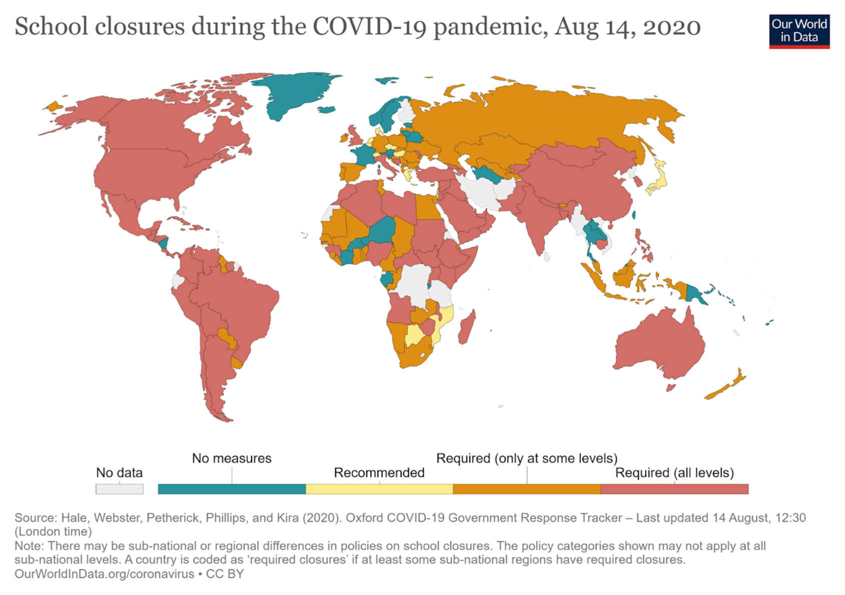
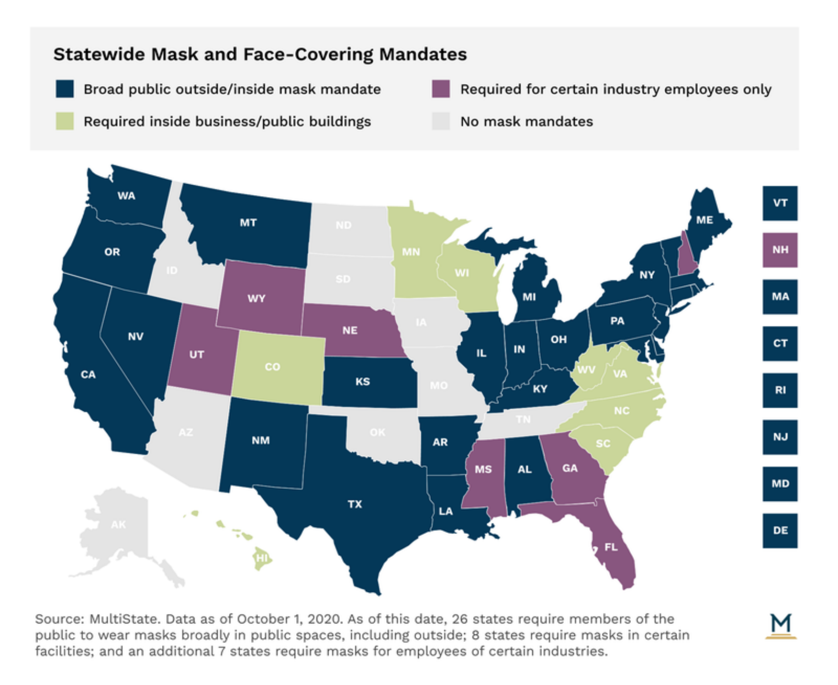
The need for, and challenges of, policy evaluation during the COVID-19 pandemic

Elizabeth A. Stuart, PhD
Bloomberg Professor of American Health
www.elizabethstuart.org
@lizstuartdc

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH


School closures during the COVID-19 pandemic, Aug 14, 2020

Source: Hale, Webster, Petherick, Phillips, and Kira (2020). Oxford COVID-19 Government Response Tracker – Last updated 14 August, 12:30 (London time)
Note: There may be sub-national or regional differences in policies on school closures. The policy categories shown may not apply at all sub-national levels. A country is coded as 'required closures' if at least some sub-national regions have required closures.
OurWorldInData.org/coronavirus • CC BY


Statewide Mask and Face-Covering Mandates

Broad public outside/inside mask mandate
Required for certain industry employees only
Required inside business/public buildings
No mask mandates
WA
OR
CA
NV
ID
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
GA
FL
WV
VA
NC
SC
PA
NY
ME
AK
HI
VT
NH
MA
CT
RI
NJ
MD
DE
M

Source: MultiState. Data as of October 1, 2020. As of this date, 26 states require members of the public to wear masks broadly in public spaces, including outside; 8 states require masks in certain facilities; and an additional 7 states require masks for employees of certain industries.

How can we learn about the effects of programs and policies?

- The country, and world, is full of variation in local policies being used to address the COVID-19 pandemic
- Could create an opportunity to learn about the effects of those policies, to inform future decision-making
- We have data on policies, outcomes, etc….what could be the problem?

Causal inference is hard

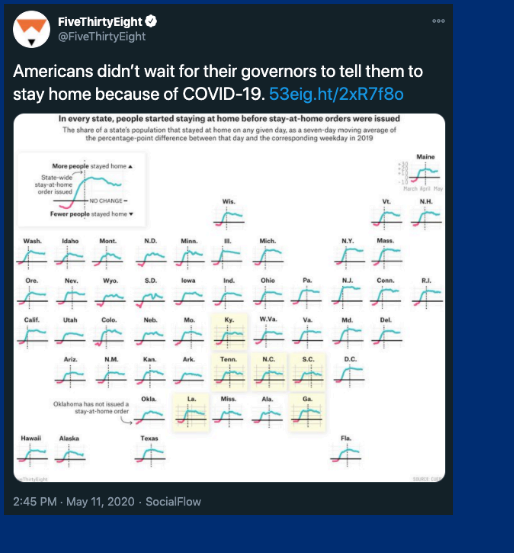
- Need to be able to compare potential outcomes for a well defined population:
  - Y(1): Outcome if treated (exposed)
  - Y(0): Outcome if control (not exposed)
- e.g., Difference in infection rates if a community has a mask mandate vs. does not have a mask mandate
- The "fundamental problem of causal inference" is that we only see one of these potential outcomes for each unit (community)

