

What does statistical testing do?

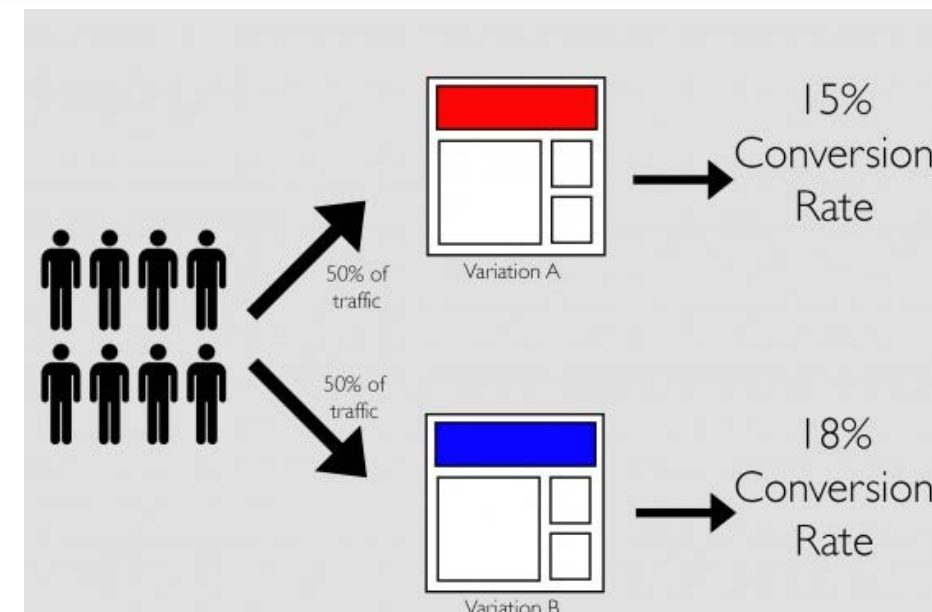
- Randomized Controlled Trial - the **gold standard** of causal inference

Pharmaceutical
Evaluation



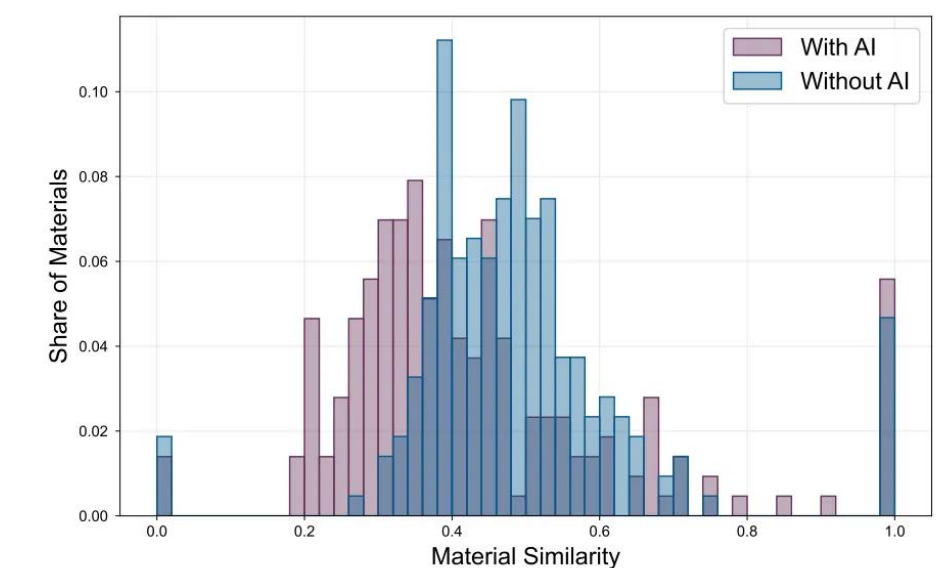
FDA Approval

A/B Testing



Code Approval

Quantitative
Social Science



Paper Approval

Causation=Approval?

