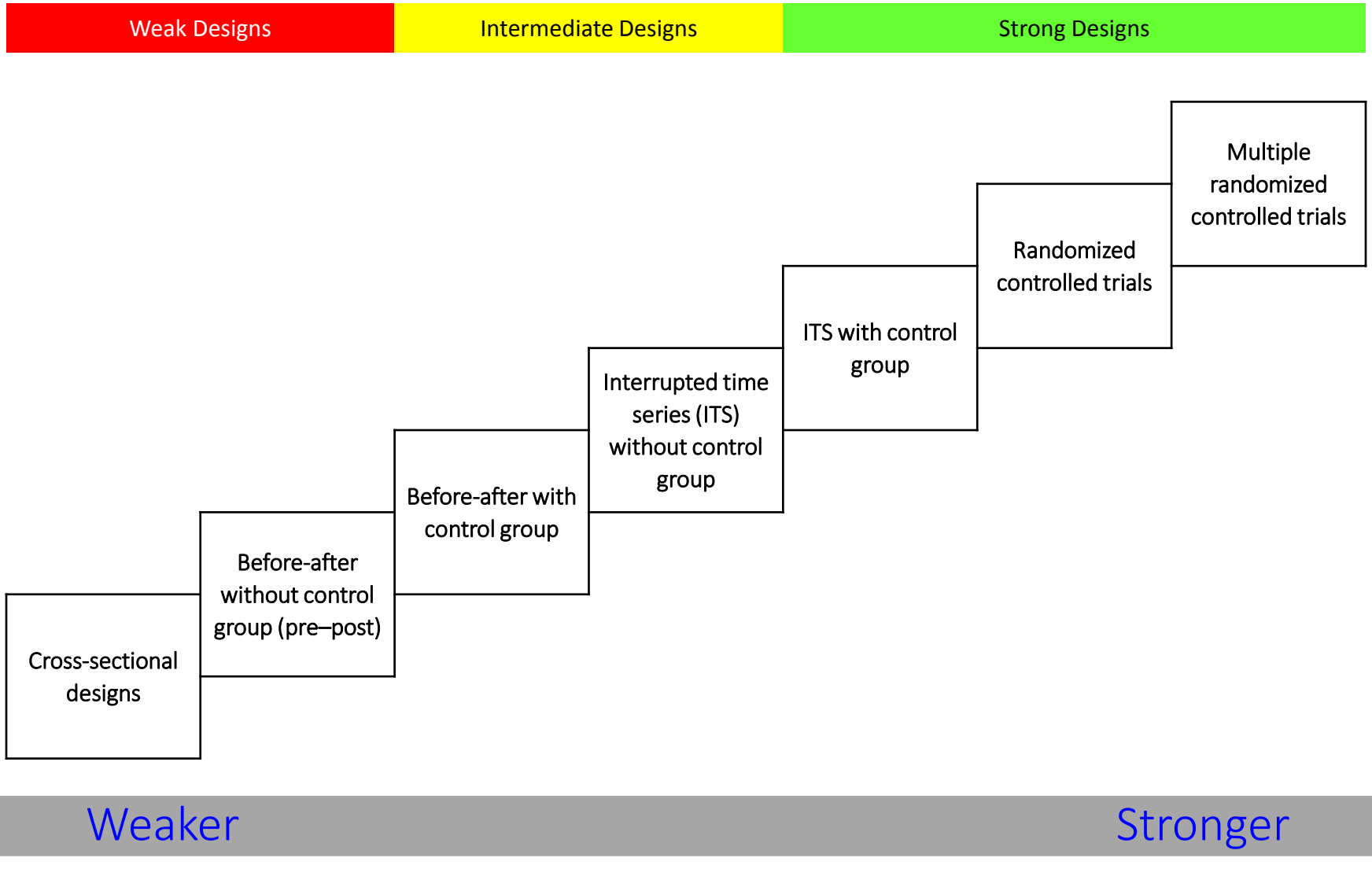
Non-Random Control Group Design



X=policy intervention O_t=Measurement at time t

- Also uses a Difference in Difference (DiD) analysis approach
- Baseline comparability is a key threat to validity

Hierarchy of research design



Adapted from: Soumerai S *et al. Prev Chron Dis.* 2015.

Methods to Improve Comparability of Non-Random Control Groups

- Propensity scores
  - Score based on logistic regression model predicting likelihood of being in study group
  - Can be used to match study and control members or to weight statistical analysis
  - Potential bias due to unobserved factors
- Instrumental variables
  - In theory, control for both unobserved & observed patient characteristics affecting outcome
  - Most common: distance to facility; regional variation, facility variation, physician variation
  - Potential bias due to residual confounding

Considerations in Design Approach

- Availability of routine data
- Structure of implementation
  - Whole system at once
  - Clusters (clinics, worksites, communities, etc.)
  - Individuals
- Retrospective vs. prospective
  - Degree of control over implementation
- Type and timing of expected effects

Additional Analysis Issues in Evaluating Natural Experiments

- Take-up
  - % offered intervention who participate
- Fidelity
  - Same intervention in all groups, all individuals
- Magnitude of exposure
  - % of planned intervention experienced
- Chronological timing
  - When exposure occurs during intervention period
- Nature of observed change
  - Statistically significant vs. clinically important

Choosing a study design: general considerations

- In conducting research studies, investigators typically want to prove that intervention X causes effect Y
- The ability of various study designs to help prove that X causes Y (i.e., "generate causal inference") varies greatly
- Investigators should use the most rigorous study design possible…
  - To get as close as possible to the truth
  - To increase policy impact
  - Journals care

Basic Steps in Creating a Study of a Natural Experiment

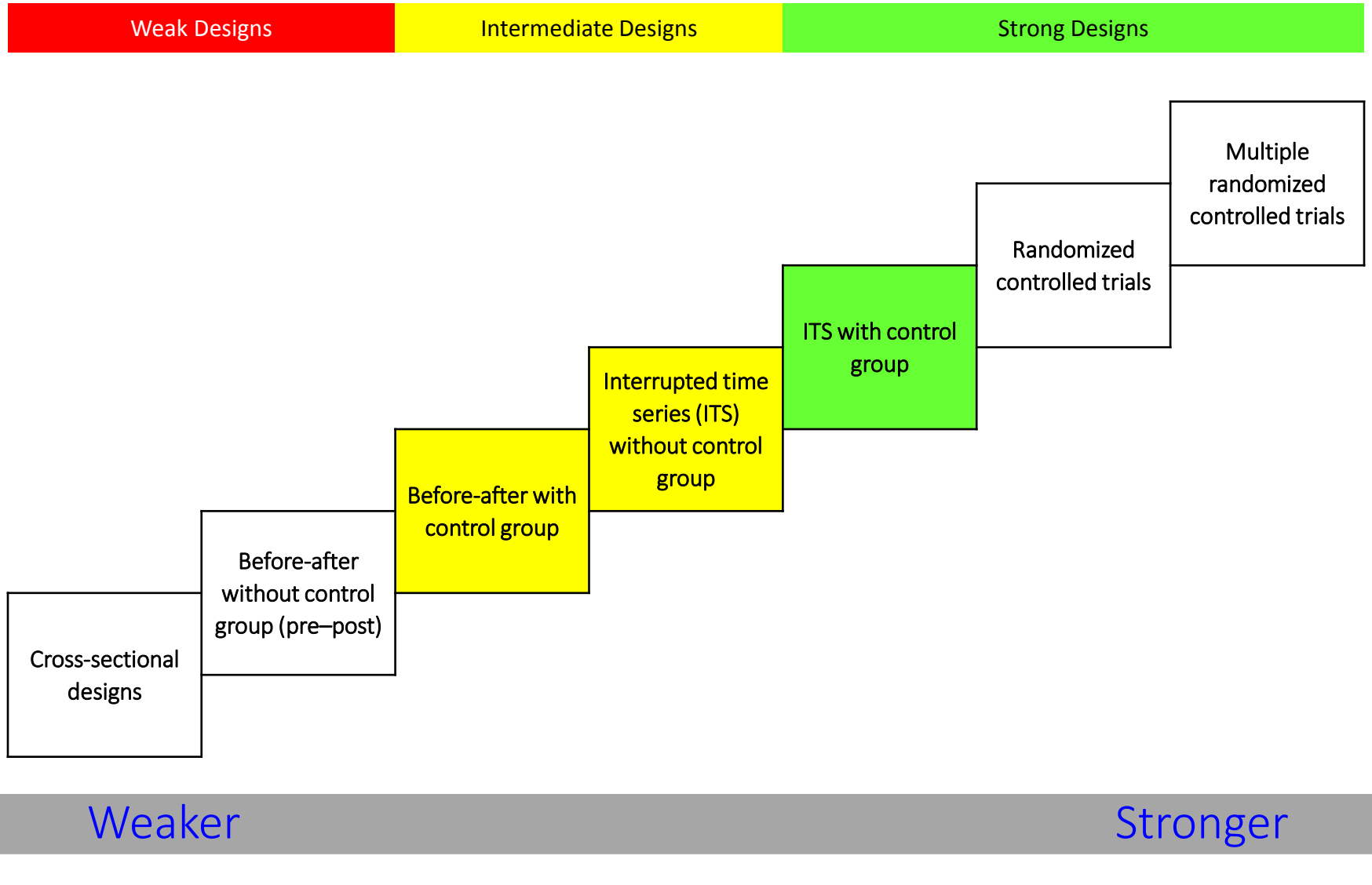
1. Identify a natural experiment, data source, and study design
2. Generate hypotheses, preferably with careful conceptual model
3. Construct denominator / intervention group
4. Construct numerator / measures that assess hypothesized effects
5. Generate control group or control group pool
6. Consider need for matching to improve internal validity