Causal inference for policy evaluation is really hard

- Can’t randomize to exposure conditions
- Often relatively few units (states, countries)
- Implementation hard to measure (does the policy mean the same thing everywhere?)
- Hard to tease out effects from other things happening, including multiple policy responses

Causal inference for policy evaluation during COVID-19 is *really really* hard

- Infectious diseases spread exponentially and have incubation periods
  - Small differences in model assumptions can have dramatic effects on results
- LOTS of policies and programs being put in place
  - Masks, schools, workplaces, stay at home, rapidly evolving treatments
- Anticipatory actions
  - e.g., staying at home before official orders to do so
- Data challenges
  - e.g., changing test availability and use
- Interactions across communities matter a lot
  - e.g., Sturgis rally



FiveThirtyEight
@FiveThirtyEight

Americans didn't wait for their governors to tell them to stay home because of COVID-19. 53eig.ht/2xR7f8o

In every state, people started staying at home before stay-at-home orders were issued

The share of a state's population that stayed at home on any given day, as a seven-day moving average of the percentage-point difference between that day and the corresponding weekday in 2019

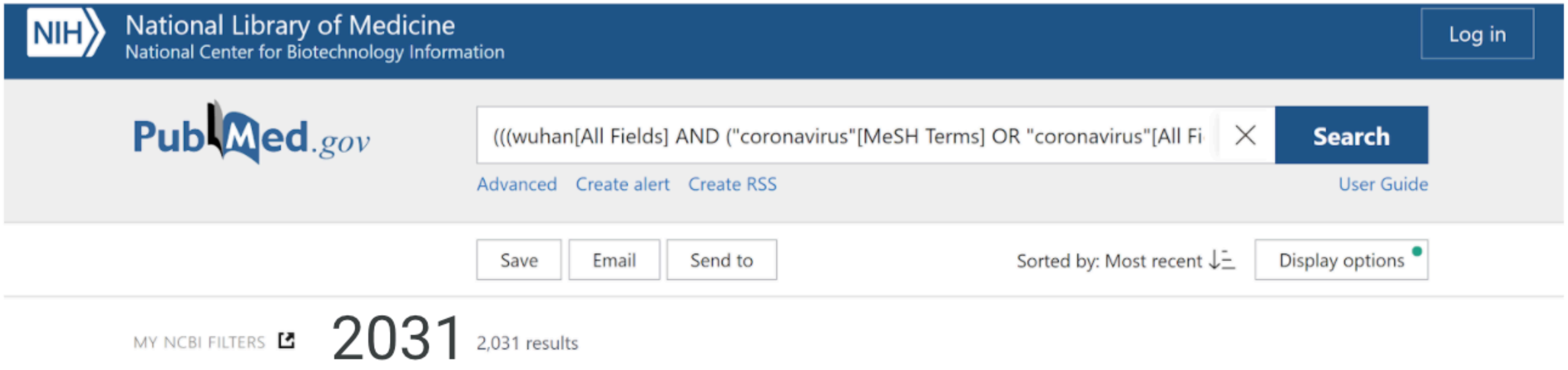
2:45 PM · May 11, 2020 · SocialFlow

But people are trying…

(((wuhan[All Fields] AND ("coronavirus"[MeSH Terms] OR "coronavirus"[All Fields])) AND 2019/12[PDAT] : 2030[PDAT]) OR 2019-nCoV[All Fields] OR 2019nCoV[All Fields] OR COVID-19[All Fields] OR SARS-CoV-2[All Fields])

AND (countries OR states OR counties OR regions)

AND (policy OR policies)