RCTs are a building block for rulemaking

Neyman's Potential Outcomes

Splawa-Neyman, Jerzy. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." 1923

	A	B	C	D	E	F	G	H
Treatment	N	N	Y	Y	N	Y	N	N
Control	Y	Y	Y	Y	N	Y	Y	Y

Neyman's Potential Outcomes

Splawa-Neyman, Jerzy. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." 1923

	A	B	C	D	E	F	G	H
Treatment	N	N	?	Y	?	?	?	N
Control	?	?	Y	?	N	Y	Y	?

Neyman's Potential Outcomes

Splawa-Neyman, Jerzy. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." 1923

	A	B	C	D	E	F	G	H
Treatment	N	N	?	Y	?	?	?	N
Control	?	?	Y	?	N	Y	Y	?

- If we can randomize, we can estimate the mean outcome in treatment and control.
- Probability theory can estimate the variance even if the data are non-Gaussian.
- The RCT is a measurement device with quantifiable uncertainty.
- *RCTs measure the size of an effect, they don't determine causation.*
- RCTs measure the effect of causes.

Neyman's Potential Outcomes

Splawa-Neyman, Jerzy. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." 1923

	A	B	C	D	E	F	G	H
Treatment	N	N	?	Y	?	?	?	N
Control	?	?	Y	?	N	Y	Y	?

randomized treatment assignments Z_i in $\{0, 1\}$, outcomes Y_i

average treatment effect:
$$ATE = \frac{1}{n} \sum_{i=1}^n Y_i(1) - \frac{1}{n} \sum_{i=1}^n Y_i(0)$$

"Estimated" average treatment effect:
$$\widehat{ATE} = \frac{2}{n} \sum_{\{i: Z_i=1\}} Y_i(1) - \frac{2}{n} \sum_{\{j: Z_j=0\}} Y_j(0)$$

Estimated Effect Size: E

Standard Error: σ

Why RCTs?

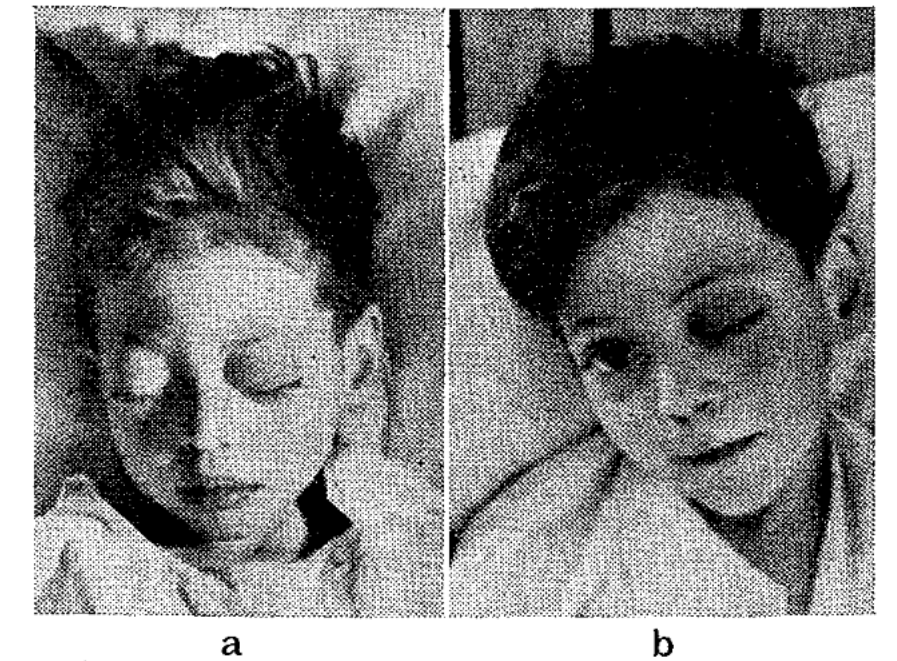
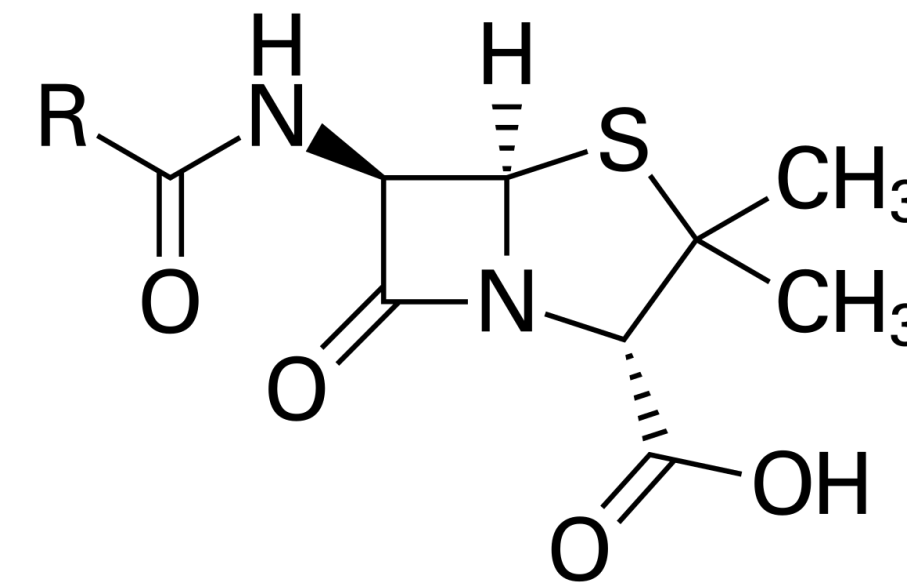
Don't start with Fisher, start with Bradford Hill.

