

5 Causal Inference Ideas for the Age of Vibes

Alex Molak



What Are Experiments Really Testing?



The NEW ENGLAND
JOURNAL of MEDICINE

CURRENT ISSUE ▾ SPECIALTIES ▾ TOPICS ▾

ORIGINAL ARTICLE

f X in ✉

Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis

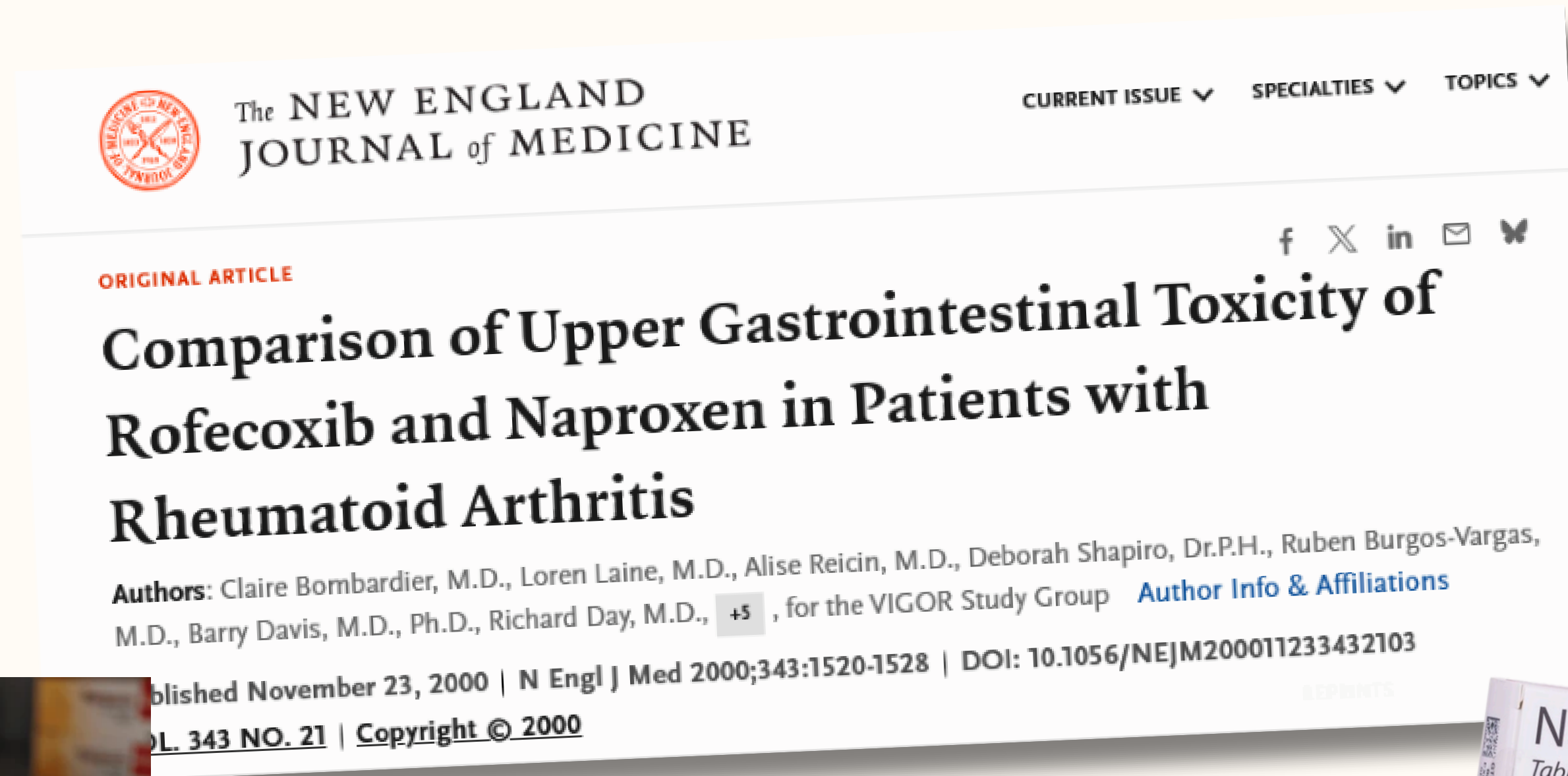
Authors: Claire Bombardier, M.D., Loren Laine, M.D., Alise Reicin, M.D., Deborah Shapiro, Dr.P.H., Ruben Burgos-Vargas, M.D., Barry Davis, M.D., Ph.D., Richard Day, M.D., [+5](#), for the VIGOR Study Group [Author Info & Affiliations](#)

Published November 23, 2000 | N Engl J Med 2000;343:1520-1528 | DOI: 10.1056/NEJM200011233432103

[REPRINTS](#)

VOL. 343 NO. 21 | Copyright © 2000

What Are Experiments Really Testing?



- Rofecoxib indeed decreased the no. of gastrointestinal events compared to Naproxen

- Rofecoxib indeed decreased the no. of gastrointestinal events compared to Naproxen
- Despite this fact, it was withdrawn from the market

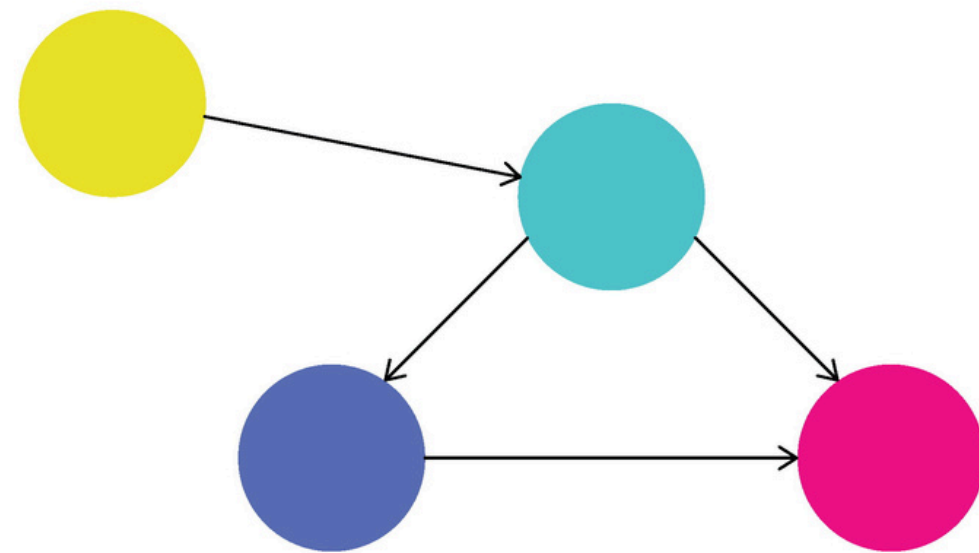
- Rofecoxib indeed decreased the no. of gastrointestinal events compared to Naproxen
- Despite this fact, it was withdrawn from the market
- ...after causing an estimated **80,000** to **140,000** excess cases of serious coronary heart disease, **44%** of which were **fatal**.



⟨packt⟩

Causal Inference and Discovery in Python

Unlock the secrets of modern causal machine learning
with DoWhy, EconML, PyTorch and more



ALEKSANDER MOLAK

Foreword by Ajit Jaokar, Course Director: Artificial Intelligence at University of Oxford





Agenda

- **What Are Experiments Really Testing?**

Agenda

- What Are Experiments Really Testing?
- **Statistical Significance vs Evidence**

Agenda

- What Are Experiments Really Testing?
- Statistical Significance vs Evidence
- **Generalization vs Collider Bias**

Agenda

- What Are Experiments Really Testing?
- Statistical Significance vs Evidence
- Generalization vs Collider Bias
- **Is Double Machine Learning Magic?**

Agenda

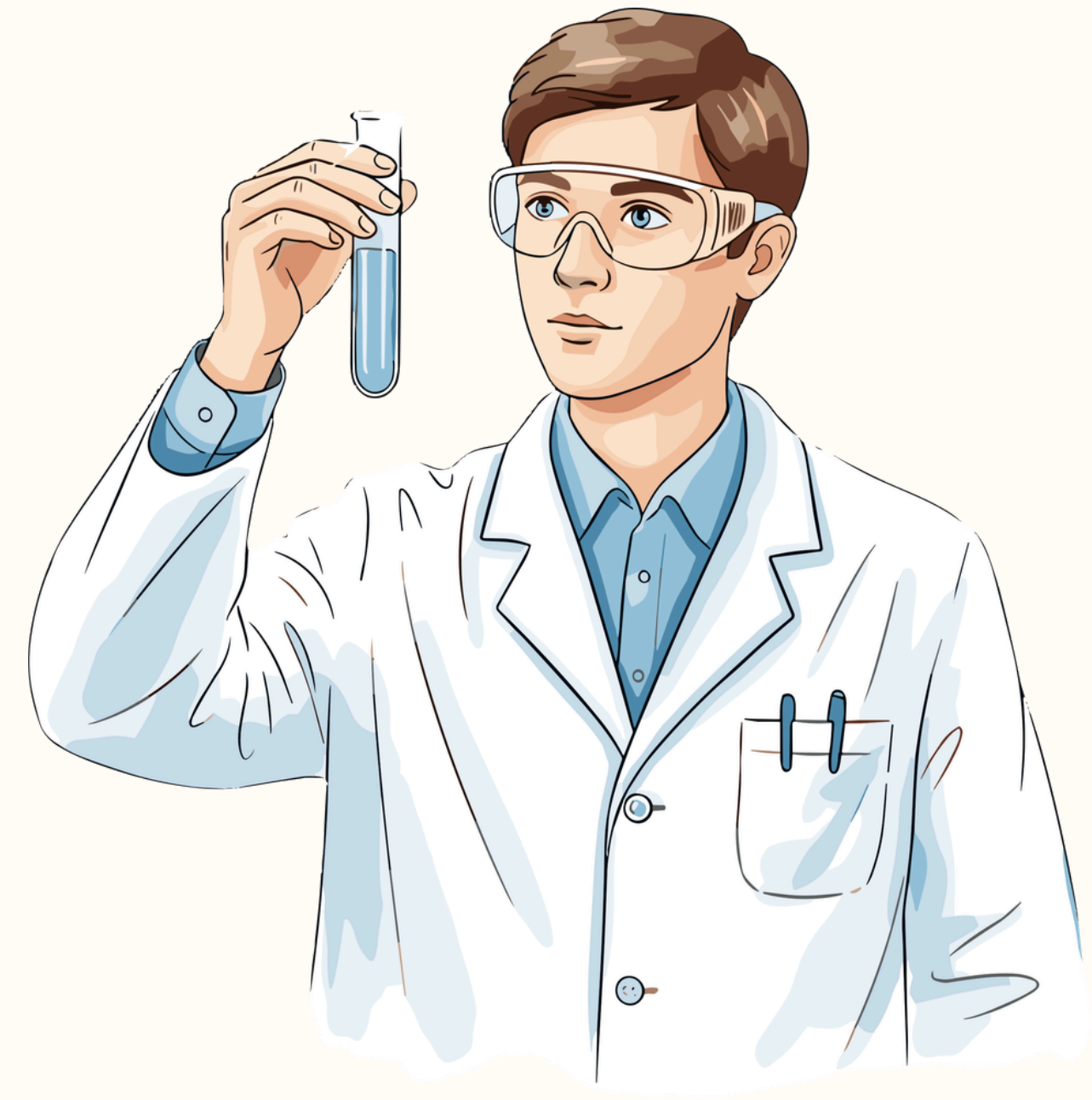
- What Are Experiments Really Testing?
- Statistical Significance vs Evidence
- Generalization vs Collider Bias
- Is Double Machine Learning Magic?
- **Three Ways to Work with Hidden Confounding**

Let's Start!