3 key factors for deciding on a study design for natural experiments

1. Stability/frequency of outcome
   - Related to frequency of outcome, duration over which assessed, how often repeated at the individual/population level, and sample size
2. Duration of baseline/follow-up or number of baseline/follow-up points
   - Longer baseline and follow-up generally better but not always needed
   - >= ~8-10 baseline/follow-up points recommended but depends on variability
3. Control group availability

Examples: 3 relatively rigorous research designs for analyzing natural experiments

3 relatively rigorous designs

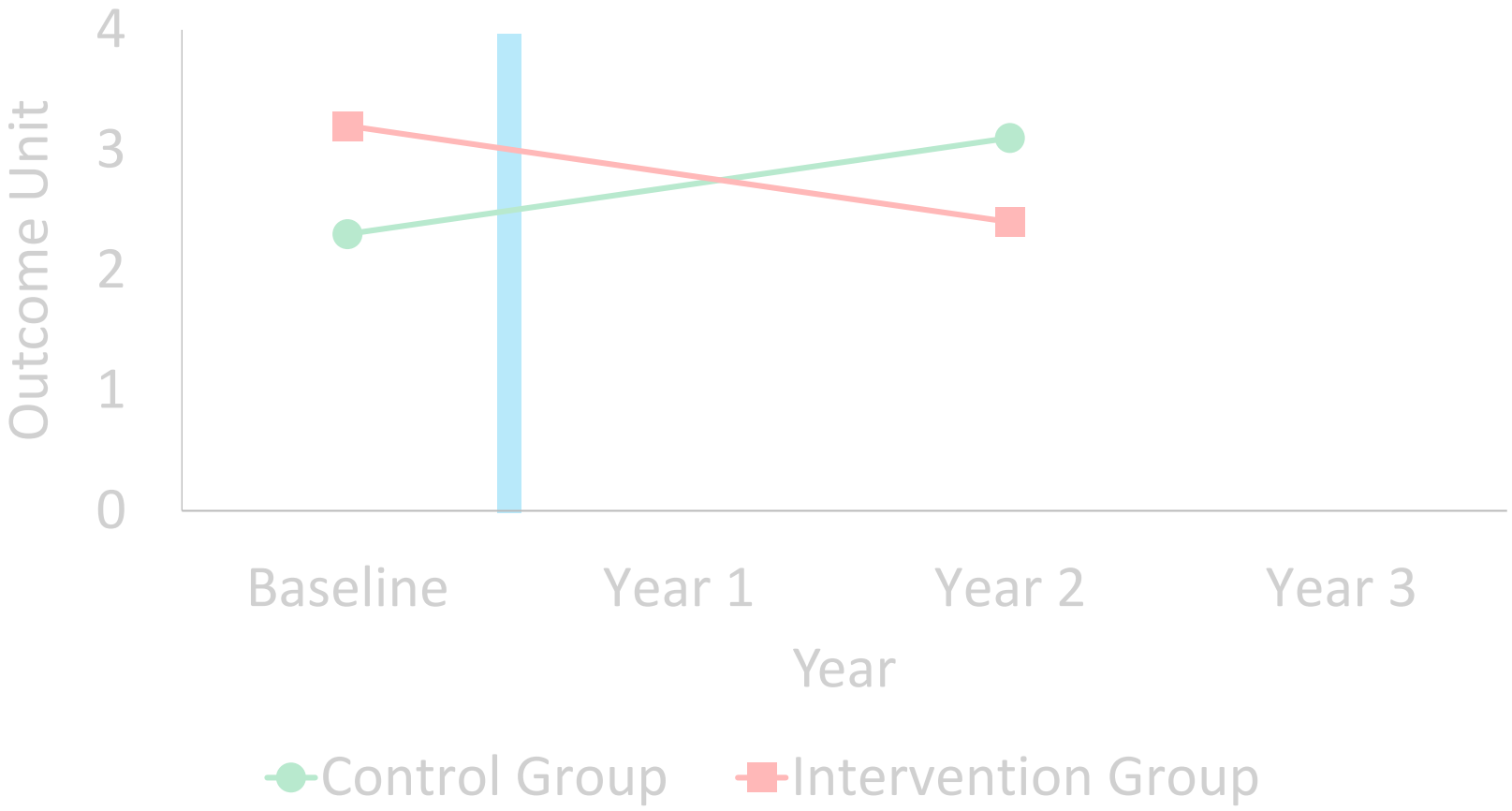


Adapted from: Soumerai S *et al. Prev Chron Dis.* 2015.

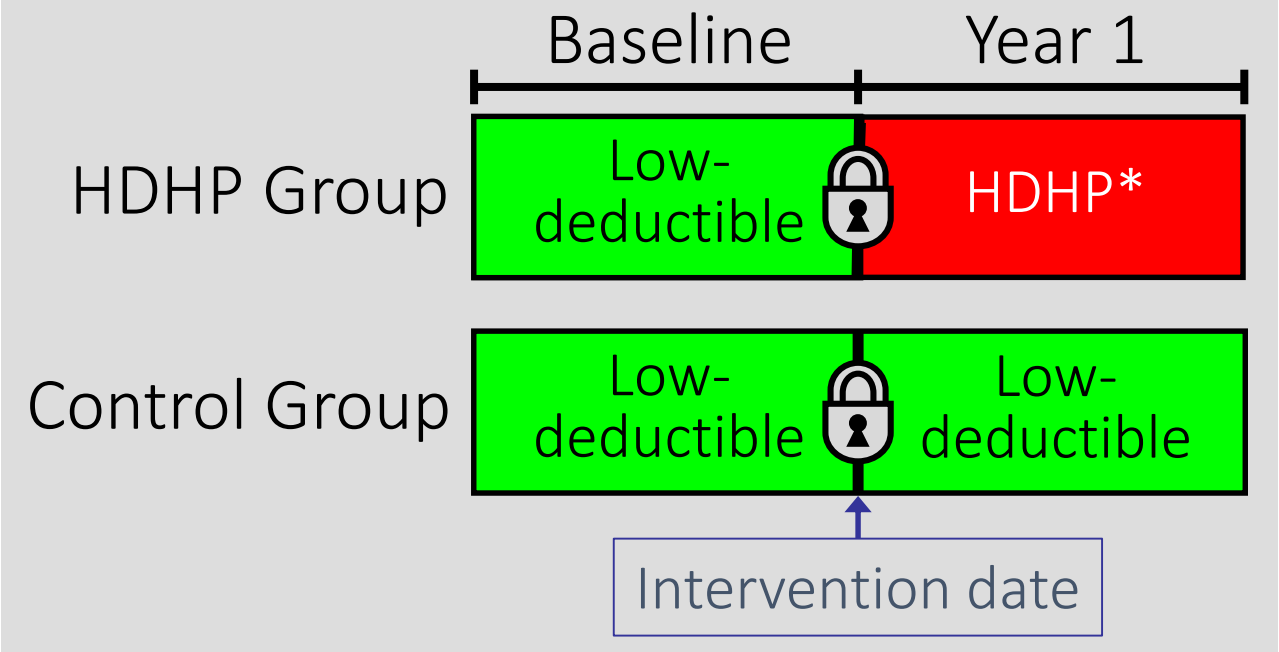
Threats to validity and reliability common to these 3 designs