Effect Size: E

Standard Error: σ

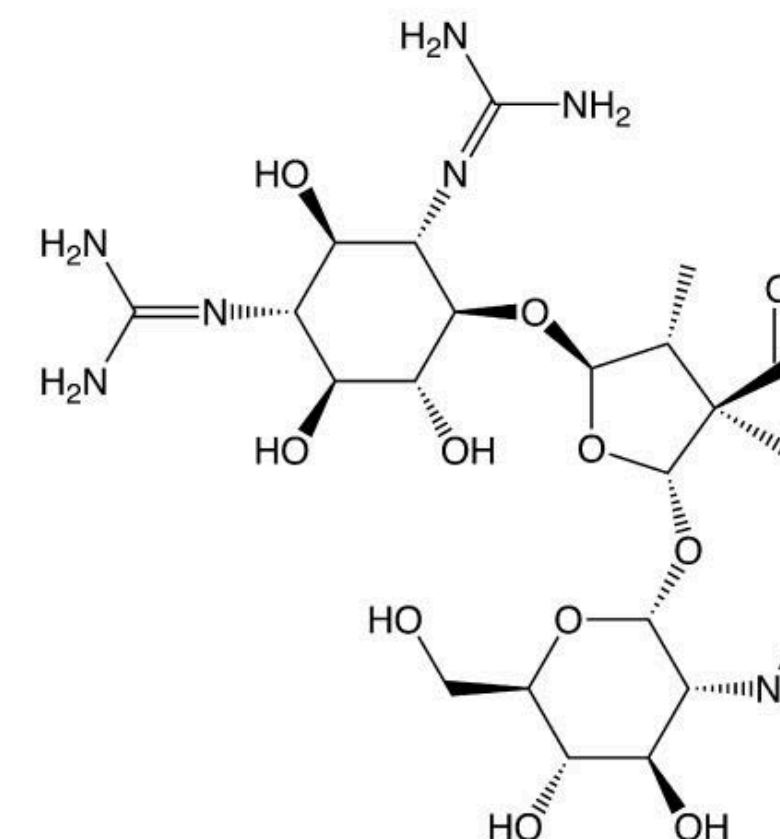
Penicillin

- Right before RCTs became a thing (1941)
- The evidence was all case studies. ($E/\sigma = \infty$)



Streptomycin

- 1947. First RCT. Designed by Hill.
- 4 deaths out of 55 in treatment, 15 deaths out of 52 patients given bed rest alone.
- ATE = 22%.
- $E/\sigma = 3$



Medicine is the only human-facing science with 5σ interventions

Penicilin

- Right before RCTs became a thing (1941)
- The evidence was all case studies. ($E/\sigma = \infty$)