What Are Experiments Really Testing?

What Are Experiments Really Testing?

Avi is a young scientist and he wants to test a new supplement that he hopes will **speed up recovery** after strenuous exercise.



What Are Experiments Really Testing?

He decides to give the supplement to his sister **Grace** and his cousin **David** after their Friday training session at the gym.



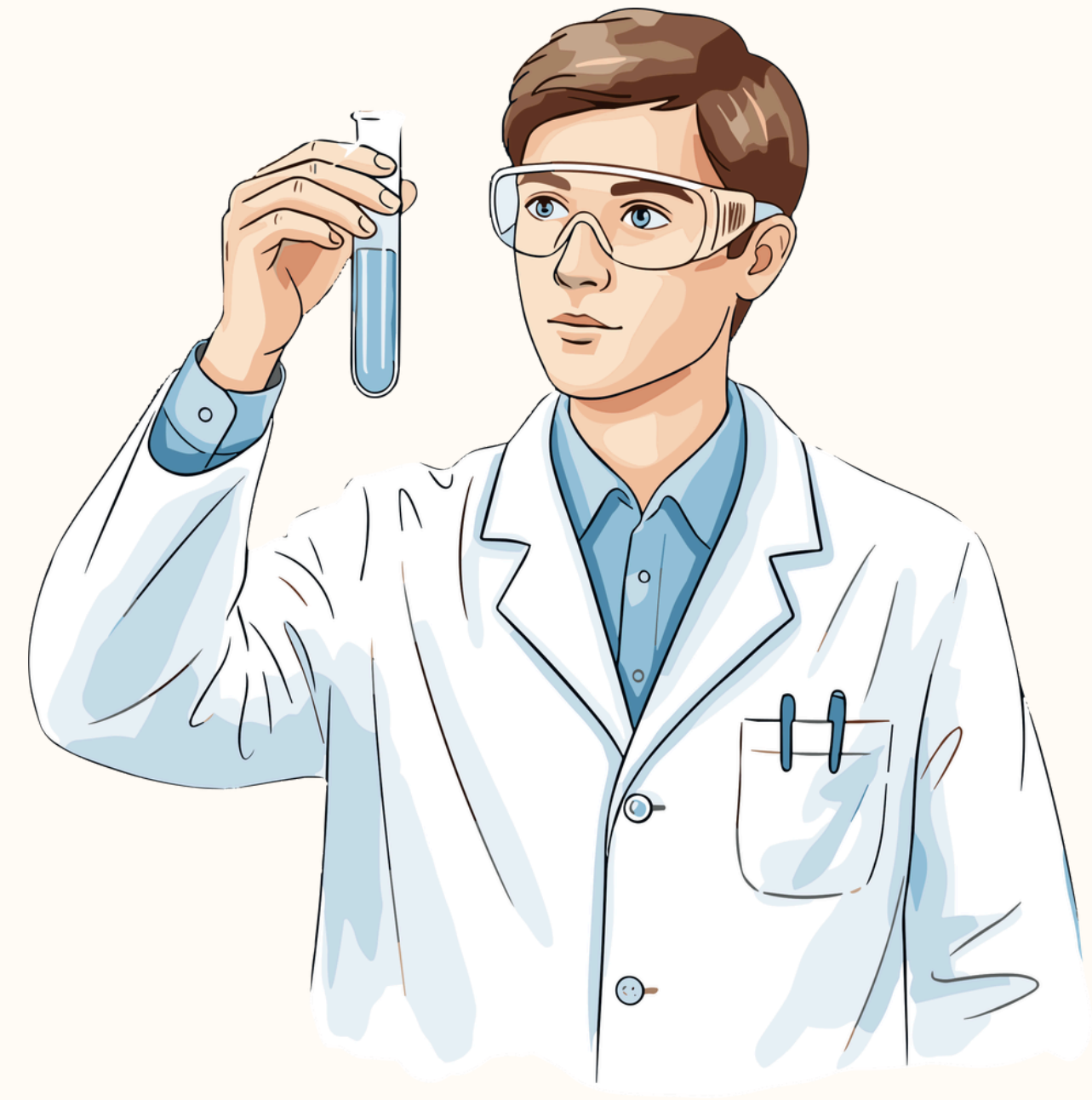
What Are Experiments Really Testing?

At the same time, he does not give the formula to his colleagues **Jill** and **Monica**.



What Are Experiments Really Testing?

Avi measures fatigue symptoms in all four over the weekend, and compares the results on Monday.



What Are Experiments Really Testing?



The results are consistent with Avi's hypothesis: **Grace** and **David** showed **less fatigue** compared to **Jill** and **Monica**.

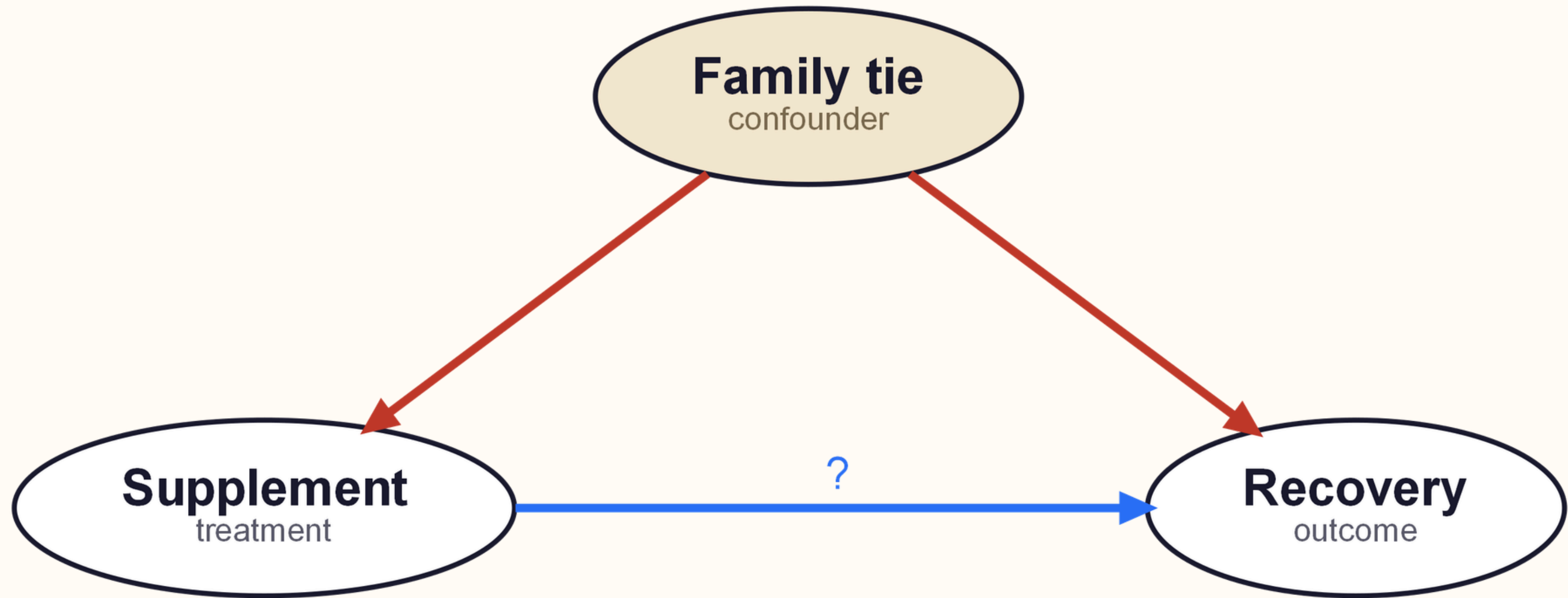


What Are Experiments Really Testing?

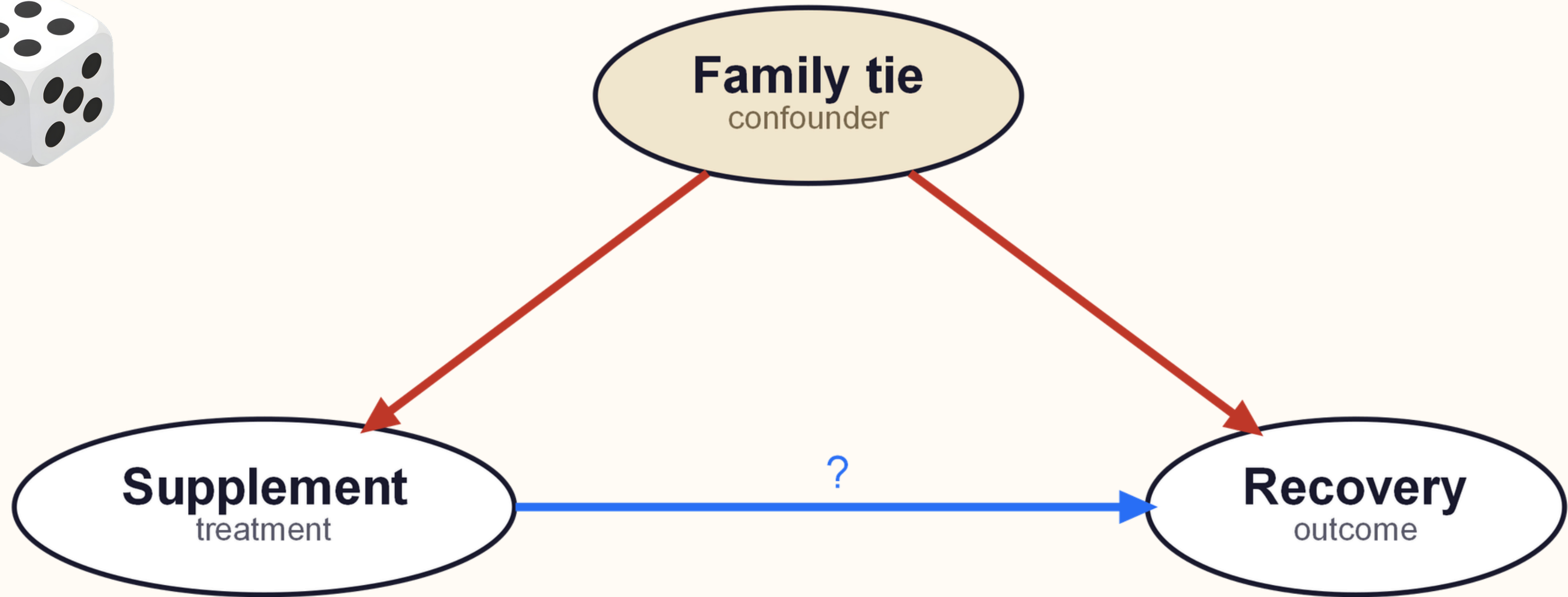
Is Avi's test **a good test** of the hypothesis that his supplement speeds up recovery (reduces fatigue symptoms) after strenuous exercise?