- Threats to validity
  - Co-interventions, selection, regression to the mean, instrumentation
- Threats to reliability
  - Data quality (short segments, unstable data, missing data or wild data points)
  - Nature of population or process (changing denominators, rare outcomes, non-linear trends)

1. Before-after with control group



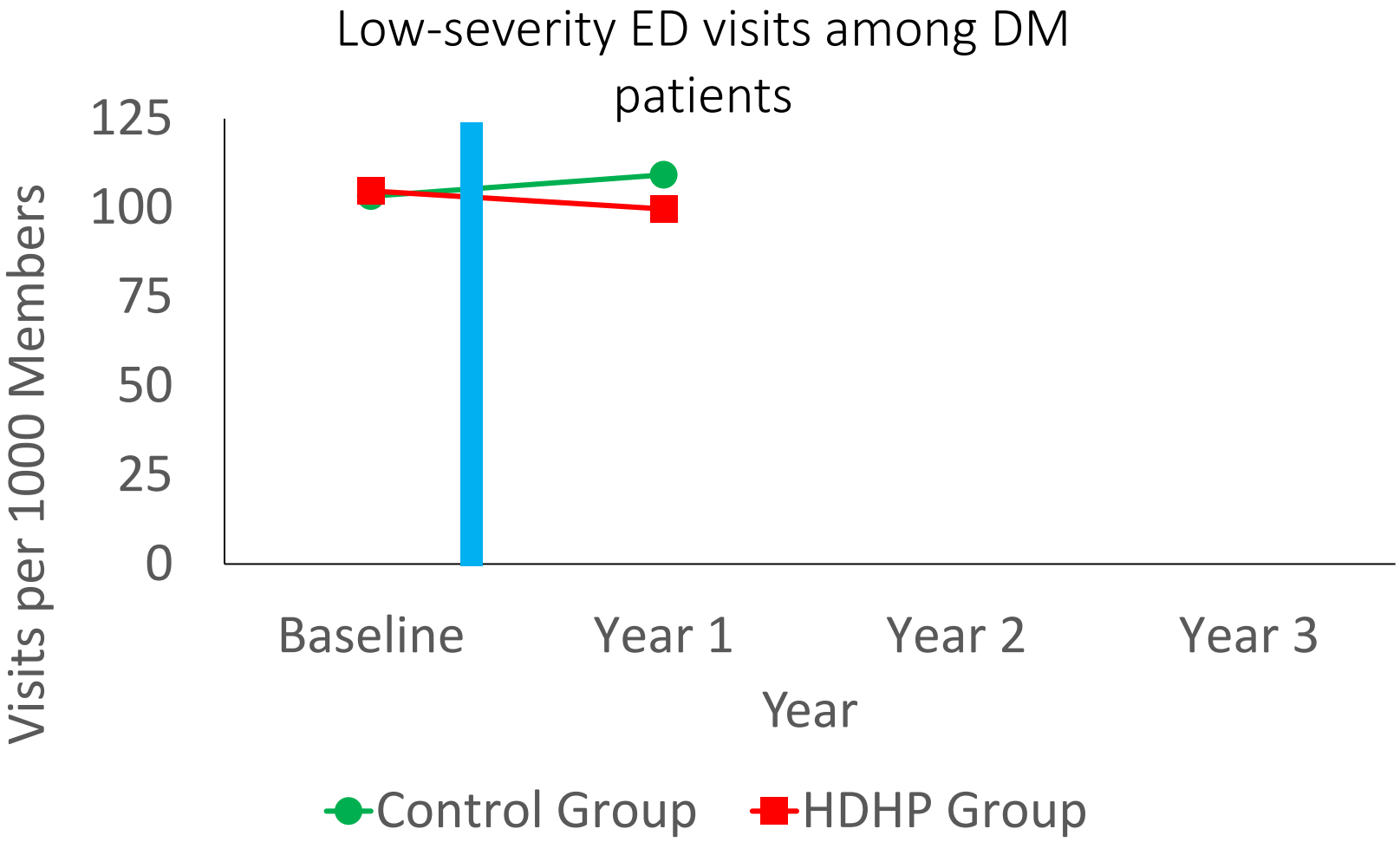
1. Before-after with control group: example



- Background: HDHPs might reduce low-severity emergency department visits among DM patients
- Natural experiment: DM patients experiencing employer-mandated switch from low-deductible plans to HDHP

*HDHP, high-deductible health plan

1. Before-after with control group: plot and estimates



Adjusted Relative DID* = -10% (-17% to -3%)

*DID, difference-in-differences

1. Before-after with control group: study design considerations

- Stability/frequency of outcome:
  - Emergency department visits relatively uncommon
- Duration of baseline/follow-up or number of baseline/follow-up points:
  - Sample size / power adequate only if restrict to 1 year before and after among “continuously enrolled”
- Control group availability:
  - Contemporaneous group with mandated low-deductible enrollment available

1. Before-after with control group: why use instead of ITS?

- If baseline trend can't be reliably captured
  - Rare events
  - Baseline time period available is too short
  - Outcome of interest is typically measured over a relatively long period; e.g. an annual basis not monthly basis
  - Data only available summarized over long periods
- Aid interpretation of ITS results

1. Before-after with control group: benefits

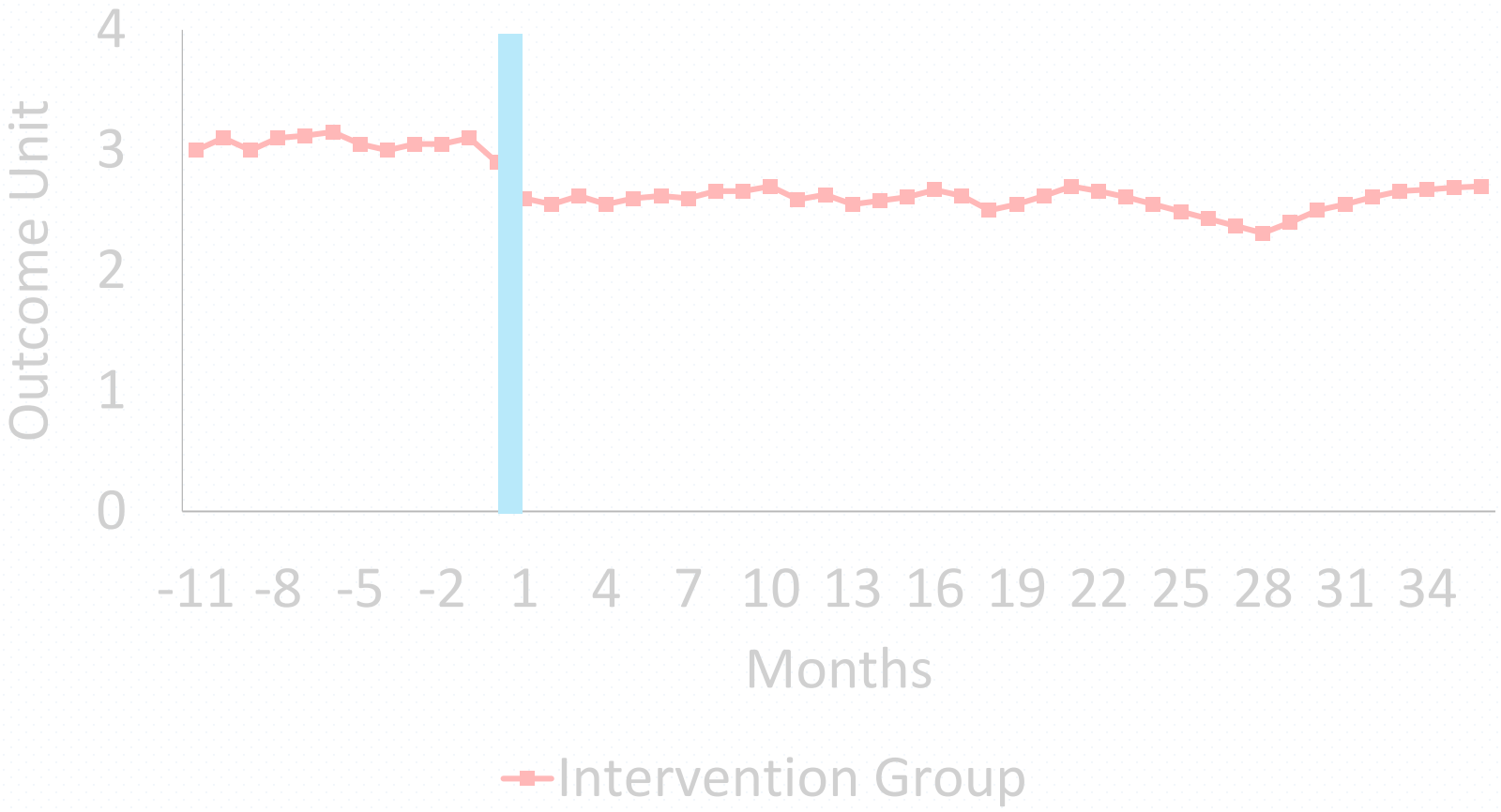
- Relatively simple to implement
- Effect estimates generally intuitive
- Can be rigorous if can prove that baseline levels and trends are similar
  - But then, why not use ITS?
- Ability to generate causal inference is better than cross-sectional and before-after without control group designs