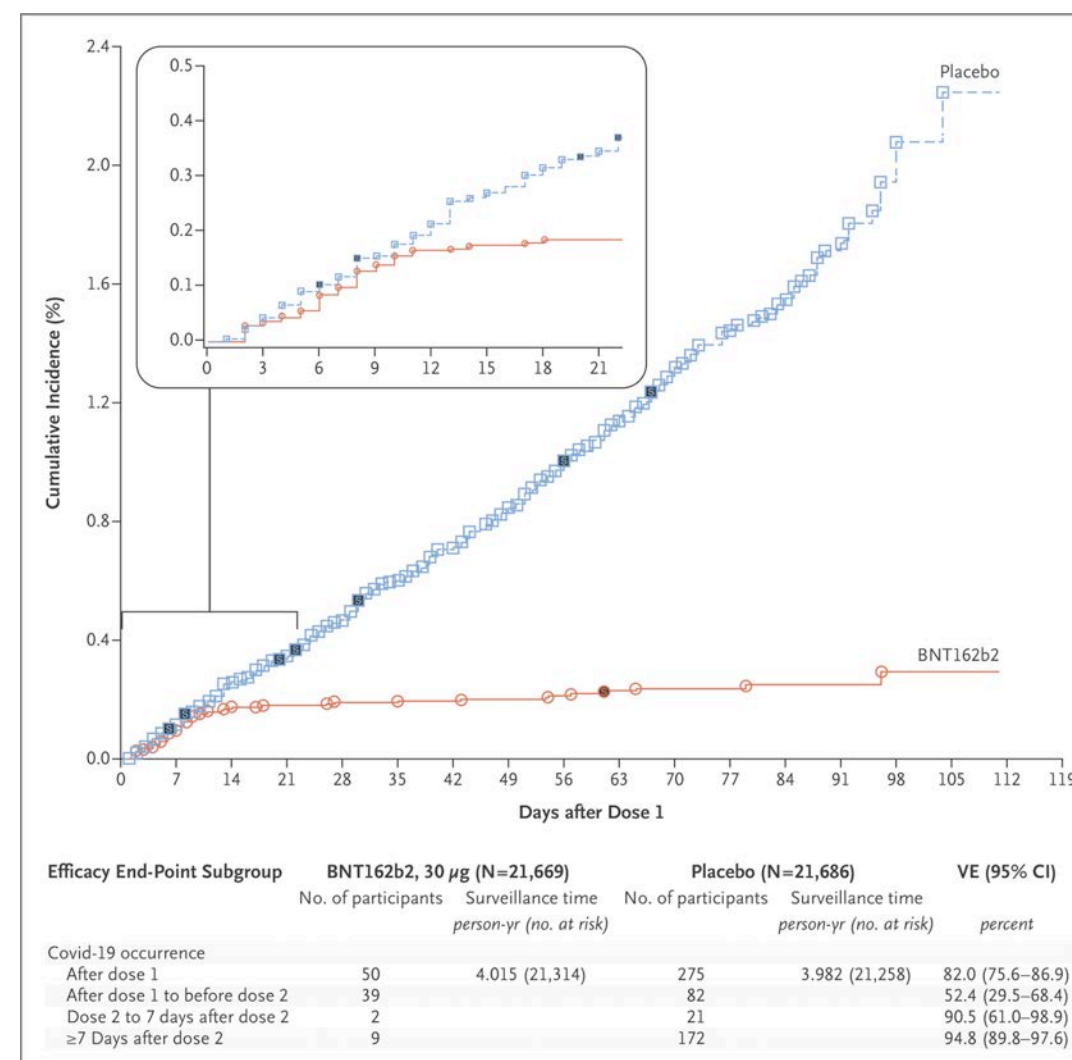
Effect Size: E

Standard Error: σ

Salk Vaccine (1954)