(((wuhan[All Fields] AND ("coronavirus"[MeSH Terms] OR "coronavirus"[All Fi

Sorted by: Most recent

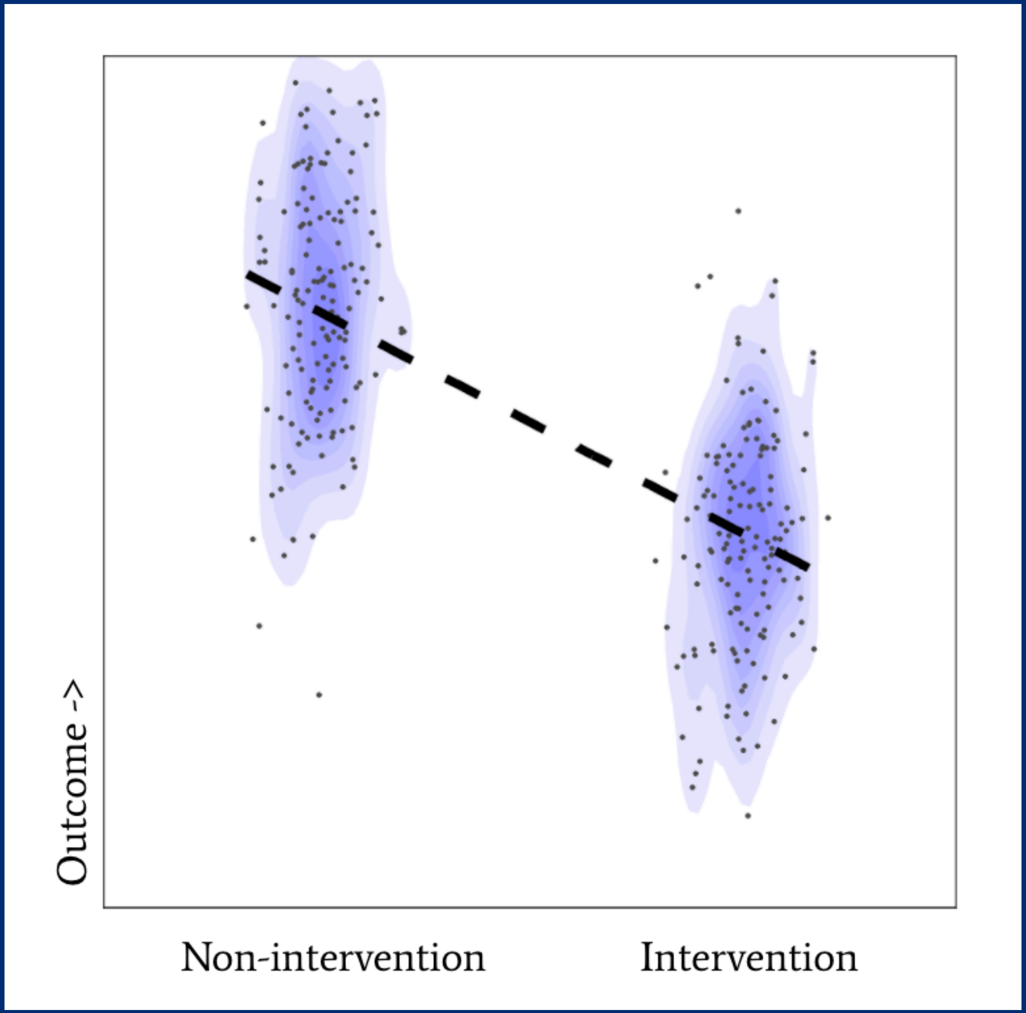
2031 2,031 results

(Search thanks to Noah Haber)

So what can we do?

What does a good (or bad) policy evaluation look like?

Cross sectional two group comparison



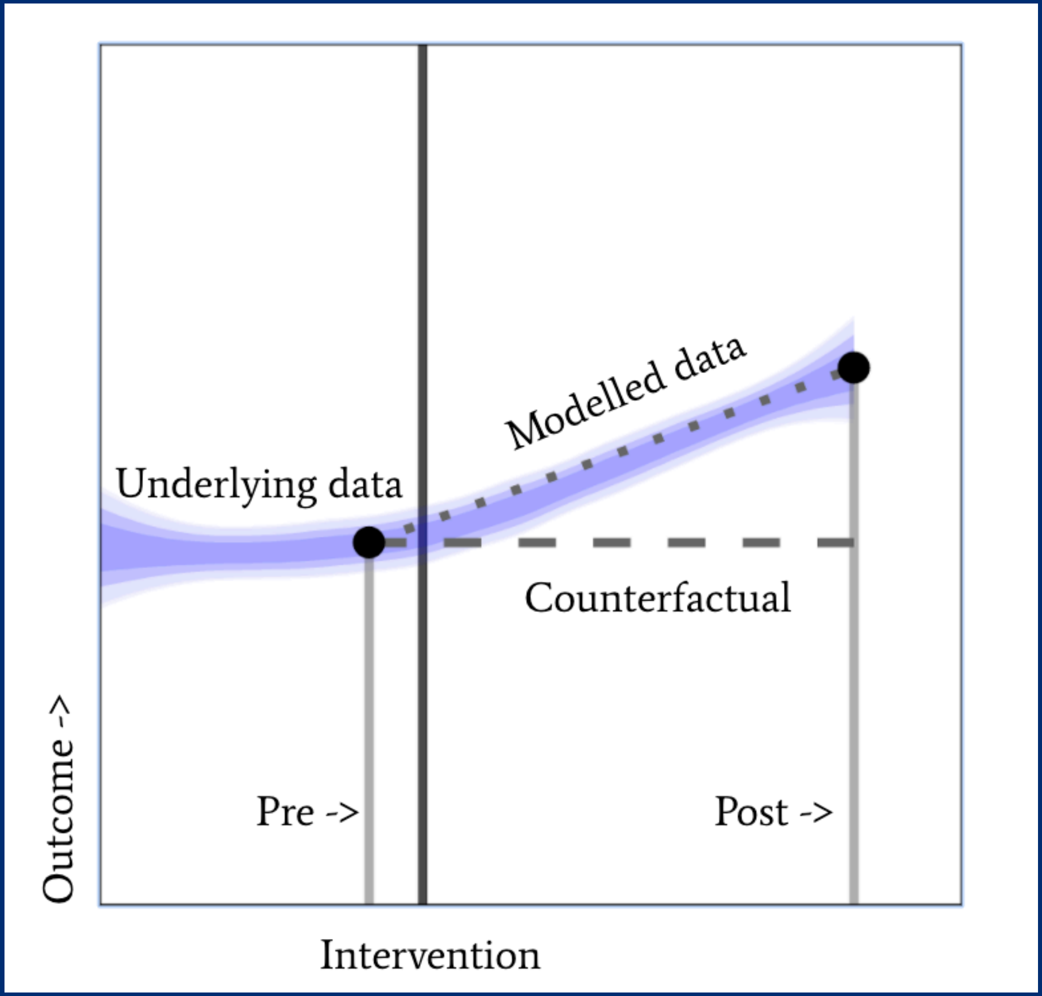
Compares outcomes between exposed and unexposed groups

Assumed counterfactual: Non-intervention location provides estimate of Y(0) for intervention location

Highly susceptible to confounding: locations that implement a policy likely quite different from those that don’t

And subtle differences (e.g., in R0) may make a big difference in policy impact estimates, esp. given exponential growth