Ryan AM *et al. Health Serv Res*. 2015.

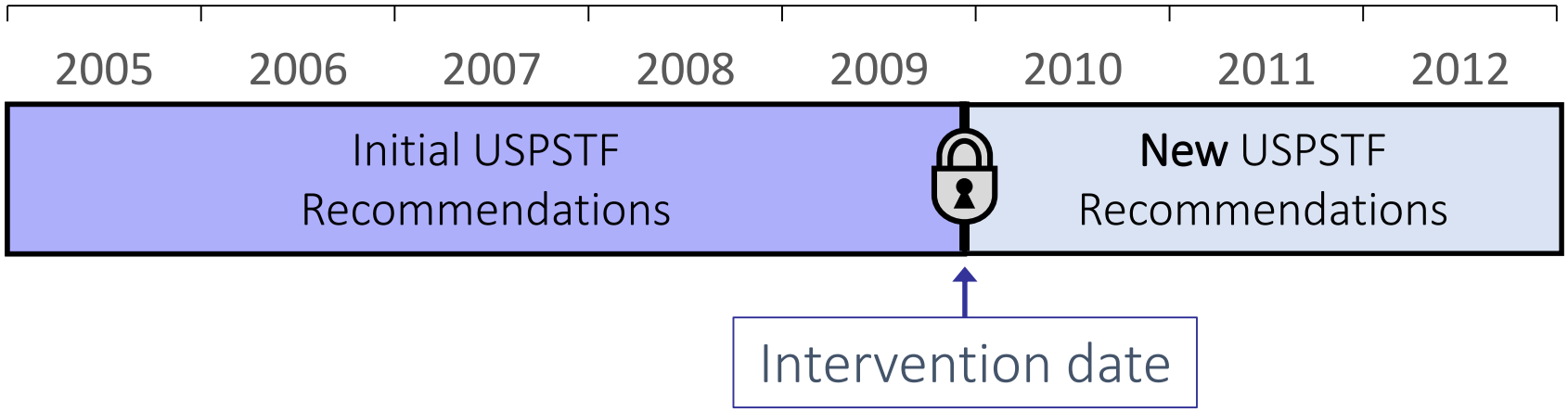
1. Before-after with control group: drawbacks

- Threats to validity and reliability as above
- But also includes the crucial potential problem of differing baseline trends
  - Hidden because summarizing outcomes over time
- Visual depiction of results not as strong as ITS
- Not recognized as highly rigorous by expert organizations such as Cochrane Collaboration

2. ITS without control group



2. ITS without control group: USPSTF/ mammography example

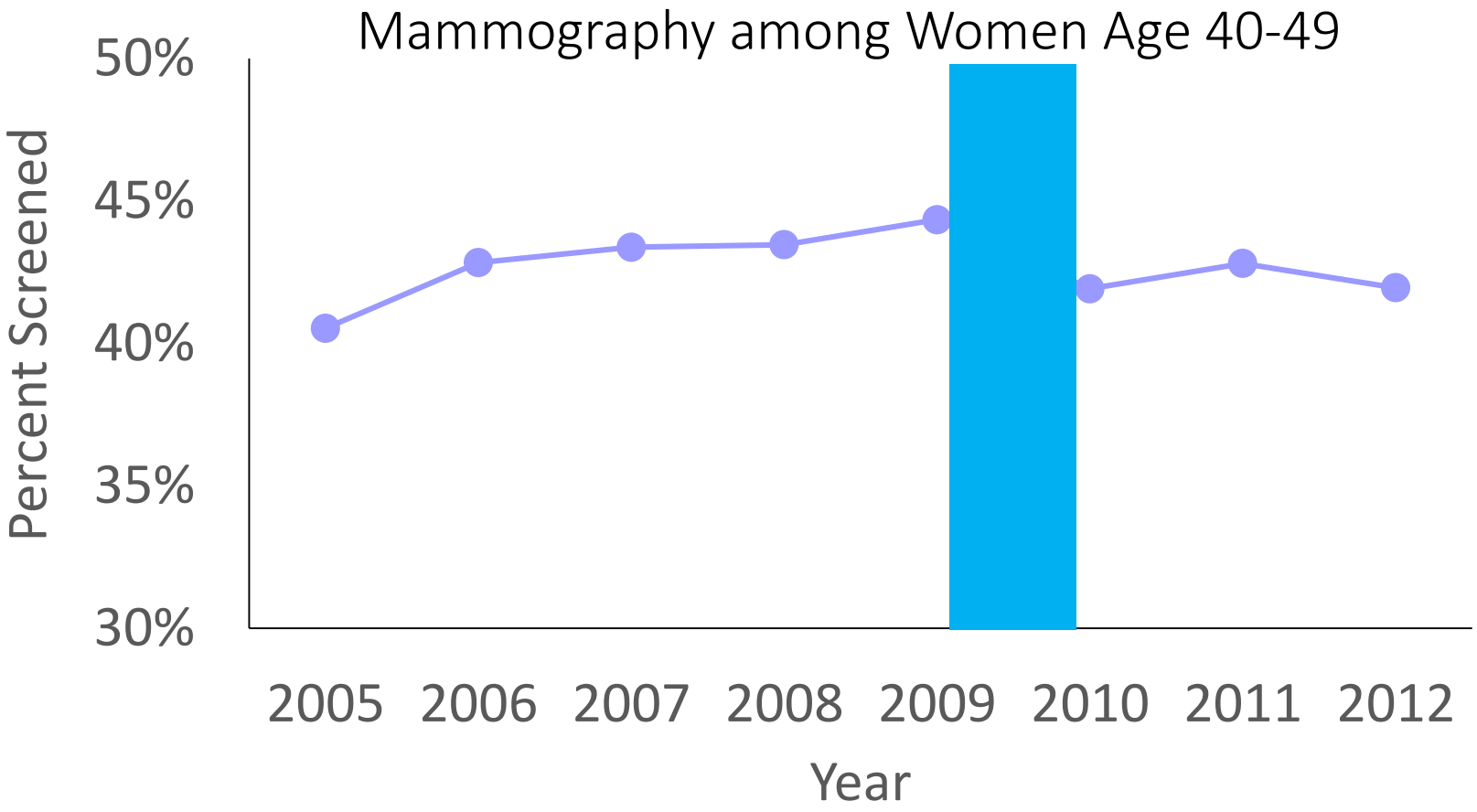


- Background: risks of annual screening mammography might outweigh benefits for women 40-49
- Natural experiment: highly publicized USPSTF guideline changes in 12/2009 for women 40-49; from q1-2 years to "personalized decision"

2. ITS without control group: study design considerations

- Stability/frequency of outcome:
  - Mammography common, but only on annual/biennial basis
- Duration of baseline/follow-up or number of baseline/follow-up points:
  - 5 baseline and 3 follow-up years: borderline too few
- Control group availability:
  - None in our dataset; none in U.S.
  - Could consider non-equivalent control group or control outcome

2. ITS without control group: plot and estimates



2012 Relative Difference = -10% (-10% to -9%)

Wharam JF *et al. J Clin Oncol*. 2015.