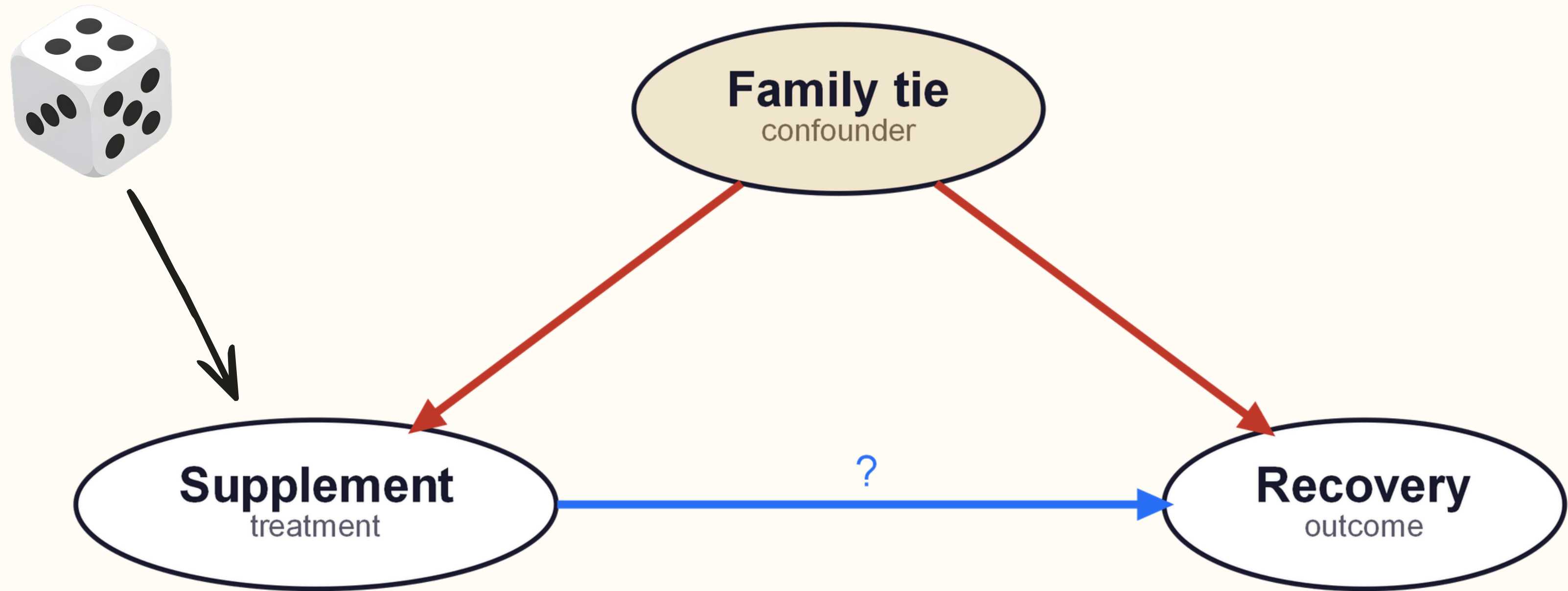
What Are Experiments Really Testing?



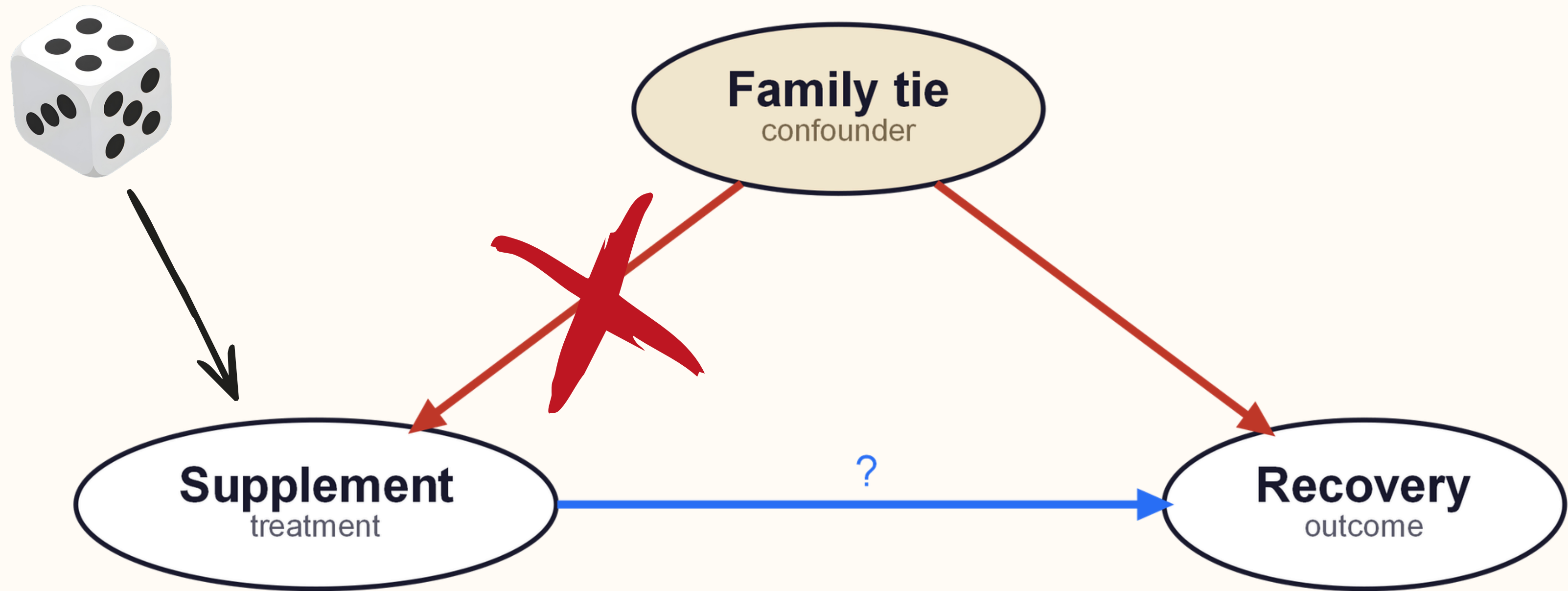
What Are Experiments Really Testing?



What Are Experiments Really Testing?



What Are Experiments Really Testing?



What Are Experiments Really Testing?

- Randomization helps us reduce the **bias** coming **from confounding** (obtain causal identification)

What Are Experiments Really Testing?

- Randomization helps us reduce the **bias** coming **from confounding** (obtain causal identification)
- From a graphical perspective, it **removes edges** from any **observed** or **unobserved** variables **affecting the treatment** variable

What Are Experiments Really Testing?

- Randomization helps us reduce the **bias** coming **from confounding** (obtain causal identification)
- From a graphical perspective, it **removes edges** from any **observed** or **unobserved** variables **affecting the treatment** variable
- It does **not change** the **edges** from variables **impacting** the **outcome**

What Are Experiments Really Testing?

Is randomization a silver bullet?



What Are Experiments Really Testing?



What Are Experiments Really Testing?

- Randomization is **helpful** for **causal identification** but it's **not magic**

What Are Experiments Really Testing?

- Randomization is **helpful** for **causal identification** but it's **not magic**
- Like any causal tool, it operates under **assumptions** (e.g., SUTVA)

What Are Experiments Really Testing?

- Randomization is **helpful** for **causal identification** but it's **not magic**
- Like any causal tool, it operates under **assumptions** (e.g., SUTVA)
- It **balances covariates**, but only in **expectation** – in any single trial they might not be balanced (but that's ok, our standard analyses take this into account)

What Are Experiments Really Testing?

- A randomized controlled trial (A/B test) is designed to:

What Are Experiments Really Testing?

- A randomized controlled trial (A/B test) is designed to:
 - Answer **interventional questions** (does X change Y?)

What Are Experiments Really Testing?

- A randomized controlled trial (A/B test) is designed to:
 - Answer **interventional questions** (does X change Y?)
 - On **average**** (between two groups of observations)***

** We can test for both ATE and CATE using randomized data, but we often encounter limitations related to statistical power in the context of subgroup analysis

*** N-of-1 trials are possible under some assumptions and can be useful, then the unit of randomization is one person indexed over time

What Are Experiments Really Testing?

- A randomized controlled trial (A/B test) is designed to:
 - Answer **interventional questions** (does X change Y?)
 - On **average**** (between two groups of observations)***
 - In a population described by the sample

** We can test for both ATE and CATE using randomized data, but we often encounter limitations related to statistical power in the context of subgroup analysis

*** N-of-1 trials are possible under some assumptions and can be useful, then the unit of randomization is one person indexed over time

What Are Experiments Really Testing?

- A randomized controlled trial (A/B test) is designed to:
 - Answer **interventional questions** (does X change Y?)
 - On **average**** (between two groups of observations)***
 - In a population described by the sample
 - It **generalizes** to the **future** only under the assumption that the underlying **data generating process** is **stable over time**

** We can test for both ATE and CATE using randomized data, but we often encounter limitations related to statistical power in the context of subgroup analysis

*** N-of-1 trials are possible under some assumptions and can be useful, then the unit of randomization is one person indexed over time

Statistical Significance vs Evidence

Statistical Significance vs Evidence

My **A/B test** is **significant** (a winner?), now what?



