Safe Inference

Say what you can and not what you can't.

Chad Hazlett

UCLA

chazlett@ucla.edu

20 July 2026

We left off with four theses

- **(1)** A causal claim is about a comparison we can *never fully observe* — so no amount of data settles it by itself.
- **(2)** *Identification comes before estimation*: we need assumptions linking what we want to know to what we can see.
- **(3)** There are no "causal methods." Credibility comes from the *assumptions and the story behind them*, not the technique.
- **(4)** Experiments are not a separate *kind* of knowledge — every design leans on assumptions; the trial just makes one hold by design.

This talk is about how we can live up to theses 3 and 4.

Story: A near-miss

- There is a brain cancer that kills half of newly diagnosed patients in about a year and 95% within 5 years.
- A surprising number of patients with an experimental therapy become long survivors.
- But a large, randomized trial was compromised, so no matter what the result or logically implied effects, it won't be believed.

Do we throw this evidence out?
Or is there defensible, life-saving information hiding in it?

Two ways to be wrong

- **Errors of Commission** — assert an effect that isn't defensible.
  - E.g. the observational result you report and believe despite confounding concerns.
- **Errors of Omission** — fail to recognize an effect that *is* defensible.
  - Here: defensible claims that could save lives but won't be accepted by community or FDA because it's not an RCT.
  - The studies you never do and papers you never write because you are afraid the effect "isn't identified."

Social science practice allows both mistakes, though we are (unevenly) sensitive to the first. We are almost entirely blind to the second.

Procedural rather than principled credibility

One villian I'll point to is a tradition of "procedural credibility" assessment (violating thesis 3).

By this I mean the habit of making crediblity judgements about causal claims based on *what investigators did* rather than *what assumptions are defensible* and where that leaves our beliefs about the causal effect.

Having heuristics based on procedures, procedural guidelines, etc. are all reasonable...But credibility ultimately derives from what we can and cannot claim, not from what we did.

Safe inference

The concept of “safe inference” attempts to capture and avoid these dual risks.

In a sentence:
Make (only) the research conclusions you can defend (or say what assumptions would need to be defended to reach that conclusion).

This is *safe*, not *silent* because it works in both direction:

1. Do not assert the indefensible.
2. Do not refuse to assert the defensible.

The flip: from point identification to safe inference

The paradigm we grew up with — point identification:

- Assume whatever it takes to land a single number; present it as if exact.
- The only "uncertainty" reported is *statistical* — sampling noise, *not* whether the assumptions hold.
- Caveats go in the discussion section, and protect no one.

Safe inference turns it over:

- Start from what we can *defensibly* assume — and read off where that leaves the answer.
- Sometimes this gives a sharp result; sometimes an honest "if-then"; sometimes a clear "we can't believe this yet."
- Uses what we have without over-claiming, and makes *which assumptions to defend* the creative frontier.

This falls under "partial identification", here we put it to particular use in the safe inference framework.

The engine: the counterfactual $Y(0)$

It is easy to go down technical rabbit holes now, but a central unifying theme is simply to ask yourself:

Where and how can we defensibly **know** or **bound** the counterfactual outcome, $Y(0)$, among the treated?

This is key because we are (or could be) interested in an effect among the treated, and the only unknown to grapple with is how they would have done without the treatment.