2. ITS without control group: why use instead of ITS with control group?

- No viable control group
  - National policy that affects everyone
  - Data not available
  - But sometimes can be clever and find a control group
- No control group needed

2. ITS without control group: benefits

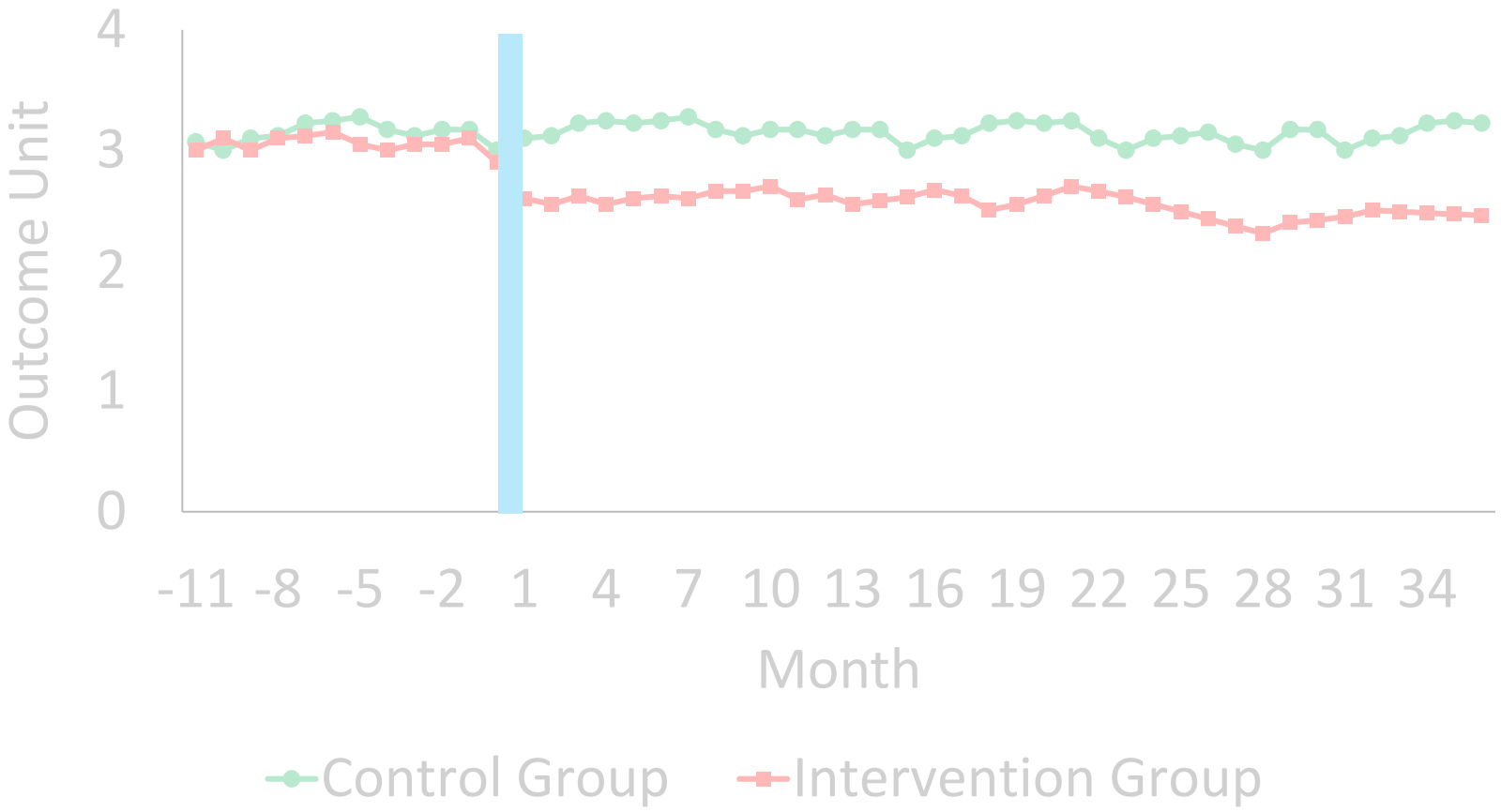
- Visual depiction of trend might be easy to interpret
- Can be rigorous if carefully constructed and / or used in the right setting
- Ability to generate causal inference is better than cross-sectional and before-after without control group designs

2. ITS without control group: drawbacks

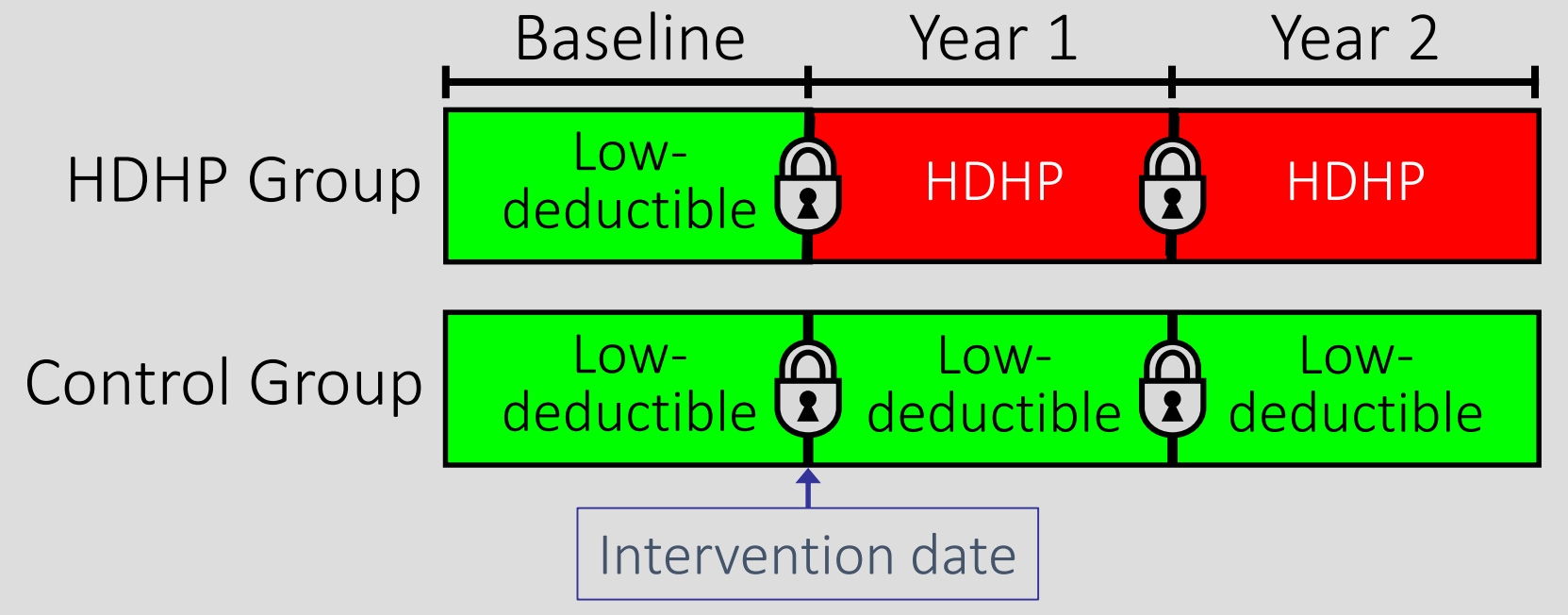
- Threats to validity and reliability as above
- In addition, greater concern that secular trends, floor/ceiling effects, history/maturation, or regression to the mean might be misinterpreted as causal effects
- Sensitive to points near end of segment
- Interpretation of parameter estimates: often not intuitive

Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. 2001.

