

# RCTs and the Gold Standard: Three Theses

**1. RCTs are not the only game in town:** Many methods live up to a gold standard.

**ALL** gold standard methods need strong assumptions to guarantee their reliability.

Which method to use depends on which assumptions you can be/become sure of.

So, gold is where you know what you need to know to apply a method reliably

# Three theses

2. Since all gold standard methods need strong assumptions to ensure their conclusion, the **conclusions** of these methods are concomitantly **very limited in scope** – they do not cover much.

## **Three theses**

### 3. Objectivity – Who is the enemy?

Nowadays: personal judgment.

**Problem:** the advantage of gold standard methods is that there are prescriptions for their reliable application. Yet there are no formal prescriptions for determining the scope of gold-standard conclusions outside the narrow range where they can be proven reliable. This holds equally for RCTs as for other gold standard methods.

**Expert judgment is essential.**

# The talk

I. Clinchers v vouchers: A distinction and its implications

II. Some examples of clinchers

III. RCTs: the vanity of rigor in biomedical trials

A. Ideal RCTs

B. Real RCTs

C. Scope

IV. Objectivity: what are we afraid of?

Bias v reliability

# I. Clinchers v vouchers

Two kinds of methods for warranting causal claims

- those that *clinch* the conclusion but are *narrow* in their range of application
  - derivation from theory
  - RCTs
  - certain econometric methods
  - ...
- those that merely *vouch for* the conclusion but are *broad* in their range of application
  - QCA (qualitative comparative analysis)
  - look for quantity and variety of evidence
  - ethnographic methods
  - show conclusion as plausible explanation of evidence
  - ...

## Costs and benefits

**Benefit:** Clinchers are deductive: *if* they are correctly applied *and* their assumptions are met, then *if* our evidence claims are true, so too will be our conclusions -- a huge benefit. *But*

**Cost:** There is an equally huge cost. These methods are concomitantly narrow in scope. The assumptions necessary for their successful application

- tend to be extremely restrictive
- they can only take a very specialized type of evidence as input
- they have only special forms of conclusion as output

It takes strong premises to deduce interesting conclusions. Strong premises tend not to be widely true.

**TRADE-OFF:** We can ask for methods that clinch their conclusions but the conclusions are likely to be very limited in their range of application

## II. Examples of methods that clinch conclusions

1. Econometric methods
2. Galilean experiments
3. Probabilistic/Granger causality
4. Derivation from established theory
5. Tracing the causal process/mechanism
6. Ideal RCTs
7. ...

These are clinchers: we can *prove* that if the auxiliary assumptions are true, the methods are applied correctly and the outcomes have the right form, the hypothesis must be true.

The point is: When it comes to clinchers – to methods from which we can rigorously derive the hypothesis from the evidence – RCTs are not the only game in town. There are lots of methods that can clinch conclusions. **ALL** of them require strong assumptions in order to assure their applicability. This restricts the range of applicability of every one of them. Which method is the best to use in a given situation depends on what we know and what we can do in that situation. And there is a vast variety of different kinds of things we know or can come to know in different situations.

**Caution:** to buy the benefits of a clinching method we must be able to ensure that it is highly probable that *all* the requisite premises obtain.

That's because of the *weakest link* principle for deductive reasoning. The probability of the conclusion can be no higher than that of the weakest premise.

- Suppose you have 10 premises, 9 of them almost certain, one dicey. Your conclusion is highly insecure, not 90% probable.
- In a deductive argument  $P(\text{conclusion}) \leq P(\text{conjunction of premises})$

I belabour this because of the benefits of clinching methods – clinchers are rigorous. It is transparent *why* the results are evidence: given the background assumptions the hypothesis follows deductively from the results. And it is transparent *when* the results are evidence: when the background assumptions are met.

- Contrast ethnographic methods and expert judgment. These can provide extremely reliable evidence. But there is no specific non-trivial list of assumptions that tell when they have done so.

**BUT** if you want credit for this benefit of a clinching method, you must be able to show that the *conjunction* of your premises has high probability *in the case at hand*.

### III.A. Ideal RCTs: briefly... to test ‘T causes O’

1. *Fixed probability and fixed causal structure.* For a given outcome O, we study a population all of whose members are governed by the same causal structure CS for O and described by the same probability P.

P is defined over an event space  $\{O, T, K_1, K_2, \dots, K_n\}$ , where each  $K_i$  is a state description over ‘all other causes’ of O except T.\*

■ ‘State description’ = (loosely)  $K_i$  *holds fixed* all causes of O other than T.

■ ‘Causal structure’ = (loosely) a description of all the causal pathways by which O can be produced.

\* This must include ‘spontaneous generation’. More formally,  $K_i$  holds fixed one variable on each pathway that does not go through T, as judged by the causal structure CS.

## Ideal RCTs: briefly...

2. *Probabilistic theory of causality (Granger causality)*. T causes O for a population with causal structure CS and probability P if  $P(O/T \& K_i) > P(O/\neg T \& K_i)$  for some  $K_i$ .

■ So, in a population where ‘all other’ causes of O are held fixed, any difference in probability of O with T present v T absent shows that T causes O in that population

■ We need this to make a connection between causes and probabilities

■ Rationale: Mill’s method of difference plus the assumption that if there are more O’s in one subpopulation than another described by the same probability there must be a causal explanation.

3. *Idealization*. In an *ideal RCT* the  $K_i$  are distributed identically between the treatment and control groups

## **Ideal RCTs are clinchers:**

If  $P(O)$  in treatment group  $>$   $P(O)$  in the control group in an ideal RCT, then trivially by probability theory  $P(O/T \& K_i) > P(O/\neg T \& K_i)$  for some  $K_i$ .