That is,

We can take interest in

$$\text{ATT} \;=\; \mathbb{E}[\, Y_i(1) - Y_i(0) \mid D_i = 1 \,]$$

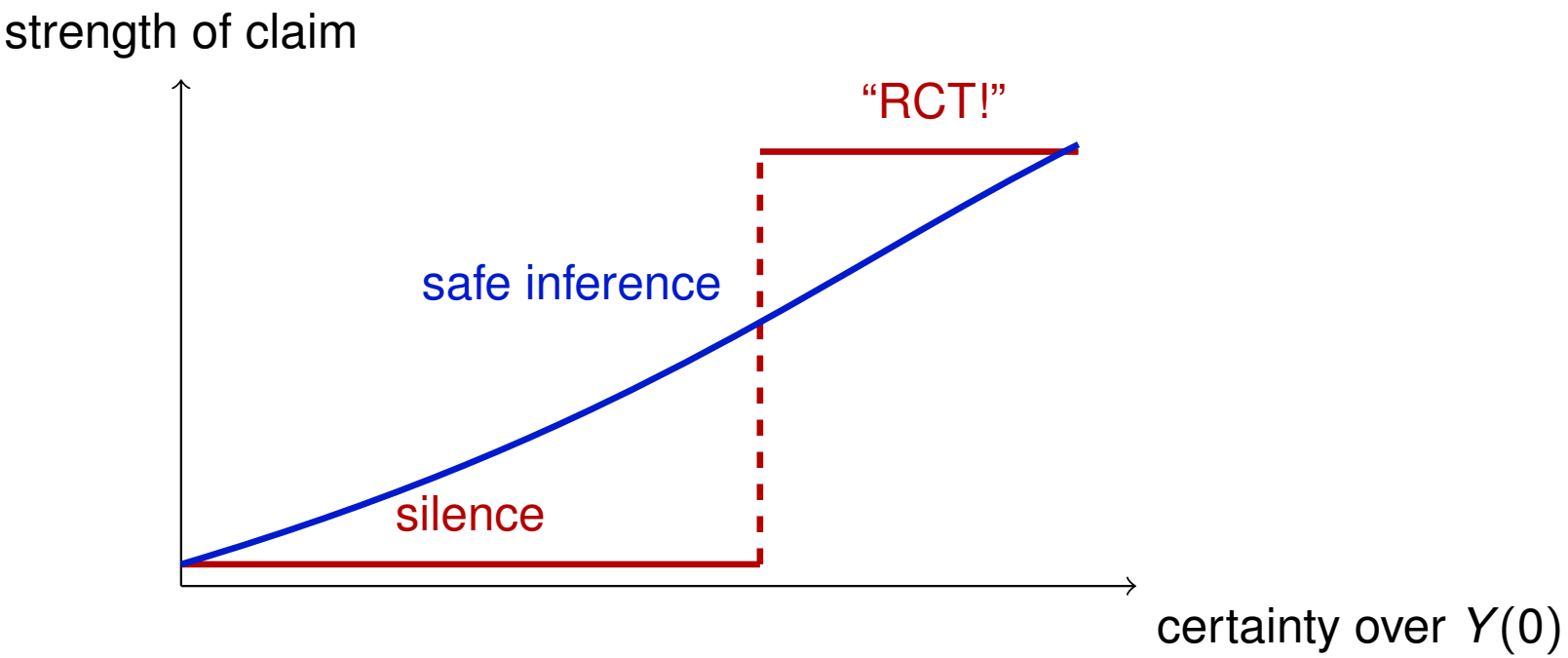
- For a treated unit: $Y_i(1)$ is *observed*; $Y_i(0)$ is the *missing* counterfactual.
- Estimator: $\widehat{\text{ATT}} = \dfrac{1}{n_1} \displaystyle\sum_{i:\, D_i=1} \left[\, Y_i - \widehat{Y_i(0)} \,\right] = \overline{Y} - \widehat{\overline{Y(0)}}.$

Every design is just a claim about $\widehat{Y_i(0)}$ or its mean, in particular:

- Think creatively about legible, arguable assumptions that lead to bounds on $\widehat{\overline{Y(0)}}$.
- Use that to report range of ATT that cannot be ruled out.
- See what you can and cannot conclude, or what some existing conclusion would imply and whether you believe it.

Smoothing the credibility curve

A useful picture here, at least regarding the comparison to RCT-focused inference:



Case 1 by example — melanoma

Metastatic melanoma (BRAF V600E). Given vemurafenib — a *single arm* (no control group): $\bar{Y}_{\textbf{treated}} \approx 48\%$ of patients' tumors substantially shrank.

What would happen untreated? These tumors essentially never regress on their own: spontaneous regression $\approx 0.23\%$; even old chemotherapy (dacarbazine) reached only $\approx 5\%$. So $\mathbb{E}[Y(0)]$ — the response rate *without* the drug — is near zero.

But these patients weren't a random draw. Could they have been unusually likely to remit even untreated? Not much.

Conservatively let $\widehat{Y(0)}$ run all the way to 10% (double the best chemotherapy). Then

$$\text{ATT} \;=\; \bar{Y}_{\text{treated}} - \widehat{Y(0)} \;=\; 48\% - \underbrace{[0\%,\ 10\%]}_{\text{conservative}} \;=\; [\mathbf{38\%},\ \mathbf{48\%}].$$

We can be quite certiain then of at least a $\sim$40-point effect, no RCT needed.

Case 1 in general: Using an almost-known $\mathbb{E}[Y(0)]$

Suppose **external control arm** (ECA) gives us $\mathbb{E}[Y(0)]_{\text{ECA}}$.

We don't assume this is our treated group's exact untreated outcome, rather we allow for a difference:

$$\delta \equiv \mathbb{E}[Y(0) \mid \text{treated}] - \mathbb{E}[Y(0) \mid \text{ECA}]$$

Then the effect on the treated is

$$\widehat{\text{ATT}}(\delta) \;=\; \bar{Y}_{\text{treated}} - \mathbb{E}[Y(0)]_{\text{ECA}} - \delta$$

- δ is *unobserved*; we consider a range that is logically plausible, arguing how far the *treated* group's own untreated outcome could sit from the observed ECA (the room left for *selection*).
- **Grim, near-deterministic prognosis** $\Rightarrow$ little room for $\delta \Rightarrow$ a single treated arm $+$ ECA is nearly an RCT.

The melanoma case is easy because the response is huge compared to the nearly certain prognosis absent treatment.

Example 2: It isn't always so.

Metastatic pancreatic cancer (RAS-mutated, previously treated). On daraxonrasib (as a *single arm*), median survival was $\overline{Y}_{\text{treated}} \approx$ **13** months.

What would happen absent treatment? *Not* as knowable.

- On standard chemo these patients live ≈ 6.7 months.
- But treated arm may have been healthier, so may have higher $Y(0)$.