3. ITS with control group



3. ITS with control group: HDHP/ DM care example



- Background: HDHPs might reduce expensive specialist visits among DM patients
- Natural experiment: DM patients experiencing employer-mandated switch from low-deductible plan to HDHP

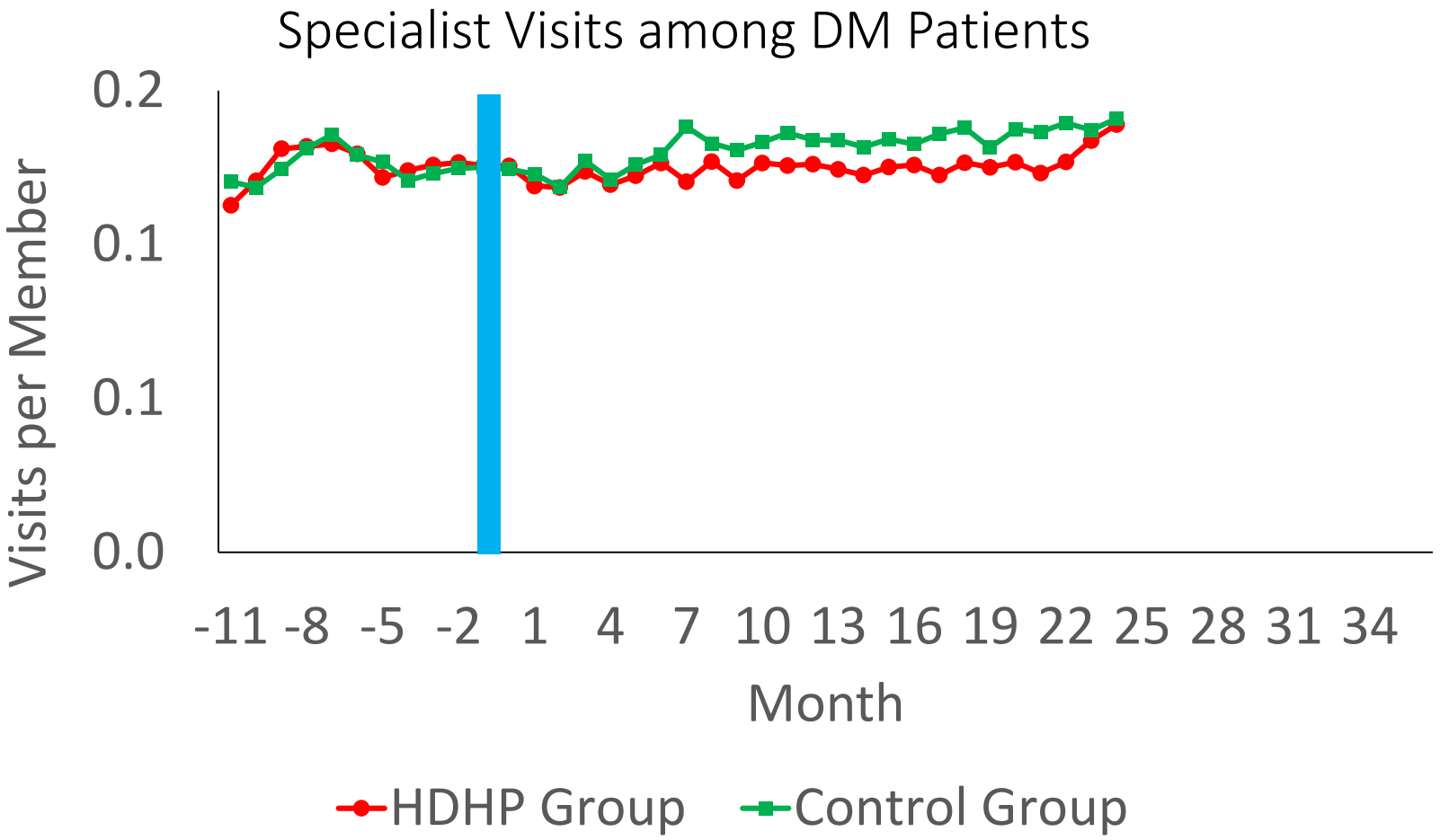
HDHP, high-deductible health plan

3. ITS with control group: study design considerations

- Stability/frequency of outcome:
  - Specialist visits relatively common in DM
- Duration of baseline/follow-up or number of baseline/follow-up points:
  - 12 baseline and 24 follow-up points among "continuously enrolled" yielded reasonable sample size
- Control group availability:
  - Contemporaneous group with mandated low-deductible enrollment available

St.Clair *et al*. *Am J Eval*. 2014.

3. ITS with control group: plot and estimates



Trend Δ: **-.0014 visits/mo.** (p<0.01); Trend ΔΔ: **.0005 visits/mo.^{2}** (p<0.01)

3. ITS with control group: benefits?

- Strongest "quasi-experimental" design
  - Research has demonstrated that well-constructed controlled ITS designs yield similar effect estimates to gold standard randomized controlled trials
  - More rigorous than above designs
  - Able to generate causal inference
- Visual depiction of results can be highly valuable
  - A picture is worth…
- Less expensive and potentially more generalizable than RCTs

Hallberg K. and Cook TD. http://teams.mspnet.org/media/data/TEAMSMayWebinar.pdf?media_000000008372.pdf

3. ITS with control group: drawbacks

- Threats to validity and reliability as above
- Sensitive to points near end of segment
- Can be complex to implement; devil in the details
- Interpretation of parameter estimates: often not intuitive

Summary: 3 key factors for deciding on a study design for natural experiments

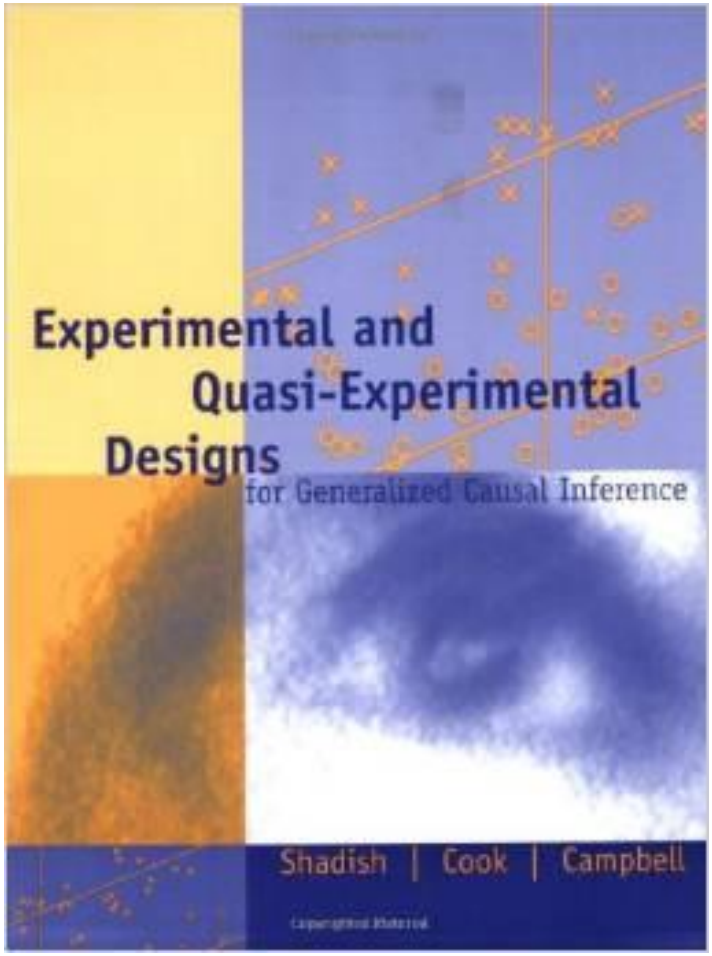
1. Stability/frequency of outcome
2. Duration of baseline/follow-up or number of baseline/follow-up points
3. Control group availability

Takeaways

- Study designs differ in their ability to generate causal inference
- In studying natural experiments, typically aim for ITS with control group
  - But other designs can sometimes be sufficient, necessary, or almost as rigorous
- Even if control group not available or immediately obvious, controlled study might be possible
- Can use ITS for pictures and to prove rigor and DID analysis to convey results intuitively