Statistical Significance vs Evidence

- **What does a significant result really mean?**

Statistical Significance vs Evidence

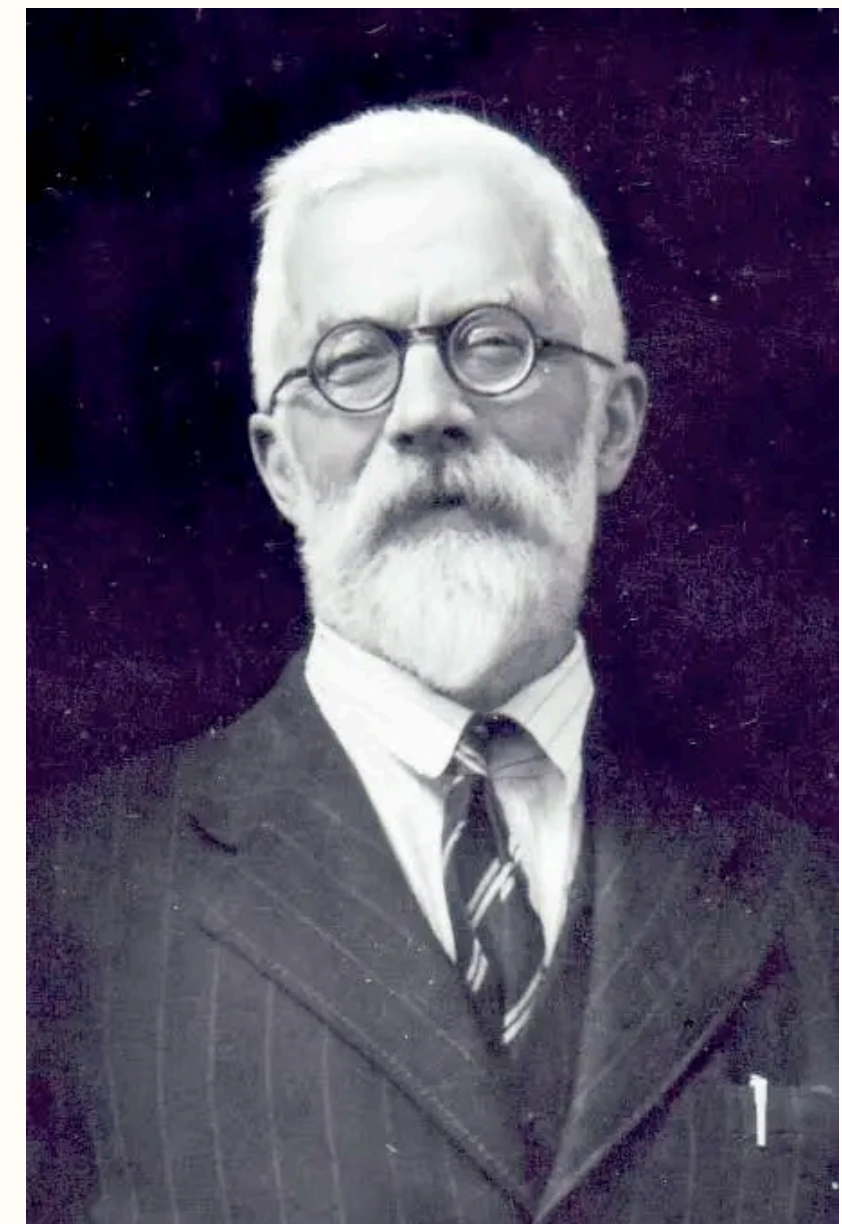
- What does a significant result really mean?
- We use **p-values** (or confidence intervals) to determine significance

Statistical Significance vs Evidence

- What does a significant result really mean?
- We use **p-values** (or confidence intervals) to determine significance
- P-value is a **frequentist term** → it refers to the probability - over **hypothetical** experiment **repetitions** - of observing the data as **extreme**, or more extreme than we observed **given the null is true**

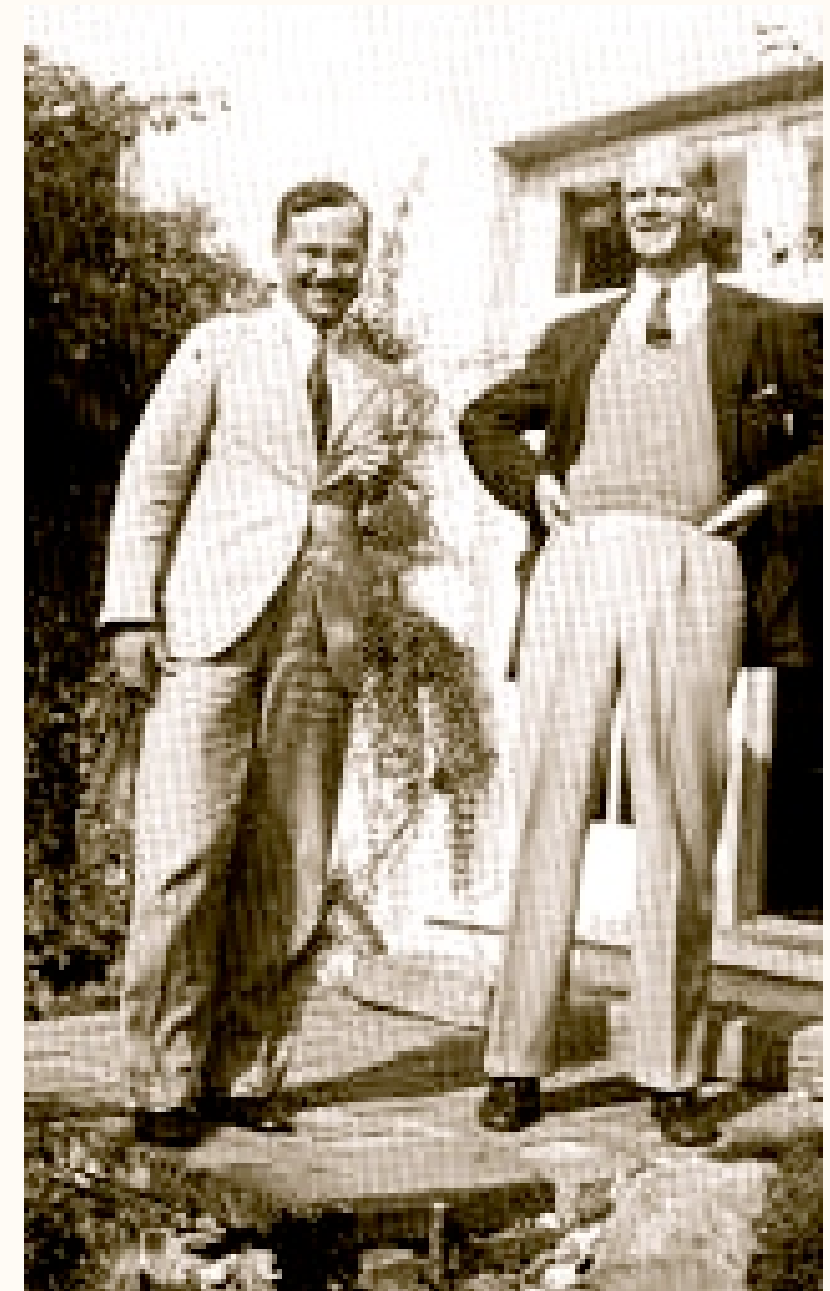
Statistical Significance vs Evidence

- What does a significant result really mean?
- We use **p-values** (or confidence intervals) to determine significance
- P-value is a **frequentist term** → it refers to the probability – over **hypothetical** experiment **repetitions** – of observing the data as **extreme**, or more extreme than we observed **given the null is true**



Statistical Significance vs Evidence

- What does a significant result really mean?
- We use **p-values** (or confidence intervals) to determine significance
- P-value is a **frequentist term** → it refers to the probability – over **hypothetical** experiment **repetitions** – of observing the data as **extreme**, or more extreme than we observed **given the null is true**



Statistical Significance vs Evidence

- But that's not all...

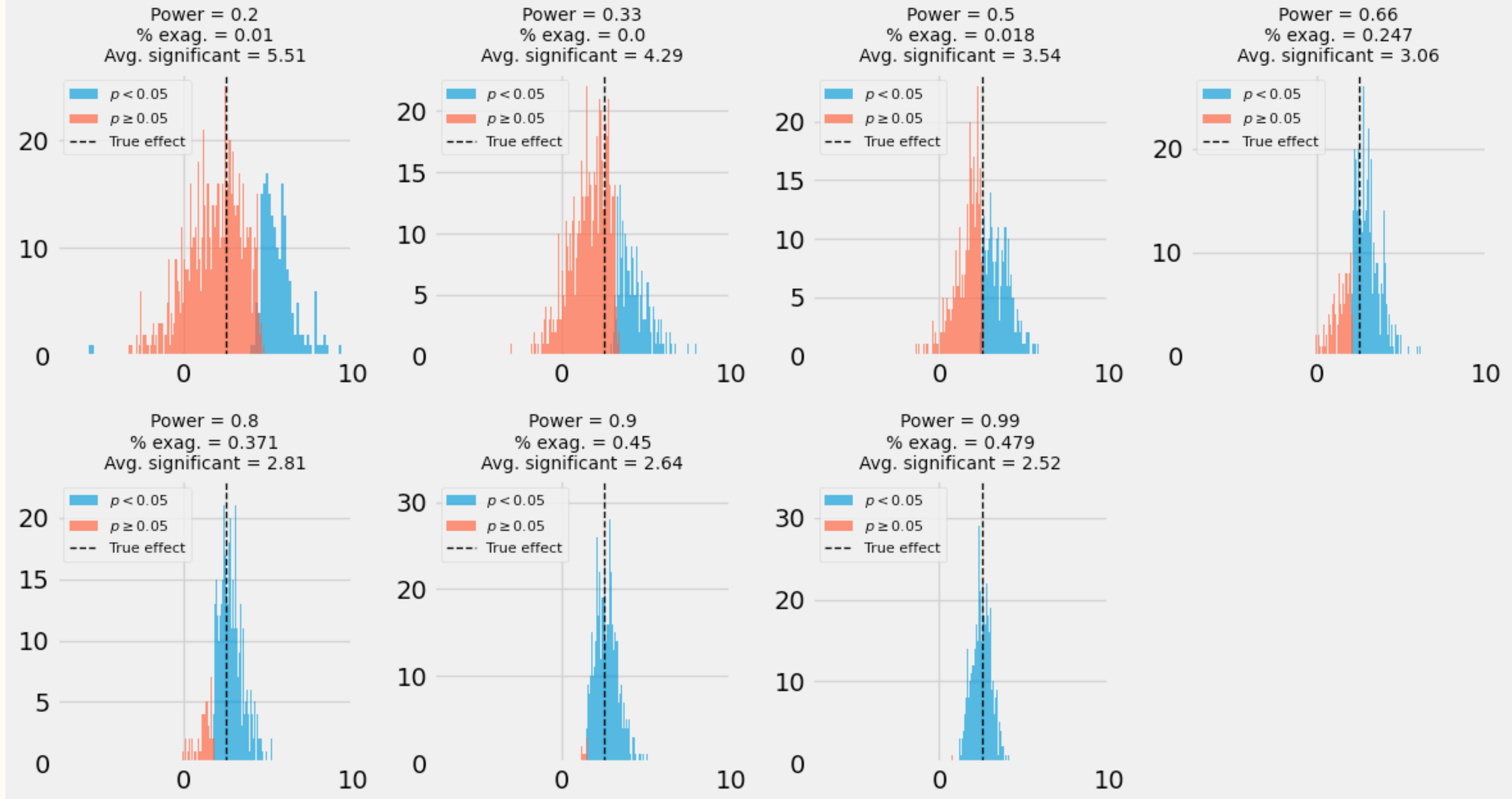
Statistical Significance vs Evidence

- But that's not all...
- In an **underpowered study**, a significant result might not be meaningful or might be misleading

- But that's not all...
- In an **underpowered study**, a significant result might not be meaningful or might be misleading
- It is likely to be **exaggerated** (Type-M error) and can have a **reversed sign** (Type-S error)

Statistical Significance vs Evidence

Distribution of estimated effects over 1000 trials per sample size



Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors

Andrew Gelman¹ and John Carlin^{2,3}

¹Department of Statistics and Department of Political Science, Columbia University; ²Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Victoria, Australia; and ³Department of Paediatrics and School of Population and Global Health, University of Melbourne

What Are Experiments Really Testing?