Simple pre/post

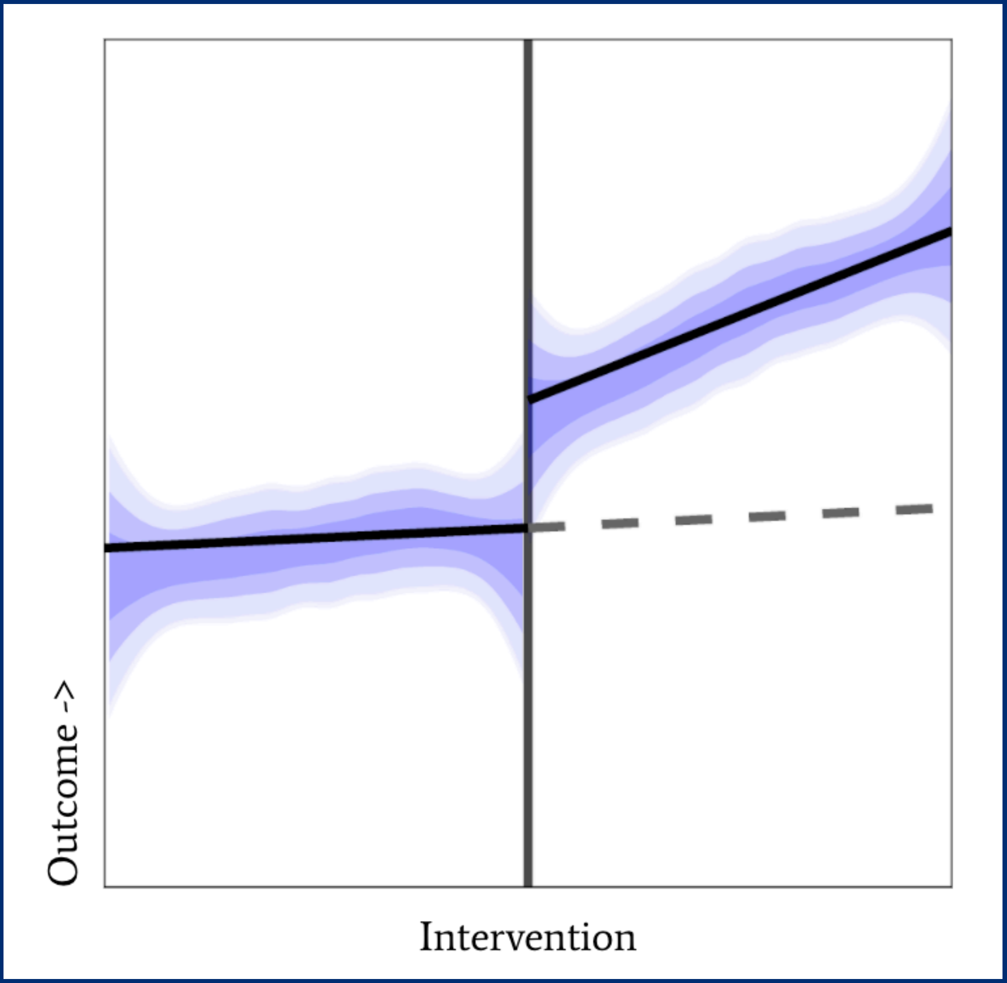


Compares outcome levels at two time points

Assumed counterfactual:
All change over time is due to the intervention

VERY questionable in infectious diseases (and many other areas without stable outcomes in the pre period)

Interrupted time series

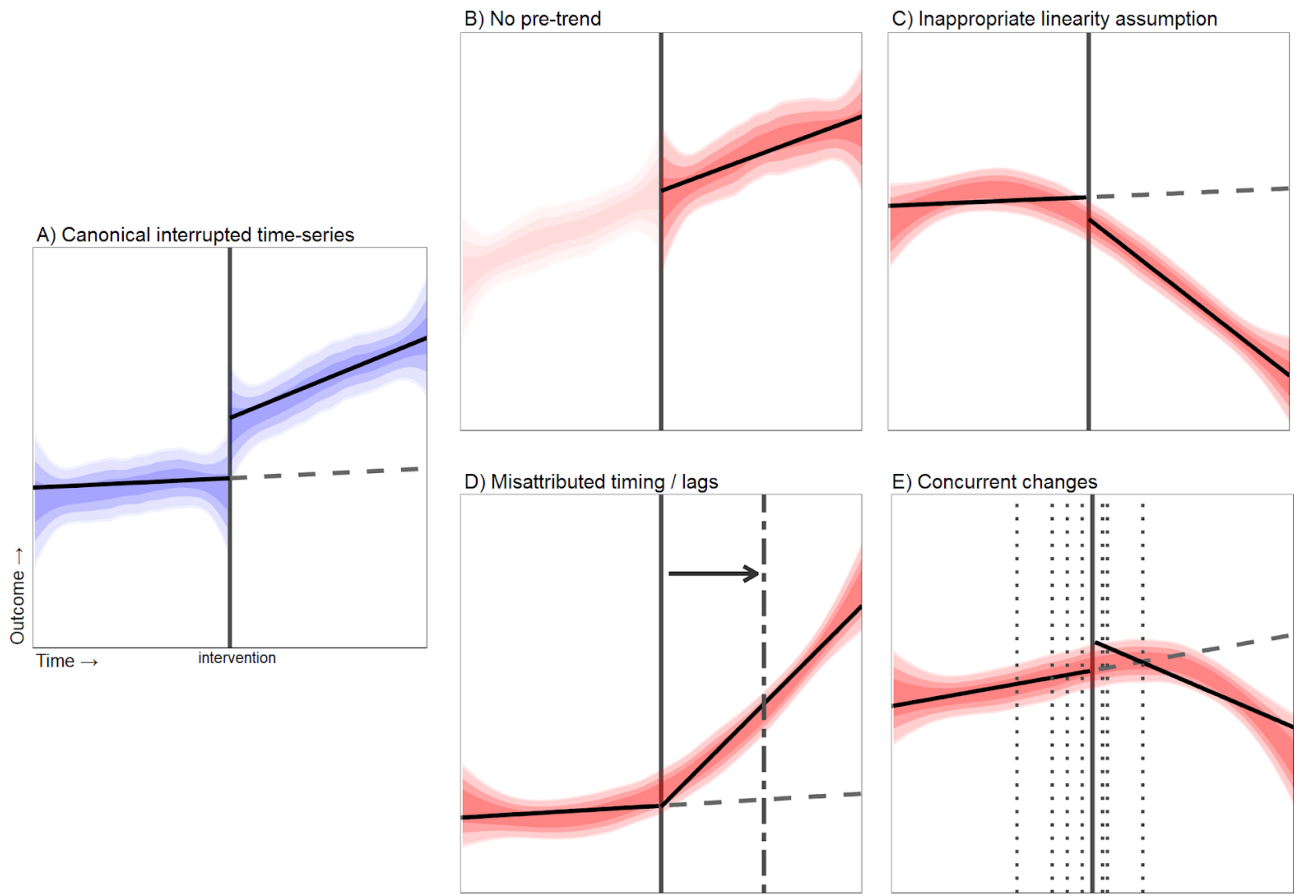


Models outcome in “pre” period and projects that out into the post period

Assumed counterfactual: outcome would have continued on the same (modeled) trajectory, if not for the intervention