Final thoughts

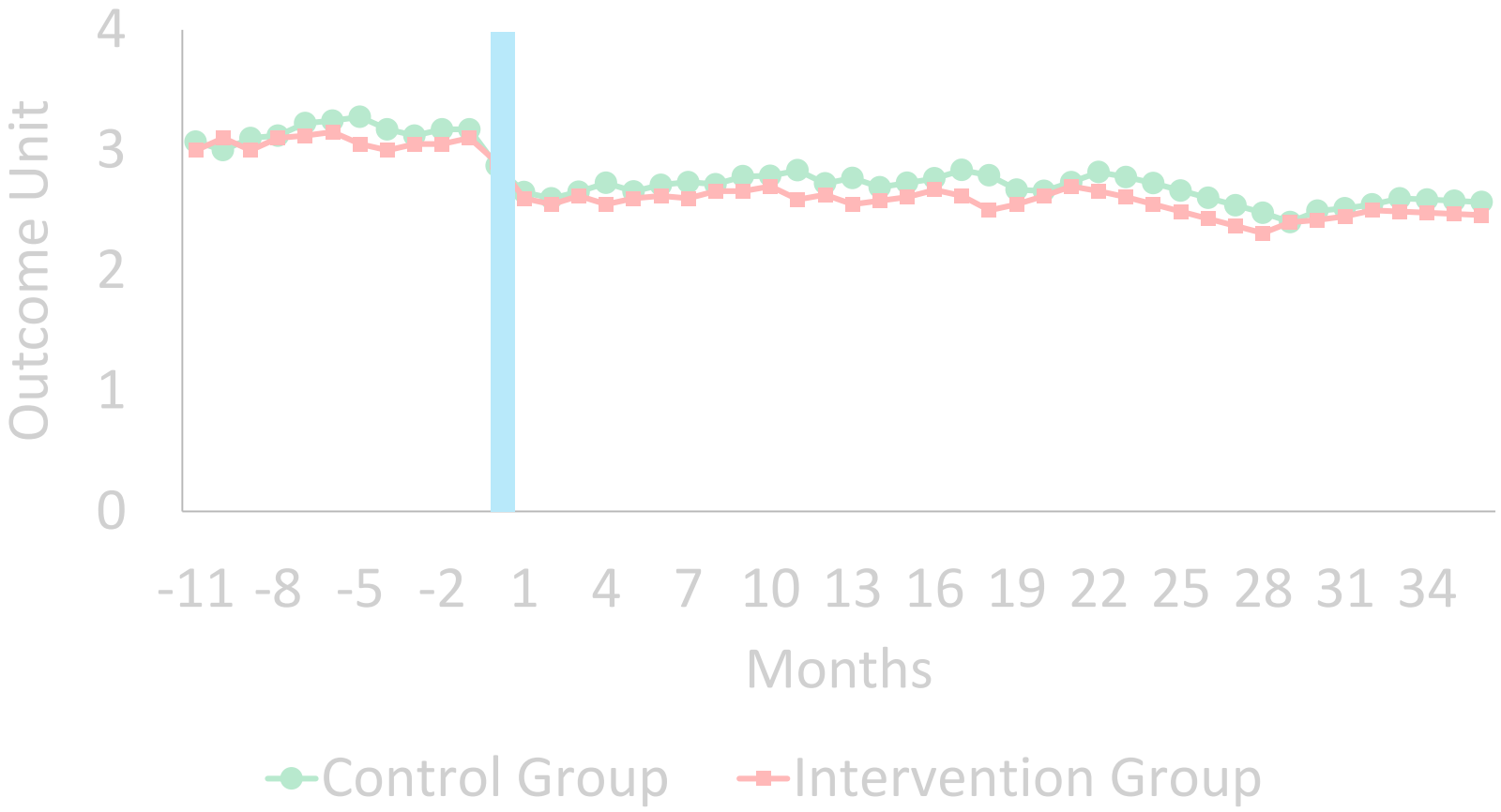
- Natural experiments can be implemented at many levels of aggregation
- Routinely collected data offer potential for robust evaluation
- ITS is strongest quasi-experimental design
- Methods exist to strengthen sample design and statistical analysis



Thank you!

KDuru@mednet.ucla.edu

3b. ITS with control group, also matched on baseline trend



3b. ITS with control group, also matched on baseline trend: basics

- Reasons for special role of baseline outcomes in matching:
  - High correlation with follow-up outcomes
  - Likely correlation with selection

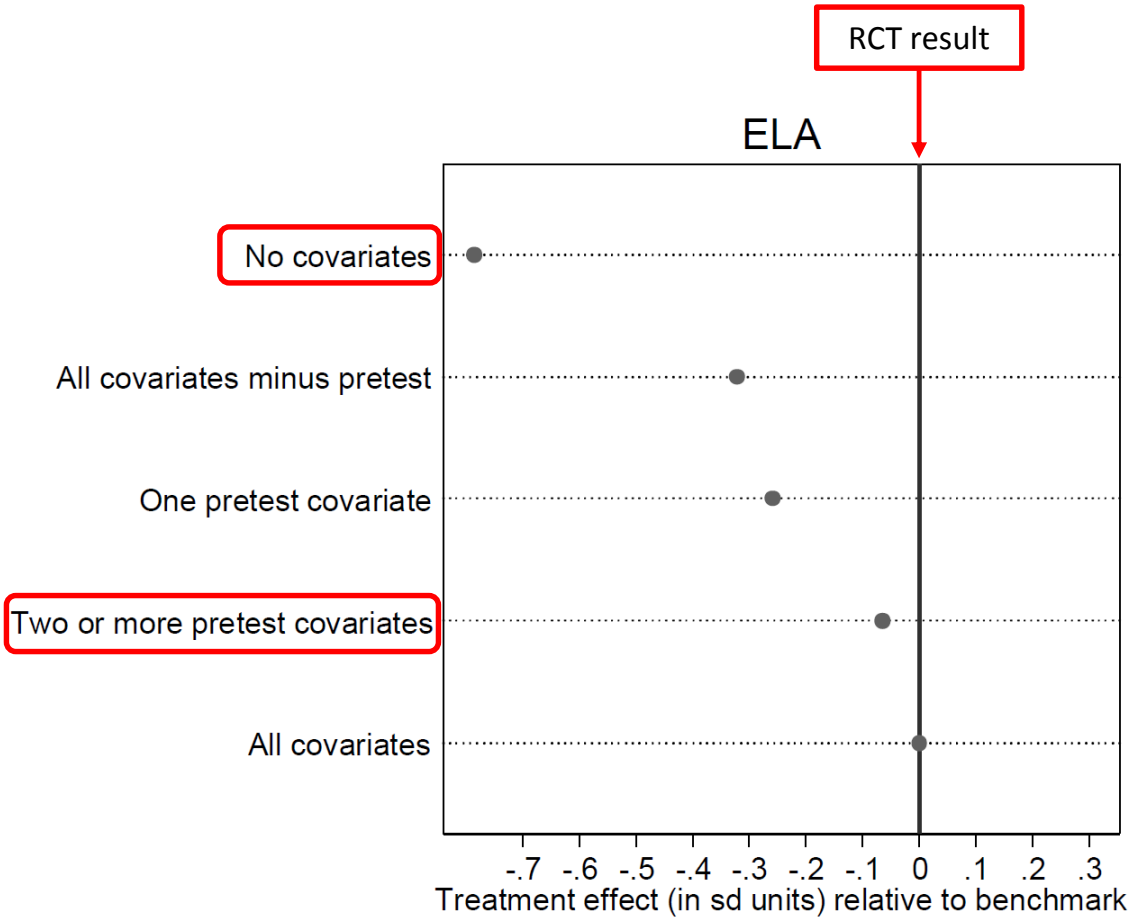
(1) Hallberg K. Cook TD. http://teams.mspnet.org/media/data/TEAMSMayWebinar.pdf?media_000000008372.pdf (2) St.Clair *et al*. *Am J Eval*. 2014.