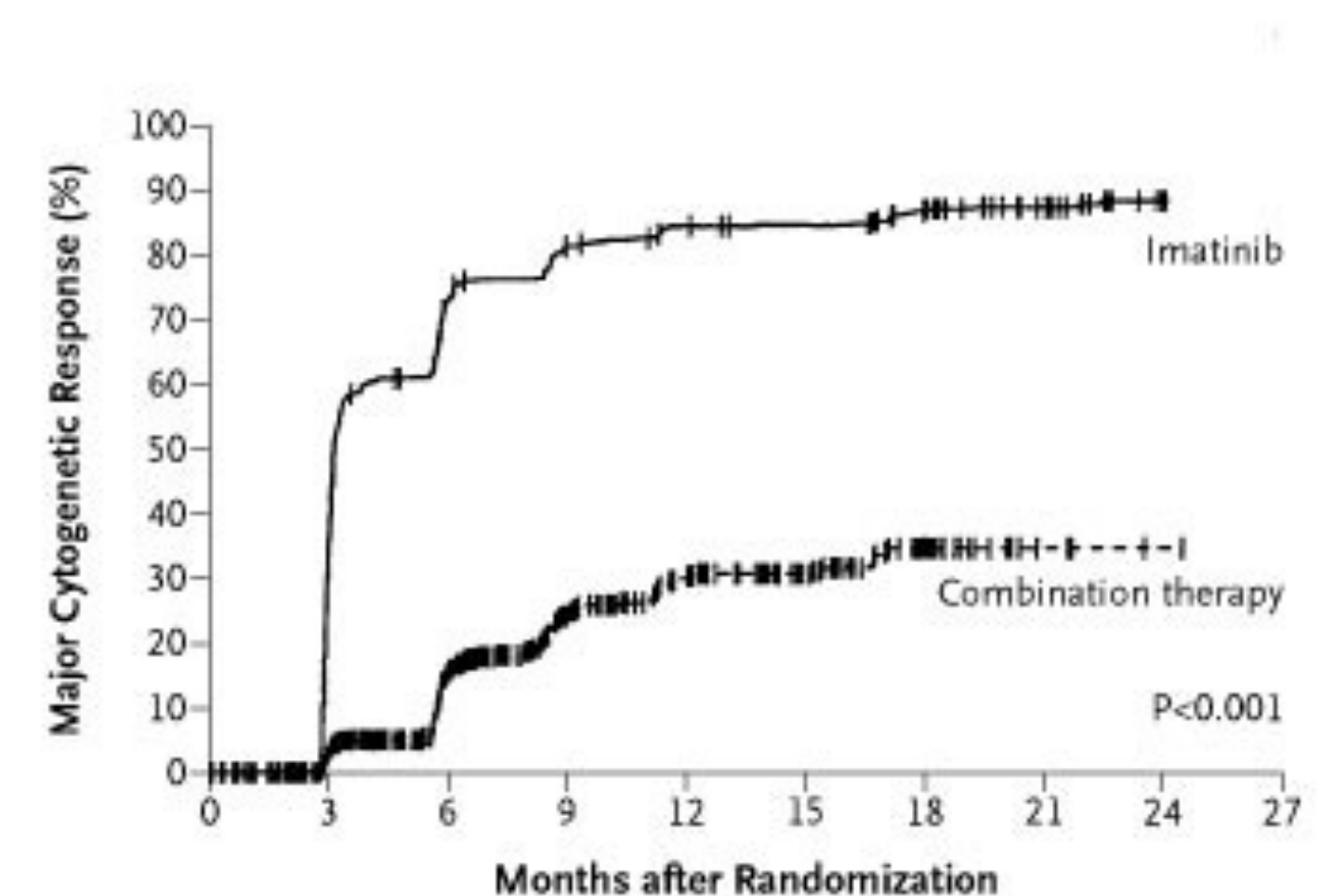
- Only quasi-randomized
- But in the randomized analysis, E/σ was 7 for paralysis.

mRNA Vaccines



$E/\sigma > 20$

gleevec/imatinib



$E/\sigma > 17$

Medicine is the only human-facing science with 5σ interventions

- *Deficiency Diseases*: vitamins
- *Infections*: penicillin, sovaldi, HIV antiretroviral therapy
- *Vaccines*: Salk vaccine, MMR vaccine, covid vaccines
- *Pain, fever*: ibuprofen, paracetamol, opiates
- *Diabetes*: insulin therapy
- *Chemotherapy*: pediatric acute lymphoblastic leukemia, many others since
- *Cancer*: Gleevec
- *Obesity*: GPL-I agonists
- *Surgery*: aortic valve replacement, organ transplants

we seem to find many “one in a million” type treatments

FDA Causality

- Sulfanilamide disaster
- Federal Food, Drug, and Cosmetic Act of 1938
- FDA given mandate to prove drugs *safe* and *effective*
- These are defined *statistically*. Evaluated statistically.
- 1962 Kefauver-Harris Amendment to the FFDCFA demands *substantial evidence* of efficacy for the approval of a new drug.
- Two randomized controlled clinical trials with results significant at the 0.05 level.
- *What does this mean in practice?*