Possibly better than pre/post, but relies on ability of the model to project accurately into the future

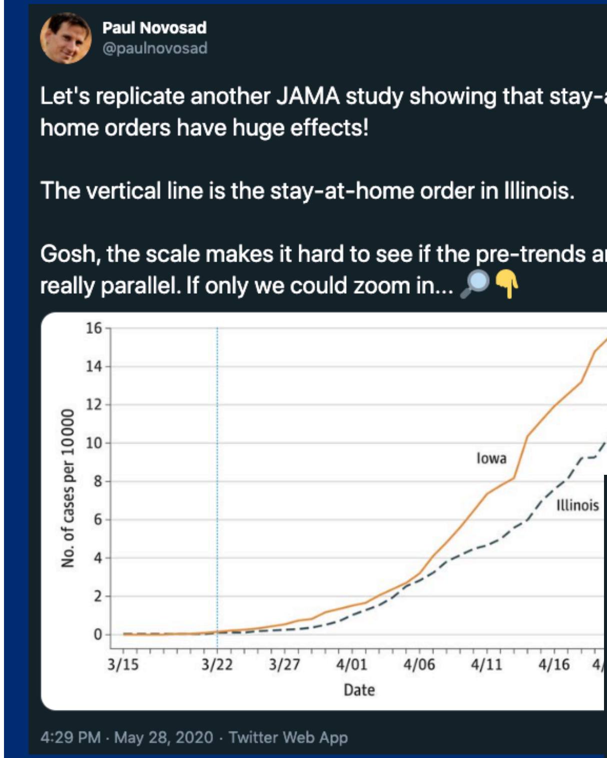
Issues that can come up in ITS in COVID-19



Linearity assumptions can be particularly challenging

Infectious disease dynamics:

- Almost never linear
- “Exponential”
-
- “S curve”
- “Flattening”



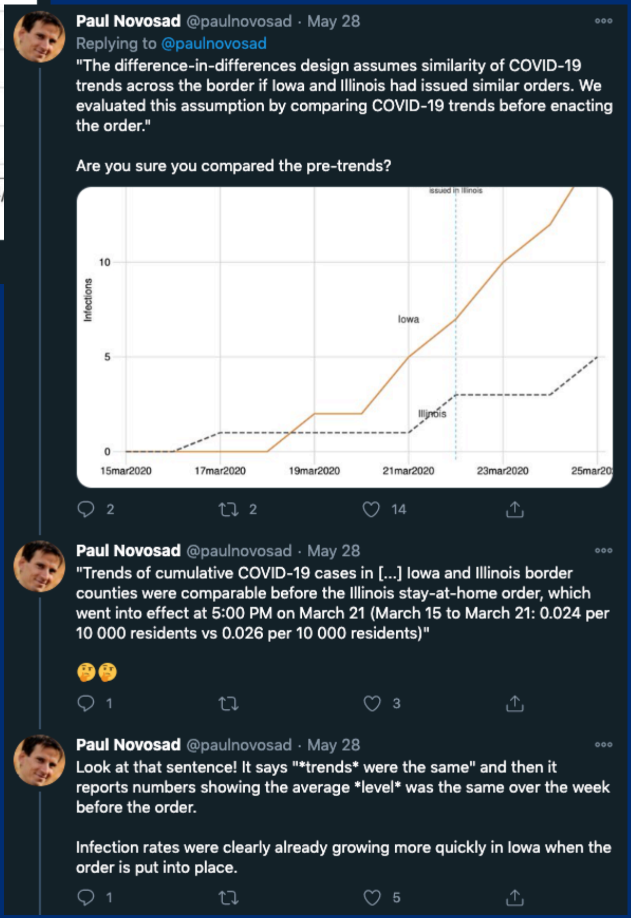
Paul Novosad @paulnovosad

Let's replicate another JAMA study showing that stay-at-home orders have huge effects!

The vertical line is the stay-at-home order in Illinois.

Gosh, the scale makes it hard to see if the pre-trends are really parallel. If only we could zoom in... 🔎👇

4:29 PM · May 28, 2020 · Twitter Web App



Paul Novosad @paulnovosad · May 28

Replying to @paulnovosad

"The difference-in-differences design assumes similarity of COVID-19 trends across the border if Iowa and Illinois had issued similar orders. We evaluated this assumption by comparing COVID-19 trends before enacting the order."

Are you sure you compared the pre-trends?

💬 2 🔁 2 ♡ 14

Paul Novosad @paulnovosad · May 28

"Trends of cumulative COVID-19 cases in [...] Iowa and Illinois border counties were comparable before the Illinois stay-at-home order, which went into effect at 5:00 PM on March 21 (March 15 to March 21: 0.024 per 10 000 residents vs 0.026 per 10 000 residents)"