Therefore: if  $P(O)$  in treatment group  $>$   $P(O)$  in control group, T causes O relative to CS,P.

Dawid, A. P. (2000) 'Causal inference without counterfactuals', *JASA* 95, 407-448.

Holland, P.W. (1986) 'Statistics and Causal Inference'. *JASA* 81, 945-960

N. Cartwright (1989) *Nature's Capacities and their Measurement*, OUP

## Remarks

From the point of view of the probabilistic theory of causality, T causes O (relative to CS,P) is consistent with T causes  $\neg O$  (relative to CS,P). This lines up with what we know of RCTs:

- RCTs deliver population-average results. A *positive* result shows that T causes O in at least one subpopulation. It could produce exactly opposite results in other subpopulations.
- Positive results are conclusive but *negative* are not: Equal probability for O in the treatment and control groups does not show that T does not cause O. It shows that *if* T causes O (because it does so in one  $K_i$ ) it must also cause  $\neg O$  (because it does so in some other  $K_i$ 's).

### **III.B. The scope of a real RCT.** What do we do to ensure the premises?

- Careful use of statistics to move from frequencies to probabilities
- ‘Random’ assignment to treatment and control groups
- Quadruple blinding
- Careful attention to drop-outs, non-compliance,...
- 

These last are to ensure premise 3 (equal probability of  $K_i$ 's in treatment and control groups). Let's accept premise 2 (link between causes and probabilities). What about premise 1?

## Premise 1

We experiment on a population of individuals who are taken to have the same *fixed causal structure* (albeit unknown) and *fixed probability measure* (albeit unknown).

**Our deductive conclusions are entirely confined to individuals with that very causal structure and probability.**

How do we know what individuals beyond those in our experiment this applies to?

### III.C. The Vanity of Rigor in RCTs

The title is borrowed from my ‘Vanity of Rigor in Economic Models’.  
Identical problems: internal v external validity.

Economists make a huge investment to achieve rigor *inside* their models. (INTERNAL VALIDITY)

**BUT**...how do they decide what lessons to draw about target situations *outside* their models? (EXTERNAL VALIDITY)

Essentially – expert judgment and scientific debate

So. We pay a lot for rigor in economic models, but the rigor is all in the front end. After that it's judgment and informal argument. Is the front-end cost worth paying given that rigor gives out long before the real conclusions are drawn?

## **And RCTs...**

Which 'target' populations have individuals with the same causal structures and probabilities as those in the experiment?

As with economic models, there can be intelligent discussion and reasoned decisions but rigor gives out.

Essentially we have good judgment and informal argument.

So. We pay a lot for rigor, but it is all in the front end. After that it's judgment and informal argument. Is the front-end cost worth paying given that rigor gives out long before the conclusions we need are drawn?

**WE NEED A CASE-BY-CASE DECISION**

## Two remarks on internal validity in RCTs

1. Random assignment is no guarantee of equal distribution of  $K_i$ 's. So what should we do to maximize the probability of premise 2?

- Manually ensure an equal distribution over all known relevant factors
- Get someone who has the background knowledge to recognize if something has gone wrong to have a close look. Not a methodologist but someone with subject-specific knowledge. ('Base-line' corrections.)

*This again involves judgement and informal argument*

2. Once we ‘manually’ ensure that all known relevant factors are distributed equally between treatment and control groups, what more does ‘random’ assignment buy?

○ John Worrall: It provides a safeguard against individuals – self-serving or other serving, consciously or unconsciously – selecting in a way that biases results.

- This is a good thing

- But when random assignment is costly we need to weigh the chances of this happening, and in a significant enough way, to jeopardize the reliability of the outcome (e.g. drug company trials where small differences matter to significance).

○ RCT advocates: Randomization is mandatory.

○ Me: That seems daft. It is always reasonable to do cost-benefit analysis and weigh the probability of various kinds of mistakes (*pace* calculation costs or moral reservations).

## IV. Who is the enemy?

Common theme: it is essential to guard against personal judgment.

Why?

- Judgements by trained experts can dramatically increase reliability
  - E.g. watching for problems in RCTs, base-line adjustments
- They are **ALWAYS** ineliminable

# Objectivity

Peter Galison and Lorraine Daston, *The History of Objectivity*

- Objectivity = the opposite of subjectivity
  - = the opposite of whatever we are worried about
- What we worry about has changed from epoch to epoch.
  - Early 19th c.: Era of ‘truth to nature’. True forms are hidden. It takes a disciplined ‘genius’ with years of careful, detailed observation to see the true forms behind variable phenomena (e.g. Goethe)
  - Late 19th c: Era of ‘mechanical objectivity’. Variable nature must be caught as it is. Personal intervention hampers this. Photographs are the paradigm.
  - Early 20th c: Era of ‘trained judgment’. Patterns are not transparent. We must be trained to see them.

Me: now -- the era of 'publicly accountable procedures'

- Mistrust of experts
- We suppose that as soon as people are involved they will act to serve human interest -- our own or others
- We believe that this is far more dangerous a problem than other sources of error
- We want procedural objectivity to ensure against personal liability

We put more emphasis on following the procedures than on getting it right.

\*\*Is this a prescription for getting better outcomes -- or is it a moral injunction: it is *bad* to go wrong by trusting to judgement?\*

## In Sum

- RCTs are not the only ‘gold standard’ method
- All methods -- clinchers and vouchers -- require good judgment to draw relevant conclusions -- and a great deal of it
- Since judgment cannot be eliminated, we had best get on with managing it
- This however requires judgment!
- So -- why think errors due to judgment will be universally more prevalent than other kinds?
- And at least once judgment is admitted we can draw conclusions based on **total evidence**.