What is a defensible δ. Suppose experts agree they "cannot rule out that carefully selected patients might have increased the expected non-treatment survival by 0 to 6 months"

- $\widehat{\overline{Y(0)}} \in [6.7,\ 12.7]$ mo.
- $\text{ATT} = \bar{Y}_{\text{treated}} - \widehat{\overline{Y(0)}} = 13 - [6.7,\ 12.7] \approx [\mathbf{0.3},\ \mathbf{6.3}]$ mo

Too fragile to be sure.

- A randomized trial would make sense.
- Though consider single-arm trial with patients chosen to limit δ!

Case 2: Selection and the stability-controlled quasi-experiment (SCQE)

You may have treated and control units, but not randomly assigned.

Suppose we have two cohorts or groups, one in which treatment is inaccesible (**low-use**, $Z = 0$) and one in which it has become available and people select non-randomly into it (**high-use**,$Z = 1$).

Example: COVID-19 patients admitted to a hospital in Feb–Mar 2020.

- Of patients admitted in February, 0% got dexamethasone
- Of patients admitted in March, 30% (very non-randomly) received it

Key: Absent this treatment how much could the mortality rate change from one month to the next? I.e. the "baseline trend".

$$\delta \equiv \mathbb{E}[Y(0)|Z = 1] - \mathbb{E}[Y(0)|Z = 0]$$

This baseline trend δ is the *only* assumption required to get the ATT.

SCQE: Arguments about δ

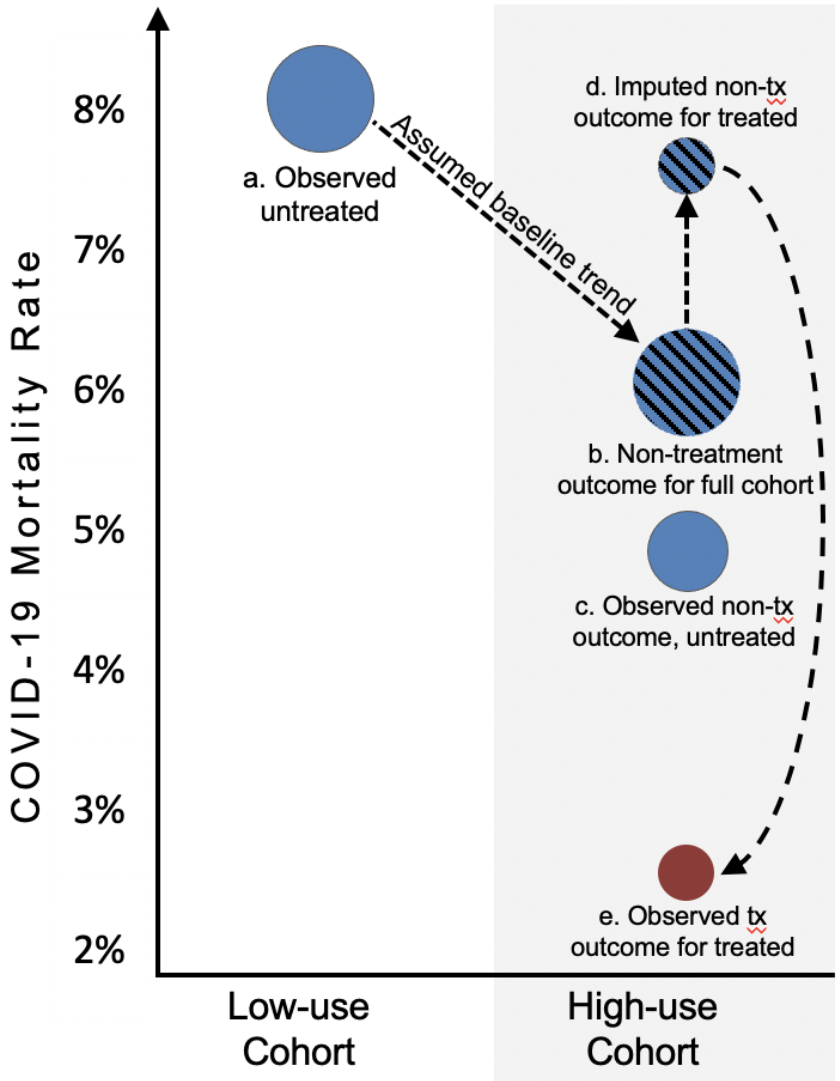
You don't know δ and it doesn't live in the data. But, you can reason about it through:

- Hard bounds (mortality rate between 0 and 1)
- Reasonable (disputable) bounds: mortality almost surely didn't double, or halve, from one month to the next; if it had we would known.
- Informed bounds: based on observed changes in patient characteristics, other treatments, and expert judgment, you might guess that mortality (absent dexamethasone) likely dropped by...

We are **not** looking for exact values for δ. Rather we will learn what we can(not) conclude based on:

- defensible ranges
- extremes you can rule out, or
- ranges where you cannot be sure either way...

How does δ+data tell you the ATT?



- We observe (a), $\mathbb{E}[Y(0)|Z=0]$. For a postulated δ, you get out (postulated) $\mathbb{E}[Y(0)|Z=1]$ (b).
- Meanwhile in later cohort, you observe $\mathbb{E}[Y(0)|Z=1,D=0]$ (c).
- Then you can back out $\mathbb{E}[Y(0)|Z=1,D=1]$ (d).
- We observe $\mathbb{E}[Y(1)|Z=1,D=1]$ (e), and that minus (d) is the ATT.