🤔🤔

💬 1 🔁 ♡ 3

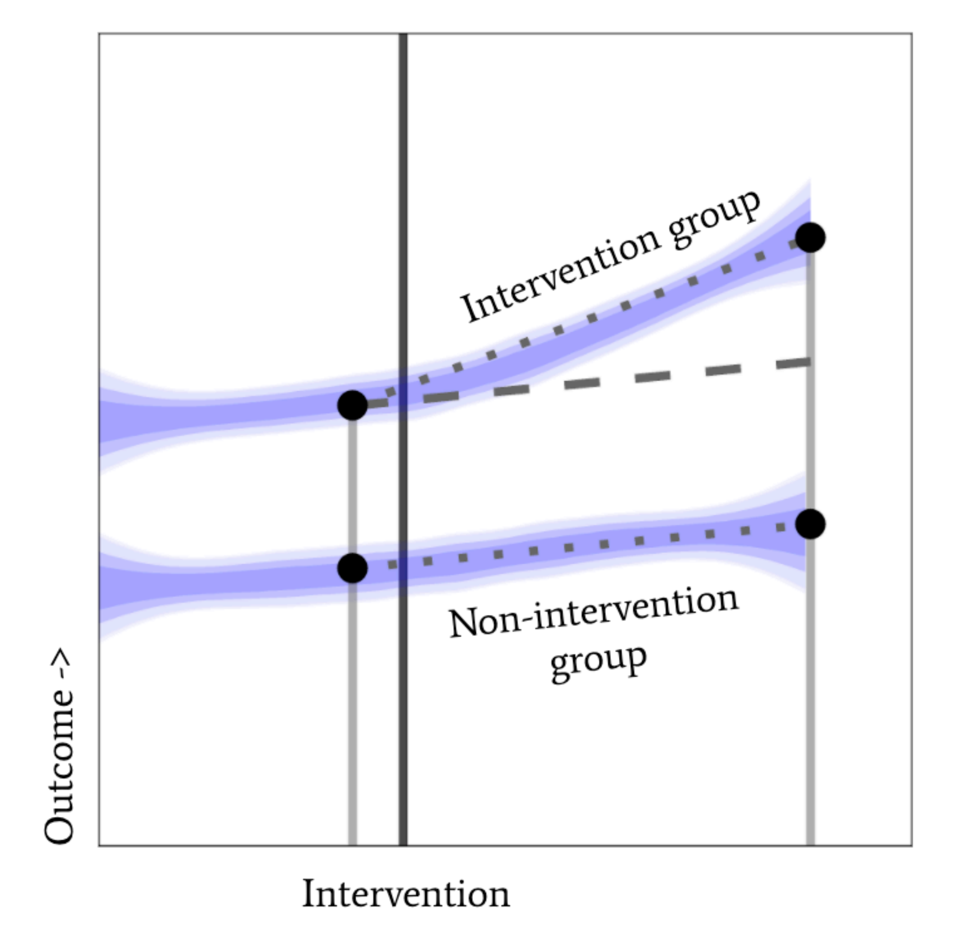
Paul Novosad @paulnovosad · May 28

Look at that sentence! It says "*trends* were the same" and then it reports numbers showing the average *level* was the same over the week before the order.

Infection rates were clearly already growing more quickly in Iowa when the order is put into place.

💬 1 🔁 ♡ 5

Comparative interrupted time series/Difference-in-differences



Models trends over time and across exposed and unexposed groups

Assumed counterfactual: Outcome would have changed in the same way pre to post as it did in the comparison group, if not for the intervention

Relies on a “parallel counterfactual trends” assumption

Note: Parallel pre-trends “make us feel better” but do not directly assess the key assumptions

Lots of nuances in CITS/DiD

- How to select comparison locations
- How to take advantage of individual level data
- Statistical power concerns (RAND, 2018)
- Challenges with staggered implementation across locations
    - e.g., standard "two-way fixed effects" models can lead to effect estimates of the wrong sign! (Goodman-Bacon, 2019)

- Plus all the challenges already mentioned…

- So a strong design, but still requires care

- BTW rapidly growing methods area, and lots of variations on this; also known as event study designs, group panel data, segmented regression, …

Case study:
Policy trial emulation

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Nested policy trial emulation

- Main idea: Think of staggered implementation of policies as nested implementation of hypothetical "target trials"
- At each policy implementation date, equate states that implement the policy with those that don't yet have the policy on a set of baseline characteristics, inc. baseline measures of the outcome
- Use traditional non-experimental study methods for each target trial
    - Ensures clear temporal ordering of covariates –> exposure -> outcomes

- Ben-Michael, Feller, and Stuart (forthcoming???)
- Extends ideas of Don Rubin, Paul Rosenbaum, Miguel Hernan on replicating a randomized trial using non-experimental data

Motivating example

- Estimating the effect of stay-at-home policies implemented in US states in late Spring 2020
- Data on policy enactment dates and COVID cases from New York Times tracker

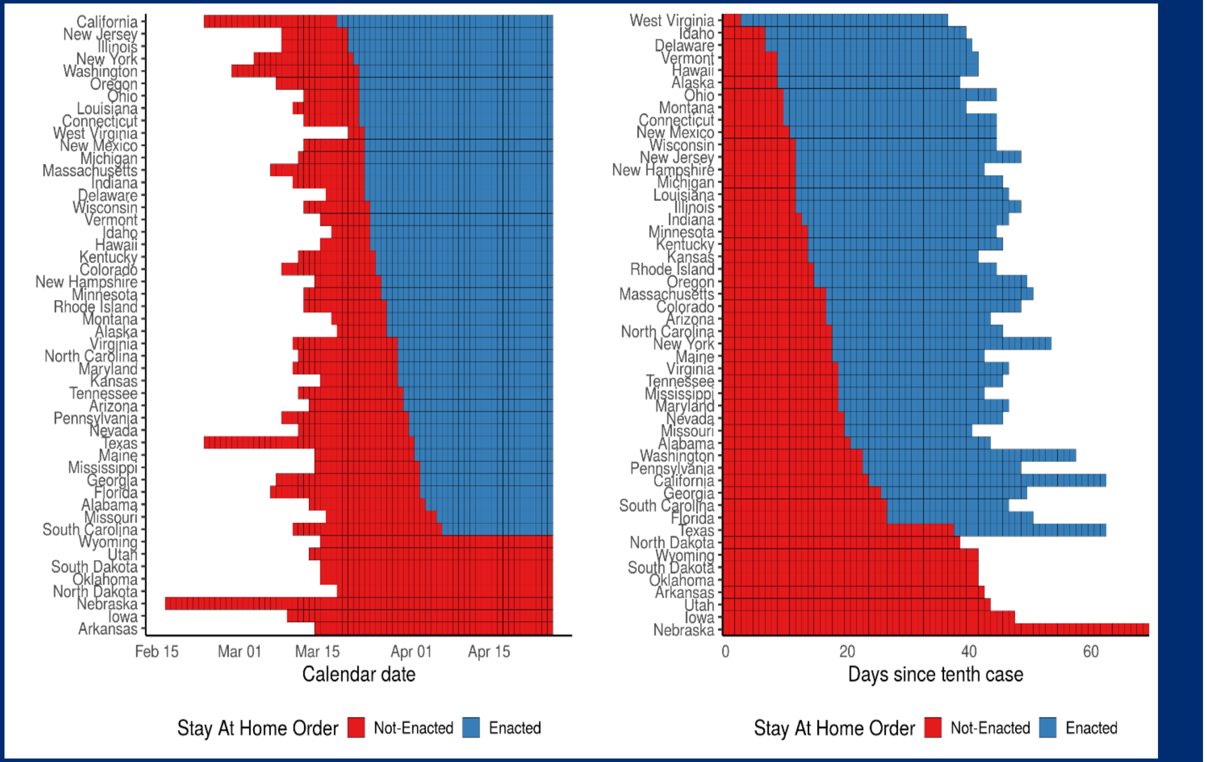
- Exposure: Implementing a statewide stay-at-home order
  - Think of analogous to intent-to-treat effect; ignores compliance
  - Also ignores spillovers and contagion