Correlations and Stories

- Let Y_i denote the measured outcome of each individual and Z_i their treatment assignment. Let $R(\mathbf{Z}, \mathbf{Y})$ denote the Pearson r-coefficient between the n-vectors \mathbf{Z} and \mathbf{Y} .

$$R(\mathbf{Z}, \mathbf{Y}) = \frac{\text{cov}(\mathbf{Z}, \mathbf{Y})}{\text{std}(\mathbf{Z})\text{std}(\mathbf{Y})}$$

- Policy:** accept the treatment if, for a specified threshold, t :

$$R(\mathbf{Z}, \mathbf{Y}) \geq \frac{t}{\sqrt{n}}$$

- If \mathbf{Y} binary, $t=1.96$: equivalent to proportions z-test at the level $\alpha=0.05$ level. Also equivalent to chi-squared test.

$$\frac{E}{\sigma} = \frac{n^{1/2}\text{cov}(\mathbf{Z}, \mathbf{Y})}{\text{std}(\mathbf{Z})\text{std}(\mathbf{Y})}$$

See, e.g., Cohen. *Statistical Power Analysis for the Behavioral Sciences* (1969)

Justifying Stories About Correlations

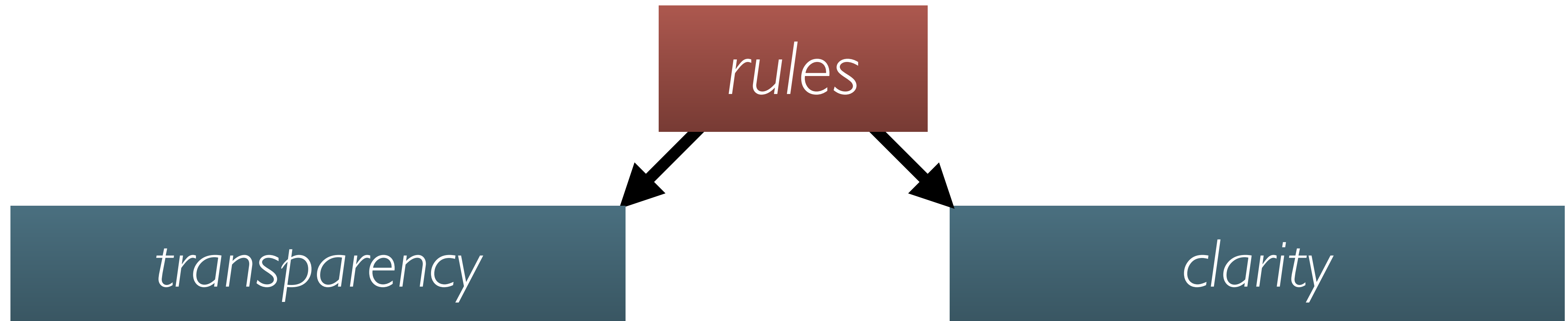
Treatment **Z**. Outcome **Y**. **Policy**: Approval if $R(\mathbf{Z}, \mathbf{Y}) \geq \frac{t}{\sqrt{n}}$

- If **Z** and **Y** are normal, $R(\mathbf{Z}, \mathbf{Y}) \approx n^{-1/2}$
- This baseline comes from statistical reasoning.
- Requires large enough sample to demonstrate higher correlation than random coin flips.
- Interocular Trauma Test is still preferable to a z-test. But we need to set a floor.

Drugs are approved if correlation between treatment and benefit is large in an RCT

Ex Ante Policy

- *Ex ante* because the rules and procedures are designed before data collection.
- *Policy* because the rules aim to inform actions to be taken after data collection.



- “Statistical games”
- *remove biases, fair assignment, clear thresholds, ...*

- Forces policy-makers to specify models, interventions, values
- *must demonstrate safety, equipoise, informed consent, ...*