Formalizing,

Recall the change in non-treatment average outcomes,

$$\delta \equiv \mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]. \tag{1}$$

By LIE,

$$\mathbb{E}[Y(0)|Z=1] = \mathbb{E}[Y(0)|Z=1, D=1]\pi_1 + \mathbb{E}[Y(0)|Z=1, D=0](1-\pi_1)$$

So we have

$$\mathbb{E}[Y(0)|Z=0] + \delta = \mathbb{E}[Y(0)|Z=1, D=1]\pi_1 + \mathbb{E}[Y(0)|Z=1, D=0](1-\pi_1) \tag{2}$$

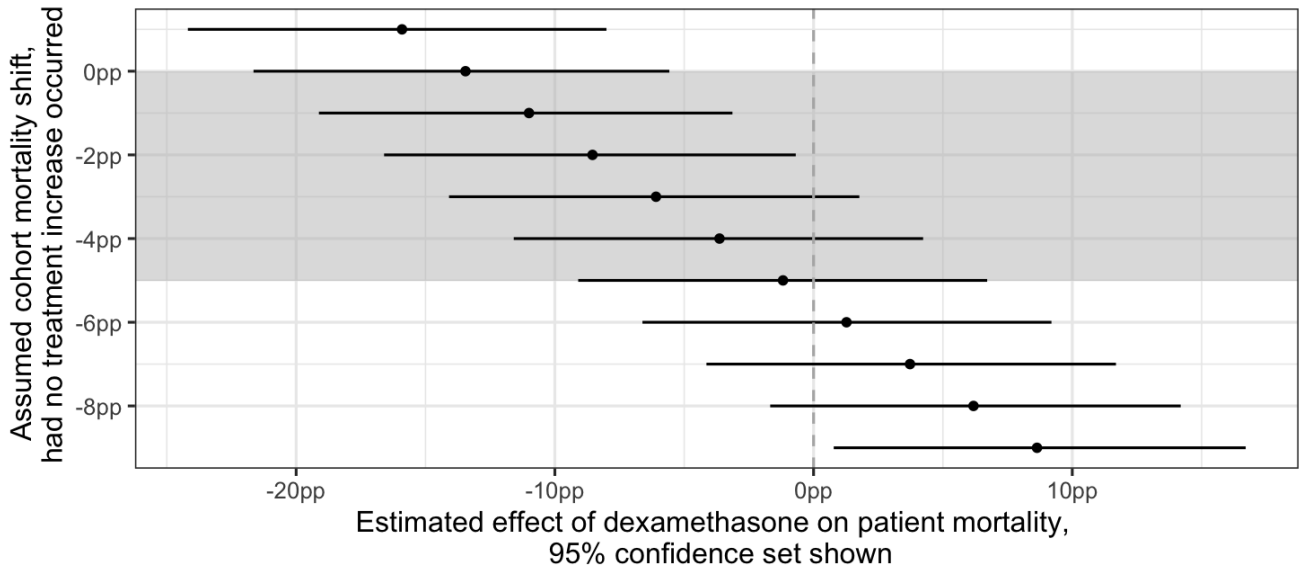
$$\mathbb{E}[Y(0)|D=1, Z=1] = \frac{\mathbb{E}[Y|Z=0] - \mathbb{E}[Y|D=0, Z=1](1-\pi_1) + \delta}{\pi_1} \tag{3}$$

So if you want the ATT at $Z=1$,

$$\begin{aligned} \text{ATT} &\equiv \mathbb{E}[Y(1)|D=1, Z=1] - \mathbb{E}[Y(0)|D=1, Z=1] \\ &= \mathbb{E}[Y|D=1, Z=1] - \left(\frac{\mathbb{E}[Y|Z=0] - \mathbb{E}[Y|D=0, Z=1](1-\pi_1) + \delta}{\pi_1} \right). \end{aligned} \tag{4}$$

Presenting results: Show it all!

We avoid *any* preferred estimate. Show the result (ATT) across wide range of values on δ, for example in the dexamethasone/COVID case we will see:



From this you can see ATT in any *ex ante* plausible δ, and note pivot points:

- Beneficial ($p < 0.05$) if baseline mortality increased, flat, or fell by up to 2.3pp (-21%)
- Non-harmful point estimate until a drop of 5pp (-46%)
- Harmful ($p < 0.05$) only if baseline mortality fell by 6.8pp (-93%), leaving mortality of just 0.4%

You already know two special cases of SCQE

Two well-known methods are special examples you can find on this plot:

- **Instrumental variables** $=$ SCQE at $\delta = 0$. The cohort/group indicator is the instrument. IV *assumes* the baseline trend is exactly zero.

- **Difference-in-differences**.
  - Examples like ours will not allow DiD (no "would be treated" units in those with $Z = 0$).
  - But if you did have that structure, SCQE at $\delta =$ the control group's observed change is precisely DiD.