- Estimand of interest: comparison of outcomes with and without the policy at post-period time t ($Y_{it}(1)$-$Y_{it}(0)$), averaged across time and periods and across the states that implemented the policy

- Outcomes: (log) number of cases, log-ratio of case counts from previous day

Defining "time zero"

- For each policy implementation date, need to formally define "time zero" to determine what is "pre" and what is "post"
- In a pandemic, how we do so matters. Two choices:
  - Calendar time
  - Case time (time since 10th case)
  - We use calendar time for now



Target trial for March 23 enactment date

- Will compare the treated cohort to the 8 “never-treated” states
  - Could potentially use those “not yet treated”, which would change over time; we don’t do that here
- Basic difference-in-differences comparison
  - Compare changes over time across groups

Assumption: Parallel counterfactual Trends

Violated if:

- Any anticipatory effects of the order
- Time-varying confounding

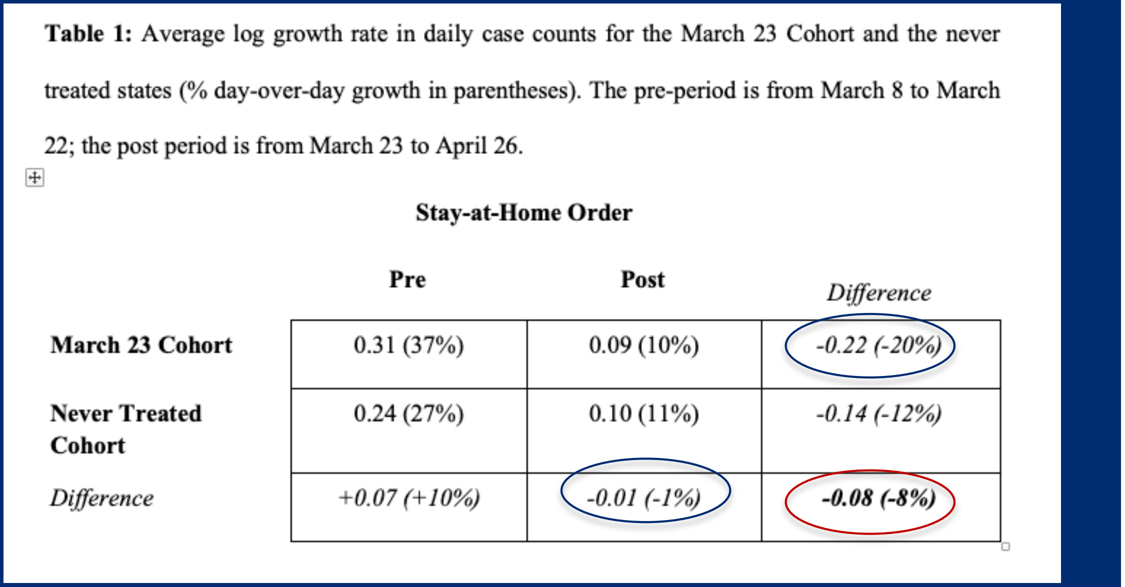


Table 1: Average log growth rate in daily case counts for the March 23 Cohort and the never treated states (% day-over-day growth in parentheses). The pre-period is from March 8 to March 22; the post period is from March 23 to April 26.