“Gold Standard” **!=** Silver Bullet



What Are Experiments Really Testing?

“Gold Standard” **!=** Silver Bullet

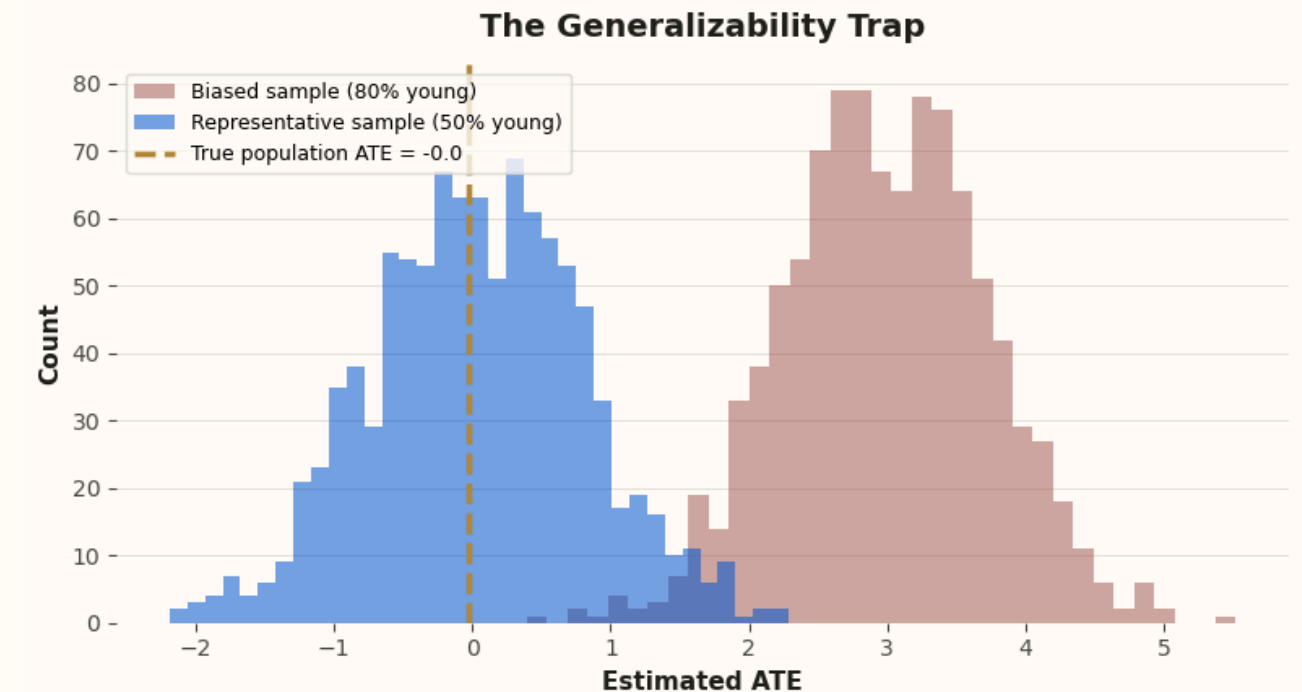
- Randomization is incredibly useful for obtaining causal identification for interventional queries, but...



What Are Experiments Really Testing?

“Gold Standard” $\neq$ Silver Bullet

- Randomization is incredibly useful for obtaining causal identification for interventional queries, but...
- Randomization does **not guarantee generalization**



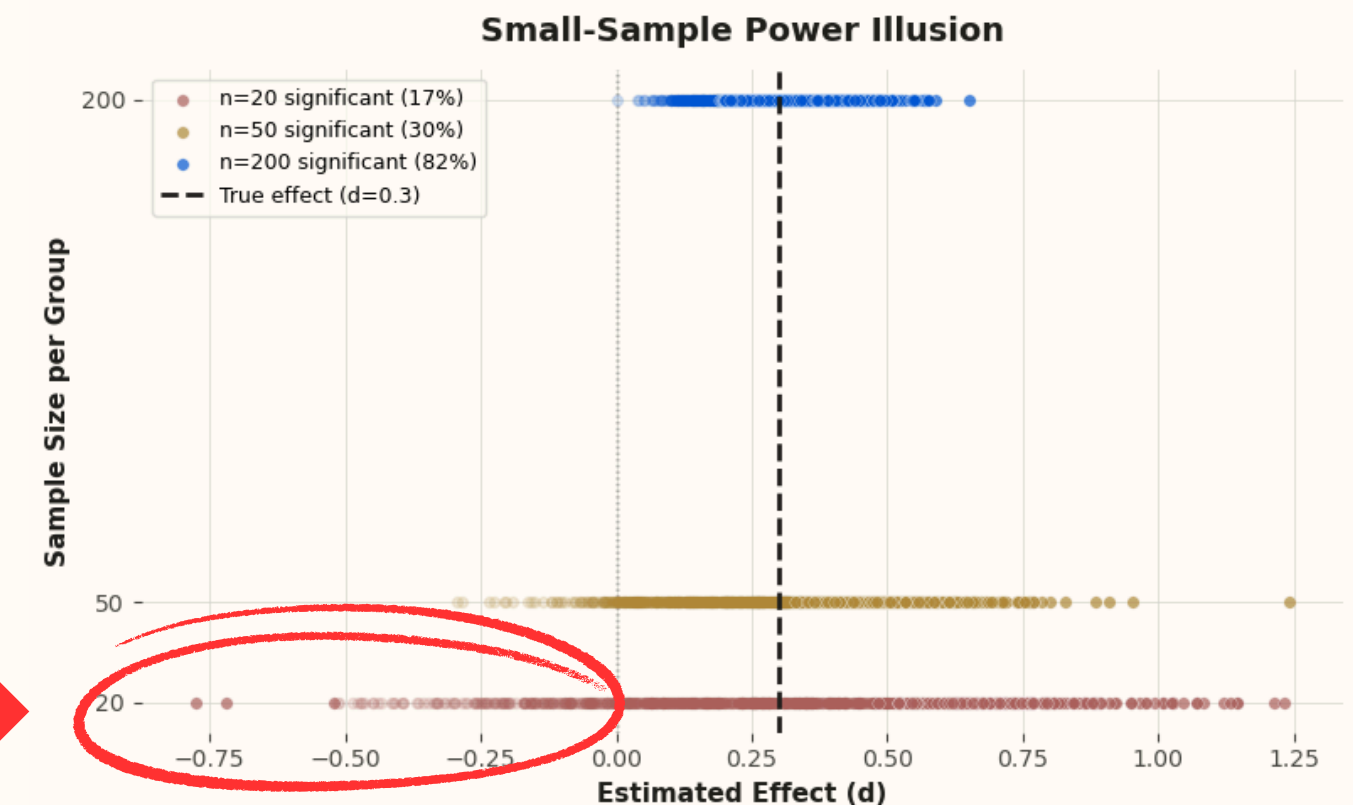
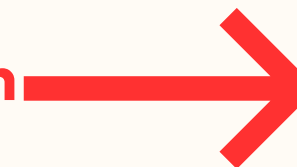
What Are Experiments Really Testing?

“Gold Standard” != Silver Bullet

- Randomization is incredibly useful for obtaining causal identification for interventional queries, but...
- Randomization does not guarantee generalization
- **Underpowered** RCTs can lead to **exaggerated** or **reversed** effects



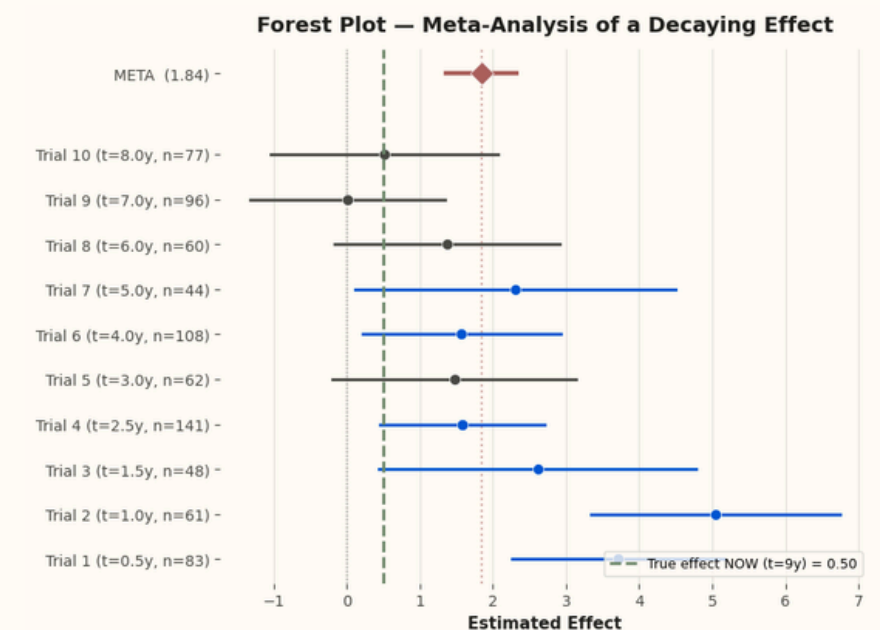
Results with reversed sign



What Are Experiments Really Testing?

“Gold Standard” != Silver Bullet

- Randomization is incredibly useful for obtaining causal identification for interventional queries, but...
- Randomization does not guarantee generalization
- Underpowered RCTs can lead to exaggerated or reversed effects
- RCTs only generalize when the **data generating process is stable**



What Are Experiments Really Testing?