Both pick one δ and pretend you believe only that. SCQE shows the whole line, then asks which δ you can defend or dismiss.

SCQE example 1: Tuberculosis prevention

- Data on tuberculosis status of 26,715 HIV-positive patients in Tanzania.
- National policy made it mandatory to provide isoniazid preventive therapy (IPT) to patients to prevent tuberculosis (2011)
- Roll-out varied by clinic but for each there is a period ($Z = 0$) before any patients given IPT, and then a period ($Z = 1$) after which some non-random fraction get it.
- After IPT policy began: **16%** of untreated developed TB vs **0.5%** of treated. Naive gap is thus **15.5pp** lower risk ($p \ll 0.001$).
- Conventionally, it looks great on magnitude and significance, and covariate adjustment barely moves it.

SCQE example 1: Tuberculosis prevention (cont.)

But none of that assures us of low bias.

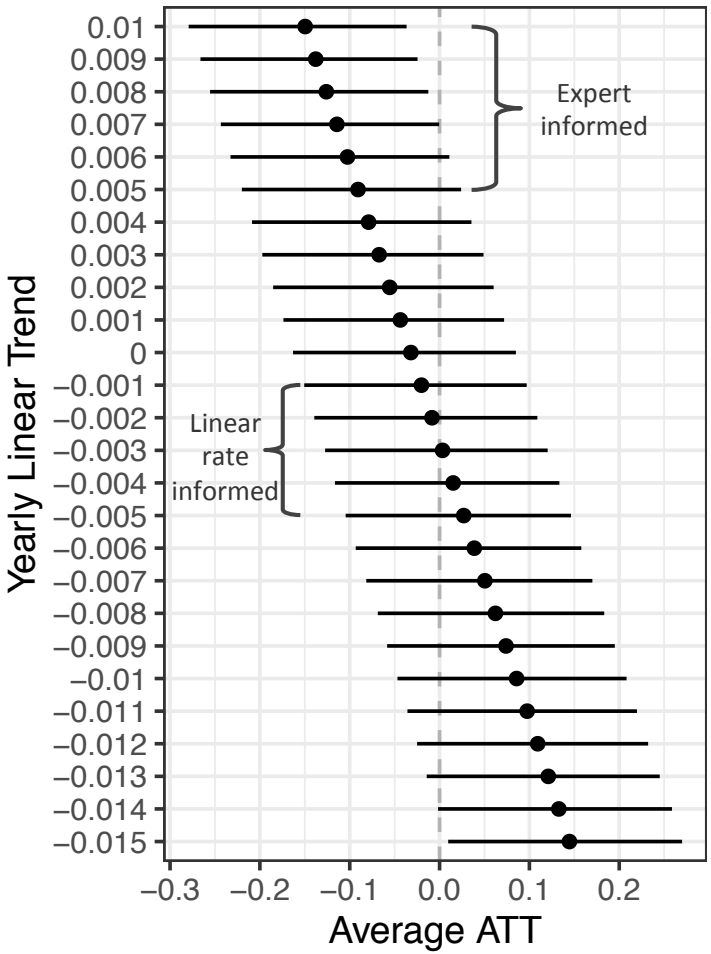
- What if IPT went disproportionately to healthier, lower-risk patients?

Let's use SCQE!

- *Ex ante* our physician colleague argued that (absent IPT) TB rates were increase at annual rate of 0.5% to 1%.
- If we try to learn rate of TB growth from periods before IPT in each clinic, we see -0.05% to -0.1%.
- We can look at those but also just see what values of δ would have to be believed to reach a given conclusion (beneficial, null, harmful)

Let's see everything at once on the SCQE plot.

SCQE result for TB/IPT:



- Over expert informed δ range, maybe helps.
- But over pre-trend informed, nothing.
- At $\delta = 0$, nothing. (-4%, not sig.)
- To claim IPT *helped*, you must argue non-treatment TB incidence would have **increased** between no-use and high-use cohorts by $\geq$ 0.7pp per year: *possible, but far from certain.*
- Conclusion: instead of claiming the **15.5pp** lower TB in IPT patients was an effect, SCQE says: you'd need strong beliefs about δ, not in keeping with the evidence, to think that was an effect.

You can do all this yourself: `scqe` on CRAN or in a web browser app at `https://chadhazlett.com/scqe-app/`

Glioblastoma and DCVax: Dismissing a life-saver?

Glioblastoma (GBM) has a grim prognosis: median survival of 14-15 months; fewer than 5% survive 5 years.

DCVax-L dendritic-cell vaccine (Liau et al., *JAMA Oncology* 2023).

Large randomized trial, compromised:

- primary endpoint (PFS) abandoned (pseudoprogression);
- ∼90% *crossover* gutted the placebo arm;
- Pivoted to a 640-patient matched **external control arm**.
- Result: 13.0% of ever-treated alive at 5 years, almost double that in the ECA (5.7%)

This was not well received...dismissed as uninformative.

Safe inference: don't dismiss on procedural grounds nor accept on flimsy evidence, ask what can(not) conclude under clear assumptions.

What would you have to believe?

This is a "one-cohort" SCQE (eligible group, 90% eventually treated).