3b. ITS with control group, also matched on baseline trend: approach

- Can simply add baseline outcome measure for each time period to the propensity score match logistic regression
- Forces intervention and control groups to have similar baseline trends

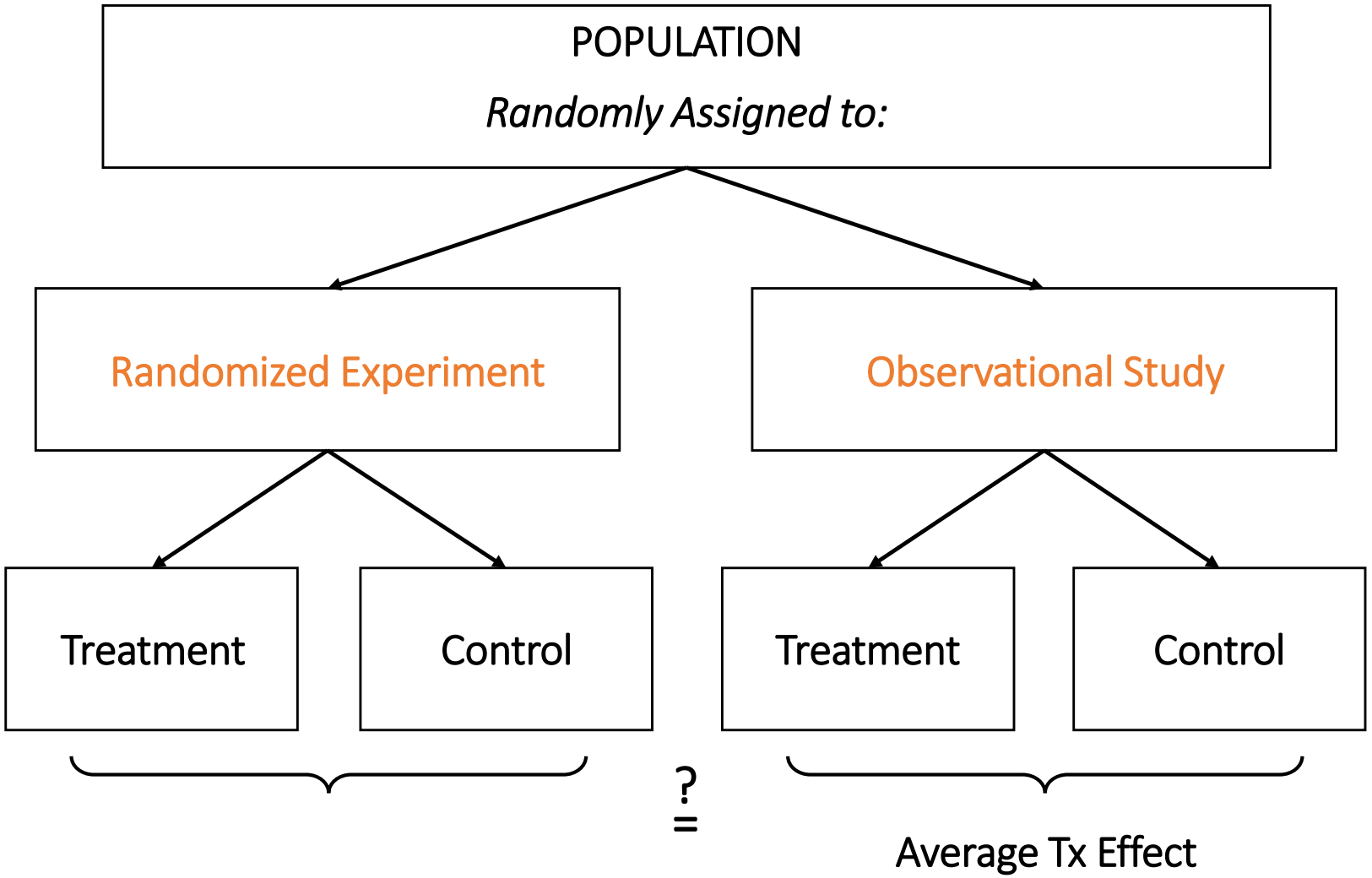
(1) Hallberg K. Cook TD. http://teams.mspnet.org/media/data/TEAMSMayWebinar.pdf?media_000000008372.pdf (2) St.Clair *et al*. *Am J Eval*. 2014.

Bias reduction: matching on only baseline outcomes



Hallberg K. and Cook TD. http://teams.mspnet.org/media/data/TEAMSMayWebinar.pdf?media_000000008372.pdf

Within-study comparison design



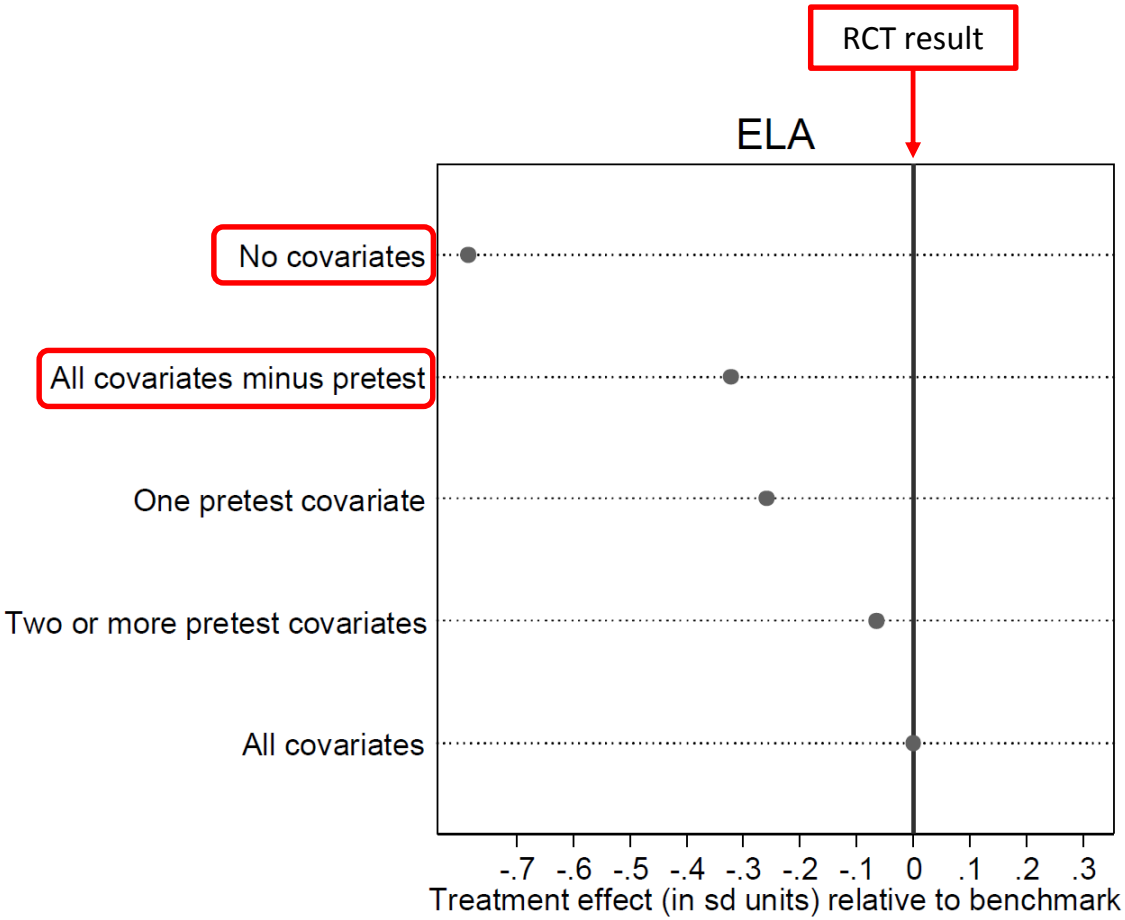
Hallberg K. and Cook TD. http://teams.mspnet.org/media/data/TEAMSMayWebinar.pdf?media_000000008372.pdf

Within-study comparison studies

- Previous within-study comparison studies have shown that ITS with control group can replicate RCT results
- More recently, St. Clair, Cook, and Hallberg have conducted the most rigorous studies to date
  - 4 arm randomized approach as on previous slide
  - In addition to determining if ITS with control group designs yield results similar to RCTs, investigators are determining optimal ITS matching approaches
  - Education setting; not medical setting

(1) St. Clair T *et al*. *Am J Eval*. 2014. (2) Schneeweiss *et al*. *BMJ*. 2004. (3) Fretheim *et al.* *J Clin Epi*. 2013. (4) Somers M *et al. MDRC.* 2013; Ferraro PJ, and Miranda JJ. *Resource Energy Econ*. 2013.

Bias reduction: matching on covariates except baseline outcomes



Hallberg K. and Cook TD. http://teams.mspnet.org/media/data/TEAMSMayWebinar.pdf?media_000000008372.pdf