Multi-Armed Bandits, Thompson Sampling, or A/B Testing? Are you optimizing for short-term headlines or long-term pills worth billions?



Ron Kohavi 

Vice President and Technical Fellow | Data Science, Engineering | AI, Machine Learning, Controlled Experiments | Ex-Airbnb, Ex-Microsoft...

June 17, 2023



What Are Experiments Really Testing?



Multi-Armed Bandits, Thompson Sampling, or A/B Testing? A
optimizing for short-term
long-term pills worth billions?



Ron Kohavi 

Vice President and Technical Fellow | Data Science, Engineering | AI,
Machine Learning, Controlled Experiments | Ex-Airbnb, Ex-Microsoft...

June 17, 2023

If you're optimizing for the short-term, and need to optimize WHILE THE EXPERIMENT IS RUNNING, use multi-armed bandit (MAB) approaches.

If you're optimizing for the long-term, to gain knowledge and insight, or to find a better treatments/features that will be launched and help you prioritize the backlog and improve the roadmap, then two-week experiment time is dwarfed by the long-term impact, then A/B testing is more appropriate.

Generalization vs Collider Bias

Generalization vs Collider Bias

Imagine Allegro has a **hiring policy** for developers such that a candidate needs to score at least **8/10** on a sum of **hard** and **soft** skills assessments



Generalization vs Collider Bias

Assuming for simplicity that hard and soft skills are independent in the general population, Allegro's **policy** would produce a **sub-population of developers** where more technically skilled people would be – on average – less socially skilled and vice versa – a popular stereotypical correlation pattern.

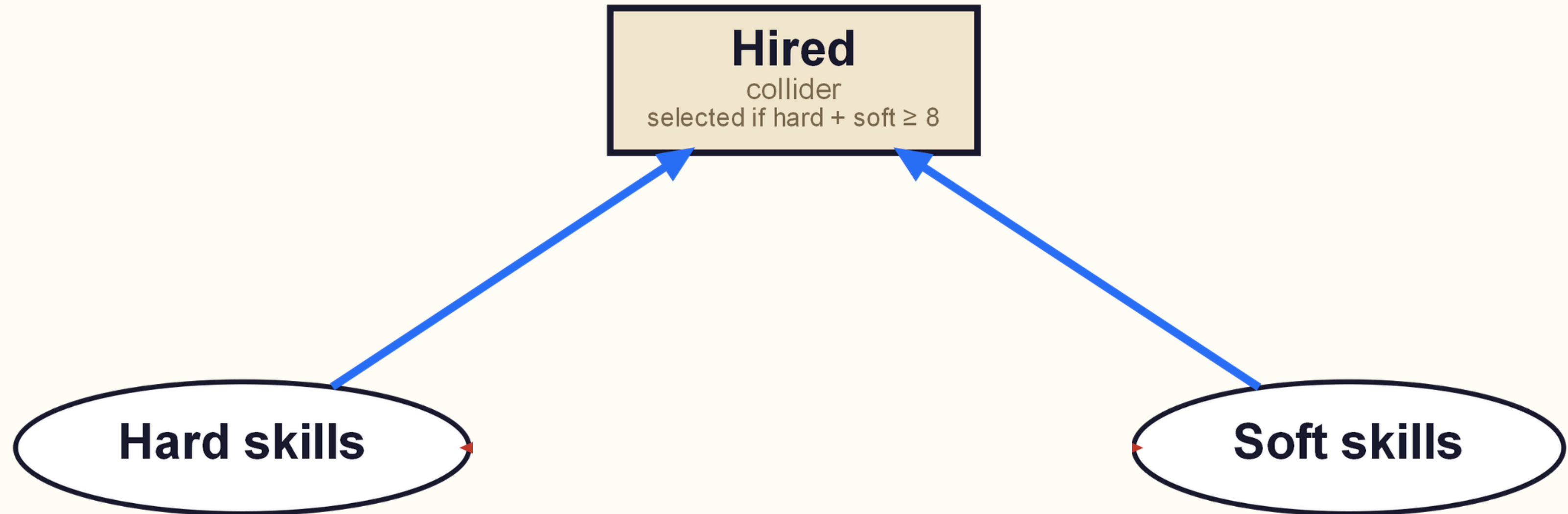


Generalization vs Collider Bias

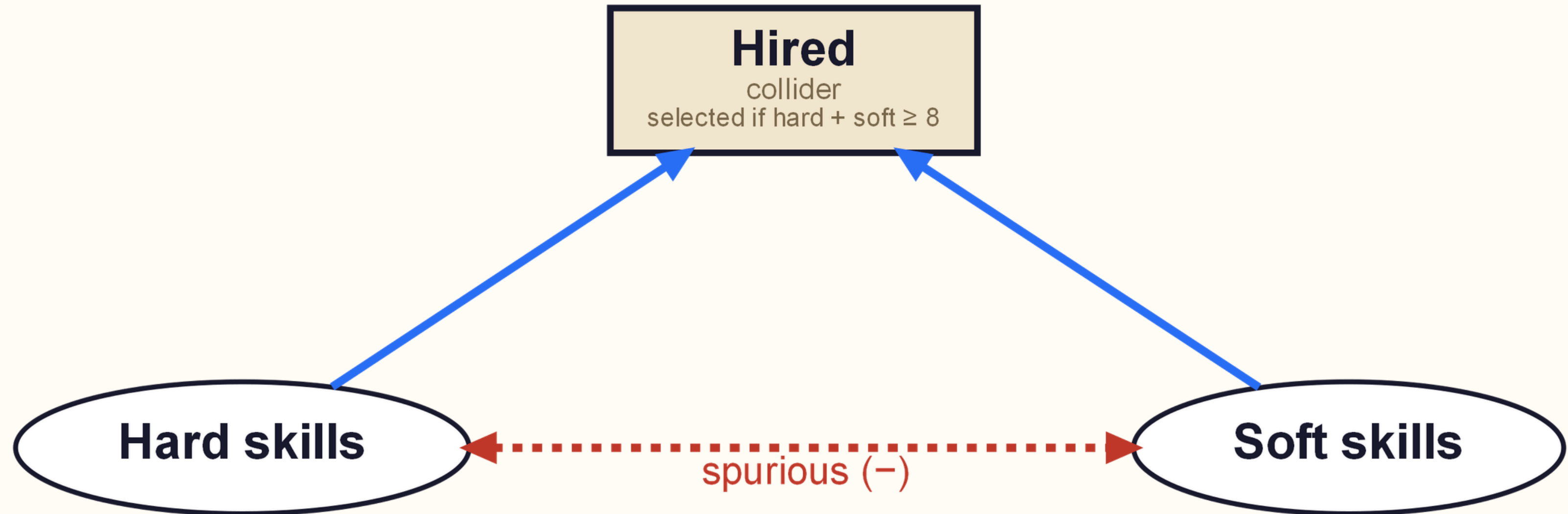
And yet this **correlation** has **no causal basis** in the general population, it has been solely produced by the **selection mechanism**