Assumption needed on: mean non-treatment in that group, $\mathbb{E}[Y(0)]$.

- Recall: 5-year (non-treatment) survival 5.7% in the ECA[1].

SCQE results in a nutshell:

- DCVax effect on 5-year survival is positive and statistically significant as long as $\mathbb{E}[Y(0)]$ in treatment eligible group is $\leq 9.4\%$.
- Positive but looses significance until $\mathbb{E}[Y(0)] > 13\%$.
- Cannot be significantly harmful unless $\mathbb{E}[Y(0)] > 17\%$, triple the ECA.

Not a slam dunk, but far from dismissable.
What sort of design should come next?
Who should have access in the meantime?

[1] The usual figure is very similar at 5%, Stupp 2005

SCQE: dexamethasone for COVID-19

Cohorts:[2]

- "Low-use" (first) cohort: days 44–101, **5.7%** took dexamethasone
- "High-use" (second) cohort: days 102–200, **46%** took dexamethasone

Outcome: 14-day mortality

- **10.9%** in the low-use (first) cohort
- **5.4%** in the high-use (second) cohort

On its surface: big uptick in treatment, big drop in mortality...promising but could it just be the *baseline trend*?

[2] Wulf, Hazlett, Hill, Chiang, Goodman-Meza, Pasaniuc, Arah, Erlandson, Montague. "Safe inference outside of randomized trials: Application of the stability-controlled quasi-experiment to the effects of three COVID-19 therapies." *Observational Studies* 11(3), 2025, 301–330.

Dexamethasone: δ considerations

Before looking at the SCQE results, let's think about where δ might be.

Baseline characteristics. High-use cohort had *fewer* ICU admissions, ventilator use, skilled-care admits within 24h — all consistent with a drop in baseline mortality.

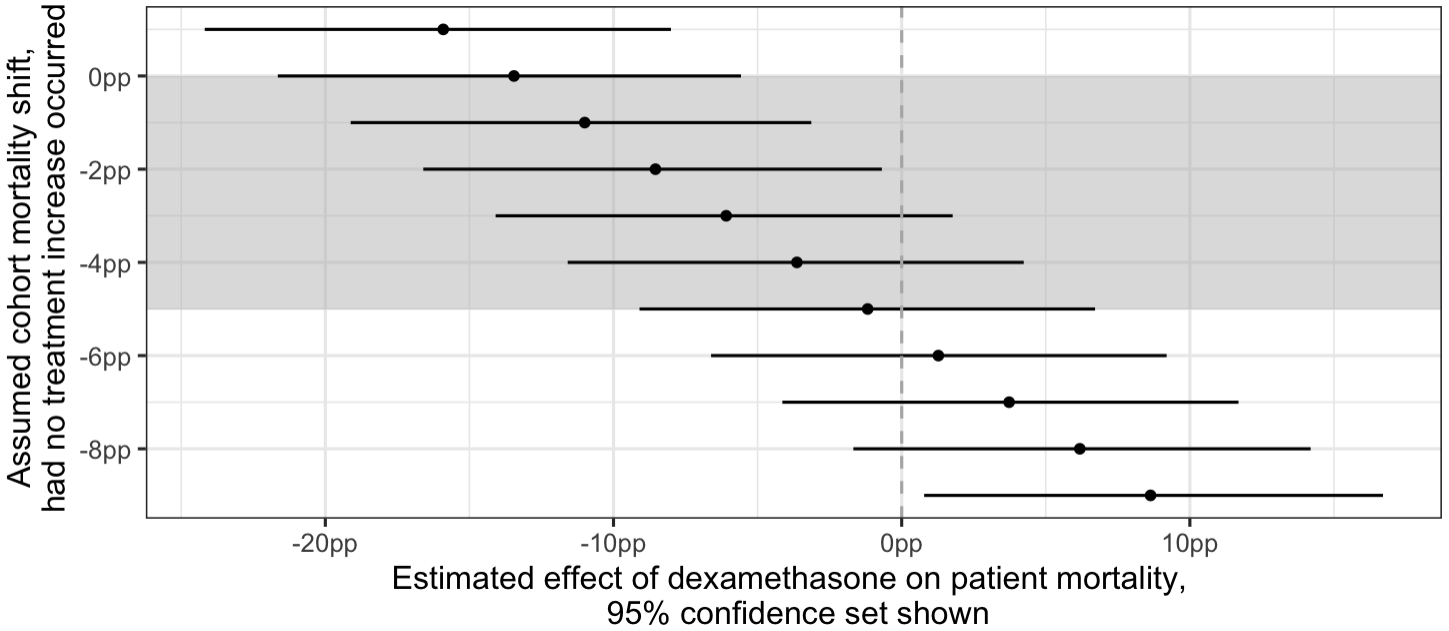
Other therapies. Higher remdesivir use in the high-use cohort (12% vs. 27%). Could reduce risk; unknown at the time.

Modeled mortality risk on observables. Predicted baseline mortality fell by 1.3–2.1pp.

Expert judgement. Physicians on the team expected baseline mortality to drop with generally improving practice over this period.

> *"All lines of evidence suggest a possible drop in baseline mortality. Taking these considerations collectively, and maintaining a conservative attitude, we postulated that a plausible range for the baseline trend ranged from no change to a reduction of up to 5pp (-46%)."*

Dexamethasone: results



- Beneficial ($p < 0.05$) if baseline mortality increased, flat, or fell by up to 2.3pp (-21%)
- Non-harmful point estimate until a drop of 5pp (-46%)
- Harmful ($p < 0.05$) only if baseline mortality fell by 6.8pp (-93%), leaving mortality of just 0.4%

Dexamethasone: what we concluded

> *"We consider [the drop in baseline required to support a harmful effect] to be extremely unlikely, particularly without a known cause that escaped the attention of physicians on the team and does not appear in the observable differences in cohorts."*

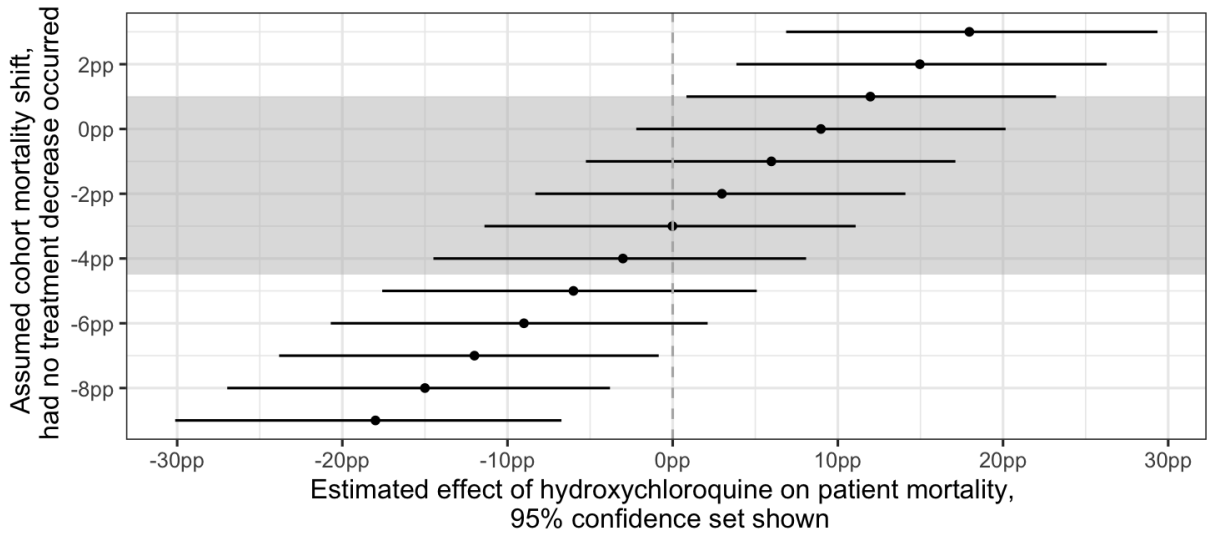
> *"Thus, we have considerable confidence that dexamethasone was at least directionally, and possibly statistically significantly beneficial, and that it was not significantly harmful."*

Consistent with the eventual results of the RECOVERY-trial.

Hydroxychloroquine (HCQ): results

Cohort shift: 36% HCQ use in high-use early cohort $\rightarrow$ 2.9% in low-use cohort. Mortality 11.6% $\rightarrow$ 8.6%.

Observables/ expertise suggested: "a wide range of baseline trends are possible, from +1pp to -4.5pp."



Wide plausible range of δ in which you would conclude HCQ was **harmful**.

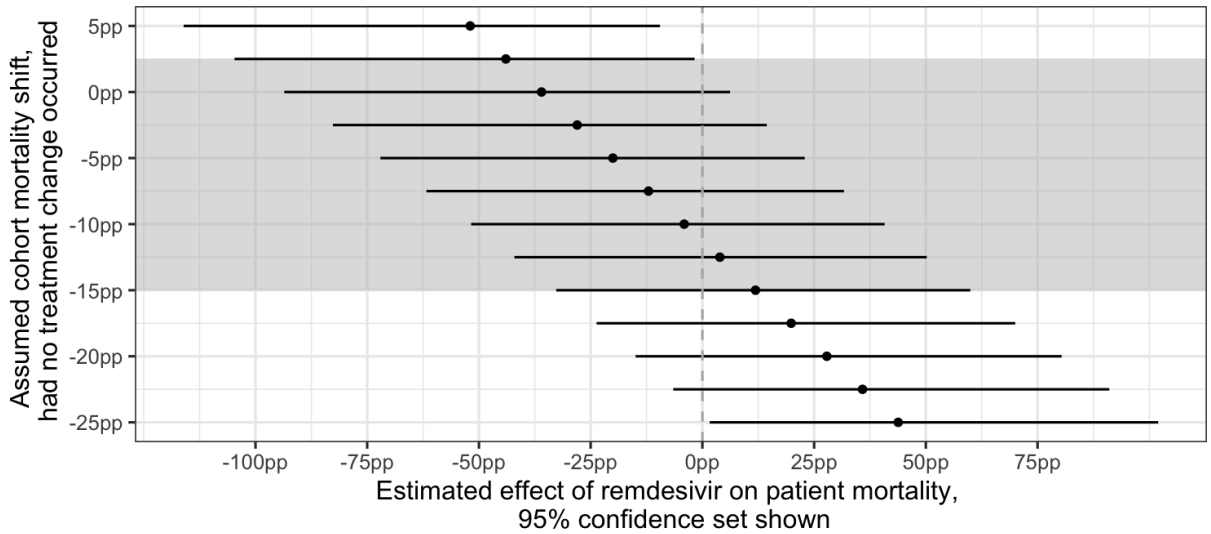
Sig. harmful if baseline mortality worsened by even 0.7pp.