Generalization vs Collider Bias

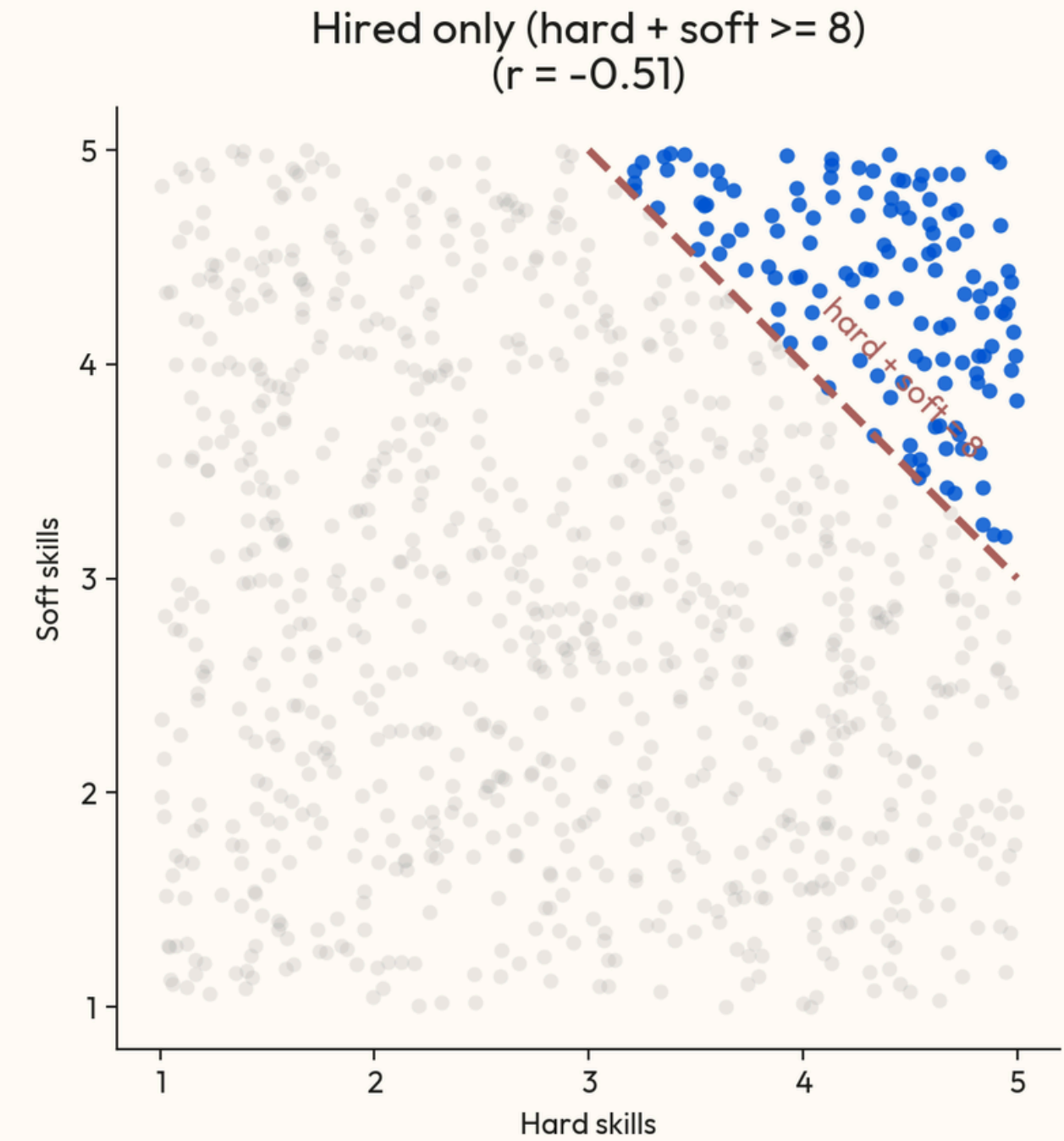
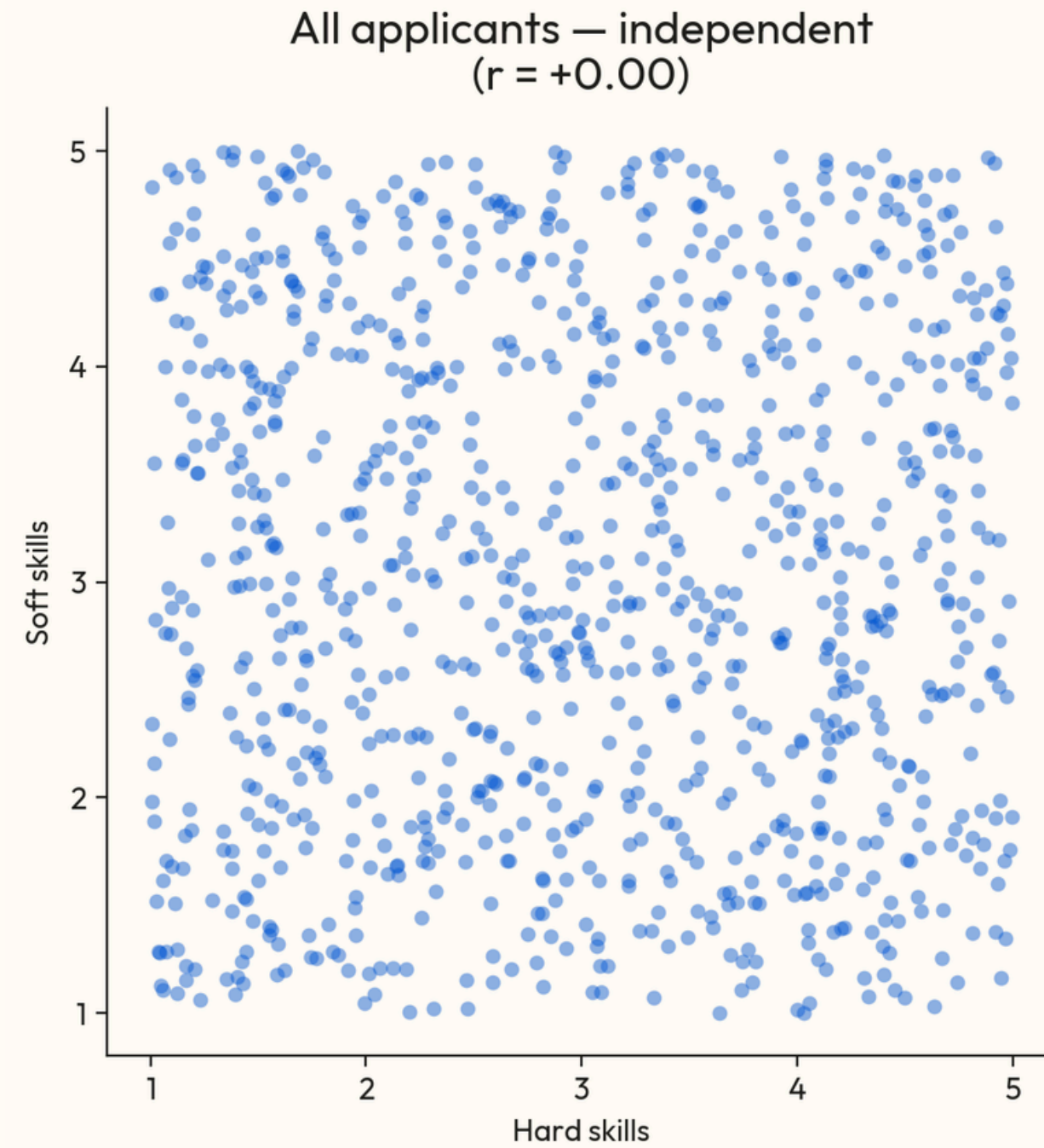


Generalization vs Collider Bias

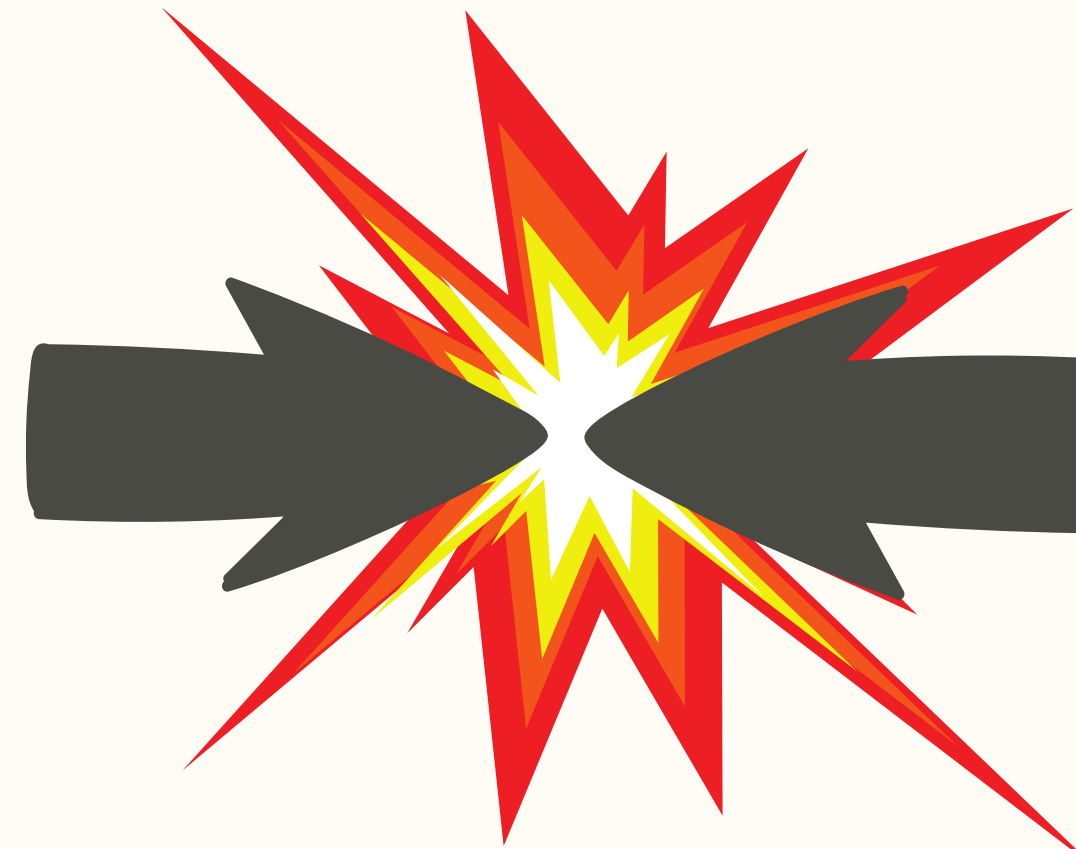


Generalization vs Collider Bias

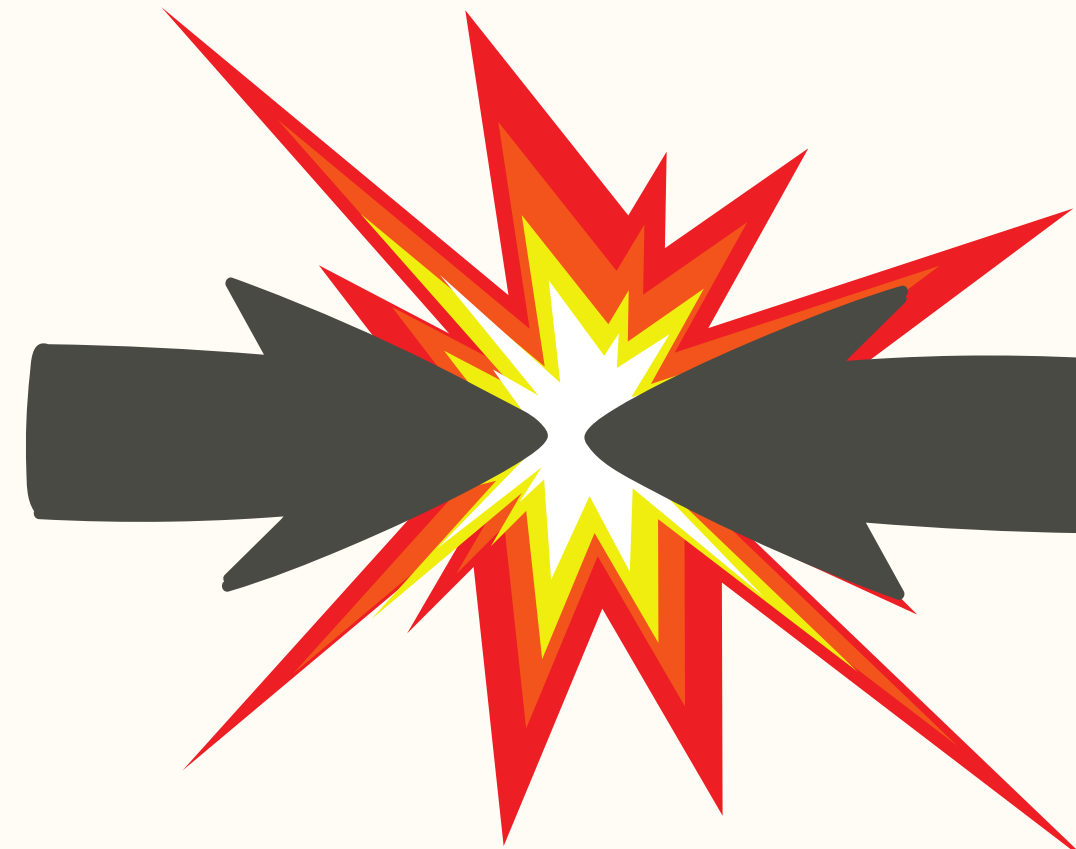
Selecting on a collider manufactures correlation