Sig. beneficial only if baseline improved by $\geq$6.7pp (-46%): not credible.

Remdesivir: results

Cohort: 0% remdesivir use in no-use cohort $\rightarrow$ 31.3% in high-use cohort. Mortality 24.5% $\rightarrow$ 13.3%.

Observables and clinical judgment: plausible range $\delta \in [+2.5\text{pp}, -15\text{pp}]$.



While benefit is plausible, cannot rule out harm (on point estimates).

Can rule out statistically significant harm – would imply 0% counterfactual mortality.

Three COVID treatments: Conclusions

- **Dexamethasone:** likely beneficial or null. Harm would require an implausibly large drop in baseline mortality.
- **HCQ:** Could be (sig.) harmful with small positive δ. Sig. benefit would require larger drop than deemed plausible.
- **Remdesivir:** sig. benefit is plausible but so is harmful (point) estimate. Not sig. harmful.

If you had to decide what to do, while waiting on RCT, you'd have some safe guidance without asserting "no confounding".

Every result consistent with eventual RCTs: rem/dex help; HCQ doesn't.

Why not just adjust, and add a sensitivity analysis?

Sensitivity analysis can also sit within safe inference, asking for assumptions on the strength of unobserved confounding.

Covariate adjustment + OVB sensitivity (Cinelli & Hazlett 2020):

- Confounding explaining as little as $\sim$**1%** of residual variance can break any result.
- Given how little we know about who got treated, we can't rule out confounding even that small $\Rightarrow$ **no conclusion survives**.

SCQE bounds the *baseline trend* instead.

- This is arguable more tangible but also happens to give us more leverage
- E.g. we can rule out significant harm from dexamethasone, since it needs an implausible $>$ 80% drop in baseline mortality.

In principle any (transparent) "postulate-and-see" approach can work; use what gives the most traction in a given setting.

Across disciplines, and settings

While I've given medical examples, safe inference is badly needed throughout the social sciences where we often don't have or wouldn't want to do RCTs.

1. The conventional regression-style, cross-sectional controlled comparison.
   - Sensitivity to unobserved confounding widely applicable, sometimes informative.
   - E.g. "From "is it unconfounded?" to "how much confounding would it take?"..." (Hazlett & Parente, JOP 2023).

2. SCQE comes up with self-selection into new policies and programs, many cases where DiD otherwise used.
   - Recent and upcoming studies using SCQE to study gun control/storage laws, pet ownership, govenment incentive programs...

3. Safe synthetic control (ongoing work)
   - Synthetic control carries tough assumptions and finite-sample biases.
   - We're exploring how tolerable assumptions can be imposed to produce safe, partial identification.

Trialism has a cost

When we value *procedural adherence* or assess credibility based on what was done rather than what we can reason about, defensible knowledge is thrown away.

A treatment that may have saved lives — **GBM** / **DCVax** — was dismissed: not because we do not know $Y(0)$ (and thus the ATT), but because the trial didn't adhere to procedural guidelines.

An RCT's value is proportional to our ignorance of $Y(0)$.

- Where $Y(0)$ is knowable by other means, demanding the trial anyway isn't rigorous, it is costly.
- Where existing evidence is not sufficient, there may be "designed one-arm trials" or "designed SCQEs" that can tighten inference still without randomization.

The cost includes the studies not conducted, results not trusted, and defensible treatments not adopted.

Safe inference: Wrap-up

Credibility is about the *assumptions*, not the procedure (thesis 3): "was it randomized?" is the wrong test — "what can we defend, and where does that leave the answer?" is the right one.

There are errors on both sides: over-claiming, *and* discarding what we could defend. We are often blind to the second.

Safe inference avoids both by saying just what we know and naming what you'd have to assume to say more.

The flip: instead of a single number that is meaningful under heroic assumptions, reveal the defensible *band*, an honest *if–then*, or just "we can't tell yet."

A program, and an invitation

We've built methods for doing this in some scenarios including sensitivity analysis and SCQE. But we need more!

The creative opening for methodological development here is in finding assumptions we actually *reason about* and that give us *traction* in a given setting, narrowing what we don't know about $Y(0)$.

Not just new ID strategies, but also new designs: Where we cannot randomize, what choices can we make to otherwise reduce uncertainty over $Y(0)$ in the treated group?

Finally: If you are interested in following developments and discussions in safe inference and "practical causal inference" more generally, see our lab and sign up for our (almost) weekly talk series at (`practicallycausal.com`)