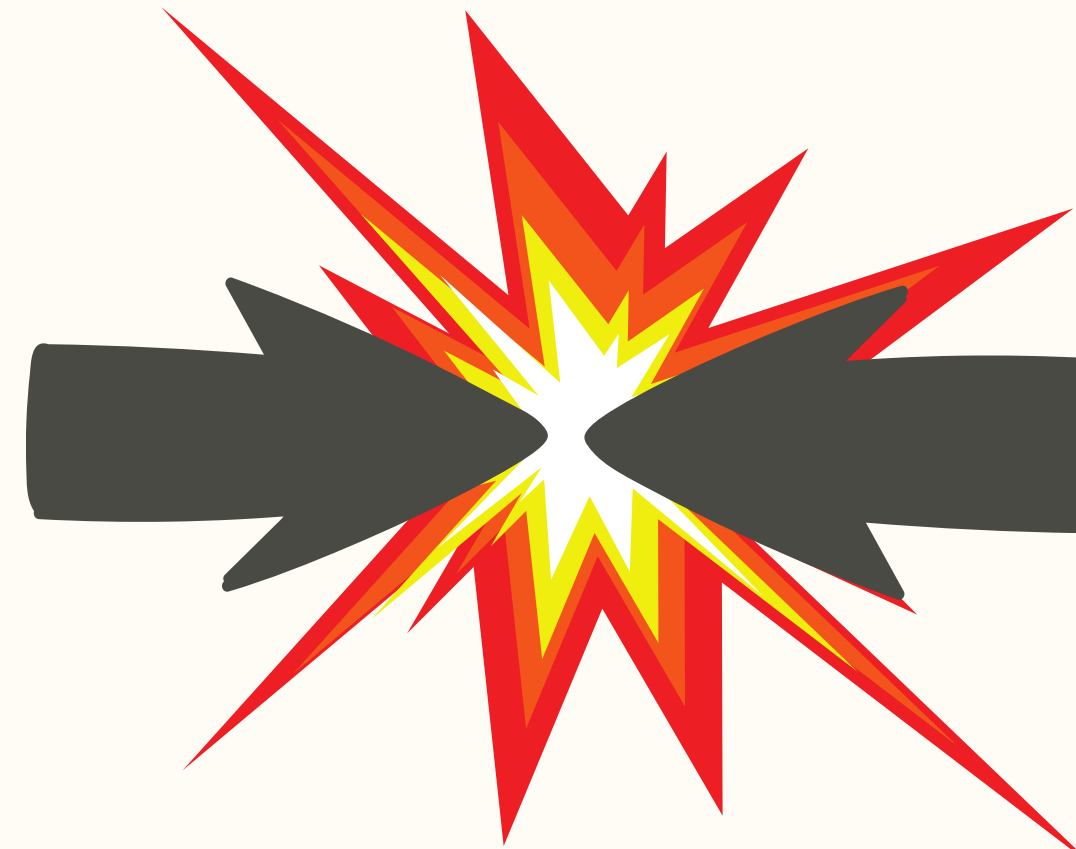
This effect is called the **collider bias** (a type of **selection bias**) and it can affect any type of study, including randomized, purely observational, and mixed



The **spurious correlation** produced by selection bias makes it difficult or impossible to generalize the study results to the **overall population**



Moreover, it may – and under true independence will – produce an **appearance of a causal effect** in-sample



Generalization vs Collider Bias

Can we use **big data** and **machine learning** to address these issues?

BIG DATA



Is Double Machine Learning Magic?

Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement*

Brett R. Gordon
Kellogg School of Management
Northwestern University

Robert Moakler
Ads Research
Meta

Florian Zettelmeyer
Kellogg School of Management
Northwestern University and NBER

September 21, 2022

Is Double Machine Learning Magic?

Is Double Machine Learning Magic?

- The authors looked at 663 large-scale RCT experiments at Facebook

Is Double Machine Learning Magic?

- The authors looked at 663 large-scale RCT experiments at Facebook
- Compared them to observational analyses (one using DML)

Is Double Machine Learning Magic?

- The authors looked at 663 large-scale RCT experiments at Facebook
- Compared them to observational analyses (one using DML)
- Used over 5,000 features in the observational analyses (a “very rich set of features”)

Is Double Machine Learning Magic?

- The authors looked at 663 large-scale RCT experiments at Facebook
- Compared them to observational analyses (one using DML)
- Used over 5,000 features in the observational analyses (a “very rich set of features”)
- Concluded that “despite having access to large-scale experiments and rich user-level data, (...) are unable to reliably estimate an ad campaign’s causal effect” using DML

Are their conclusions **useful**?



Is Double Machine Learning Magic?

- Although the authors mentioned the number of features in the study, they did **not mention** how these features were **selected**

Is Double Machine Learning Magic?

- Although the authors mentioned the number of features in the study, they did **not mention** how these features were **selected**
- From the causal point of view this **selection** step is **critical** – it can make or **break causal identification**

Is Double Machine Learning Magic?

- Although the authors mentioned the number of features in the study, they did **not mention** how these features were **selected**
- From the causal point of view this **selection** step is **critical** – it can make or **break causal identification**
- For instance, including a **collider** will almost certainly introduce bias due to **violated identification**

Is Double Machine Learning Magic?

- Although the authors mentioned the number of features in the study, they did **not mention** how these features were **selected**
- From the causal point of view this **selection** step is **critical** – it can make or **break causal identification**
- For instance, including a **collider** will almost certainly introduce bias due to **violated identification**
- The same will be true if we include **mediators** (or their descendants) or **not include confounders**

Is Double Machine Learning Magic?

- **Confounders** are often mentioned as the most challenging source of bias, because just one unobserved confounder can significantly bias the results

Is Double Machine Learning Magic?

- **Confounders** are often mentioned as the most challenging source of bias, because just one unobserved confounder can significantly bias the results
- Is there anything we can do about variables that for one reason or another we cannot observe / record?

Is Double Machine Learning Magic?

- **Confounders** are often mentioned as the most challenging source of bias, because just one unobserved confounder can significantly bias the results
- Is there anything we can do about variables that for one reason or another we cannot observe / record?
- Fortunately, **yes!**

Is Double Machine Learning Magic?

- **But before...**

Is Double Machine Learning Magic?

- **But before...**
- Let's clarify what **Double Machine Learning** (DML) is and what it isn't

Is Double Machine Learning Magic?

- **But before...**
- Let's clarify what **Double Machine Learning** (DML) is and what it isn't
- DML is also known as **de-biasing machine learning**

Is Double Machine Learning Magic?

- **But before...**
- Let's clarify what **Double Machine Learning** (DML) is and what it isn't
- DML is also known as **de-biasing machine learning**
- But de-biasing is focused on **particular types of bias**

Is Double Machine Learning Magic?

- **But before...**
- Let's clarify what **Double Machine Learning** (DML) is and what it isn't
- DML is also known as **de-biasing machine learning**
- But de-biasing is focused on **particular types of bias**
- They cannot reverse **neither confounding nor collider bias**

Is Double Machine Learning Magic?

- **But before...**
- Let's clarify what **Double Machine Learning** (DML) is and what it isn't
- DML is also known as **de-biasing machine learning**
- But de-biasing is focused on **particular types of bias**
- They cannot reverse **neither confounding nor collider bias**
- DML addresses **estimation-related biases** not identification-based biases

Is Double Machine Learning Magic?

- **But before...**
- The Facebook paper authors **did not distinguish between these two sources of bias**

Is Double Machine Learning Magic?

- **But before...**
- The Facebook paper authors **did not distinguish between these two sources of bias**
- It seems they **implicitly expected** that DML could in principle address identification-induced biases (otherwise why the study)?

Is Double Machine Learning Magic?

- **But before...**
- The Facebook paper authors **did not distinguish between these two sources of bias**
- It seems they **implicitly expected** that DML could in principle address identification-induced biases (otherwise why the study)?
- DML is a great method, but **it's not magic** – it can only work when the provided data allows for **causal identification** – it's either randomized or selected in a way that correctly reflects the structural relationships between the variables

Is Double Machine Learning Magic?

- Now, back to **hidden confounding**...



Three Ways to Work with Hidden Confounding

Three Ways to Work with Hidden Confounding

- **Hidden confounding** is indeed challenging and often **practically unavoidable** in observational studies



Three Ways to Work with Hidden Confounding

- **Hidden confounding** is indeed challenging and often **practically unavoidable** in observational studies
- So... what can we do about it?



Three Ways to Work with Hidden Confounding

- **Sensitivity analysis**



Three Ways to Work with Hidden Confounding

- Sensitivity analysis
- **Partial identification**



Three Ways to Work with Hidden Confounding

- Sensitivity analysis
- Partial identification
- **Hidden confounding detection**



Three Ways to Work with Hidden Confounding

- Sensitivity analysis
- Partial identification
- **Hidden confounding detection**



Three Ways to Work with Hidden Confounding

Ananke: A module for causal inference

Ananke, named for the Greek primordial goddess of necessity and causality, is a Python package for causal inference using the language of graphical models.

Ananke provides a Python implementation of causal graphical models with and without unmeasured confounding, with a particular focus on causal identification, semiparametric estimation, and parametric likelihood methods.



Three Ways to Work with Hidden Confounding

Ananke: A module for causal inference

Ananke, named for the Greek primordial goddess of necessity and causality, is a Python package causal inference using the language of graphical models.

Ananke provides a Python implementation of causal graphical models with and without unmeasured confounding, with a particular focus on causal identification, semiparametric estimation, and parametric likelihood methods.



Three Ways to Work with Hidden Confounding

Ananke: A module for causal inference

Ananke, named for the Greek primordial goddess of necessity and causality, is a Python package causal inference using the language of graphical models.

Ananke provides a Python implementation of

U
es
causal-falsify

pypi v0.3.1 docs passing downloads 5k license MIT

causal-falsify: A Python library with algorithms for falsifying unconfoundedness assumption in a composite dataset from multiple sources.

This library implements algorithms proposed in our two papers based on testing independence of causal mechanisms:

- Detecting Hidden Confounding in Observational Data Using Multiple Environments – NeurIPS 2023 ([pdf](#))
- Falsification of Unconfoundedness by Testing Independence of Causal Mechanisms – ICML 2025 ([pdf](#))



What can an RCT teach us?

The NEW ENGLAND
JOURNAL of MEDICINE

CURRENT ISSUE ▾ SPECIALTIES ▾ TOPICS ▾

ORIGINAL ARTICLE

Comparison of Upper Gastrointestinal Toxicity of
Rofecoxib and Naproxen in Patients with
Rheumatoid Arthritis

Authors: Claire Bombardier, M.D., Loren Laine, M.D., Alise Reicin, M.D., Deborah Shapiro, Dr.P.H., Ruben Burgos-Vargas, M.D., Barry Davis, M.D., Ph.D., Richard Day, M.D., [+5](#), for the VIGOR Study Group [Author Info & Affiliations](#)

Published November 23, 2000 | N Engl J Med 2000;343:1520-1528 | DOI: 10.1056/NEJM200011233432103

Vol. 343 NO. 21 | Copyright © 2000

REPORTS

ROFECOXIB

Naproxen
Tablets
For oral use
500mg
28 tablets
WOCKHARDT



Q&A



- Book
- Newsletter
- Podcast
- Courses
- and more!