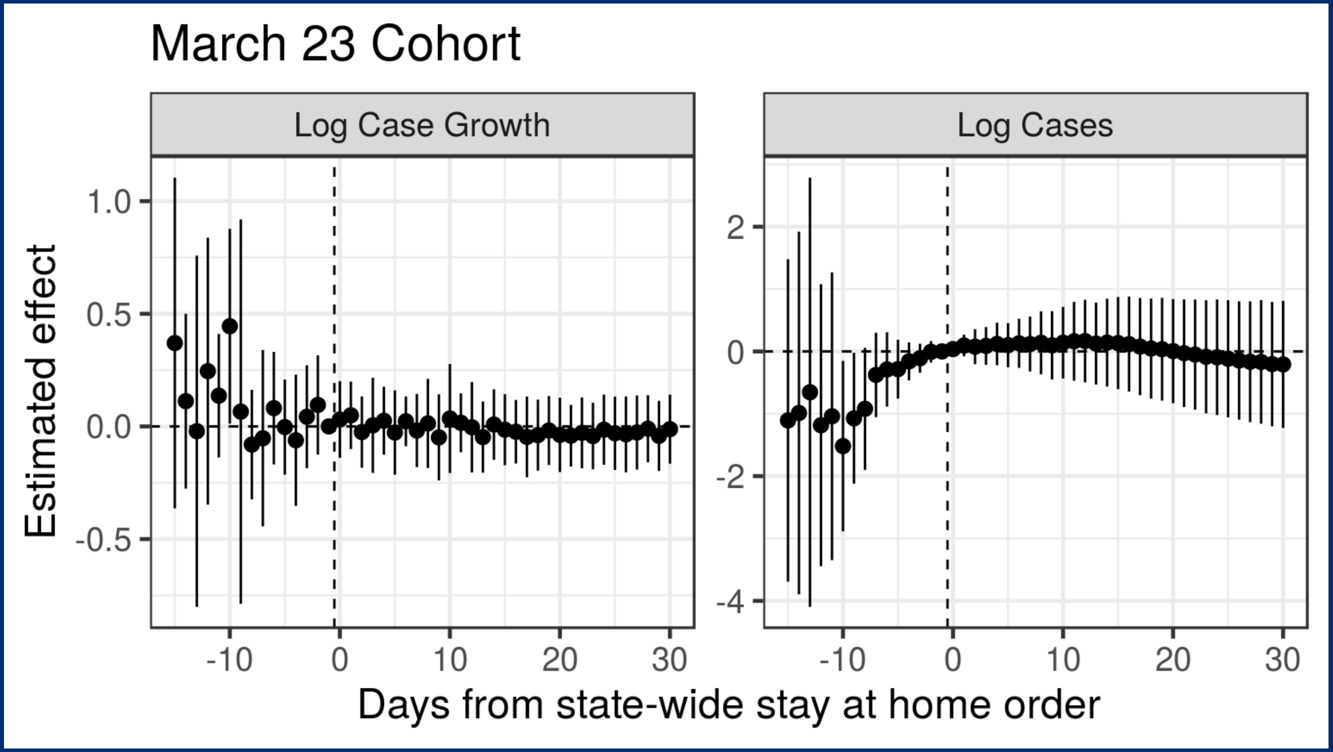
Stay-at-Home Order

	Pre	**Post**	*Difference*
March 23 Cohort	0.31 (37%)	0.09 (10%)	*-0.22 (-20%)*
Never Treated Cohort	0.24 (27%)	0.10 (11%)	*-0.14 (-12%)*
Difference	*+0.07 (+10%)*	*-0.01 (-1%)*	***-0.08 (-8%)***

Diagnostics

Can basically estimate the "effects" of the policy for each time period before and after the policy change

Like a balance check in the pre period: want to see no "effect"
[doesn't look great for either, especially log cases!]

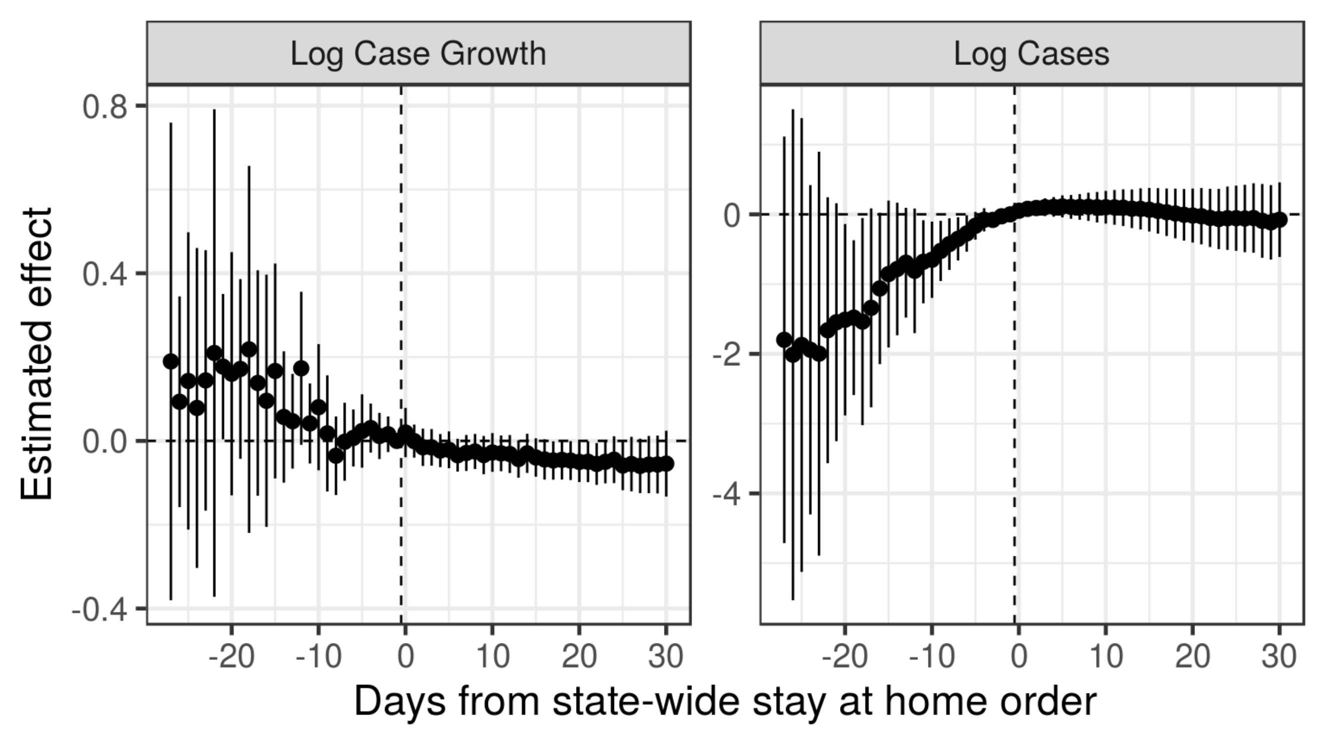


Nested target trials

Now basically repeat that for each policy implementation date and aggregate results across trials

Known as "stacking" or "event study" analysis

Equivalent to Abraham and Sun (2020), Callaway and Sant'Anna (2019) without any covariates



Discussion

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Additional thoughts…

- Important to be thoughtful and careful with policy evaluation
- Potentially highly impactful

- Policy trial emulation allows careful thought of the comparisons being made, and care regarding pre and post time periods, confounding, etc.
    - Transparent comparisons and diagnostics

- Could combine the non-parametric approach shown here with a parametric model to model impacts over time

- Recommend avoiding models that simply fit regressions to the longitudinal data, with fixed effects for state and time
    - No clear “design,” unclear comparisons and diagnostics, potential bias
    - “Design” the policy evaluation by thinking about the target trial that you would implement if possible

And…

We didn’t even consider various complications!

Differences in testing across time and space…

Differences in implementation and compliance….

Lots of other policies happening…

No formal links to models of infectious disease dynamics…

Timing of implementation challenging to determine…

How do we balance these challenges with the need to generate answers to important policy questions?

- “Be clear about what is knowable” – Goodman-Bacon and Marcus (2020)
- Acknowledge the challenges
- Conduct diagnostics and sensitivity analyses
- Collaborate across fields
- Build a body of evidence: don’t just rely on one study

Acknowledgements

Noah Haber (Stanford): @NoahHaber [esp. for many of the graphics and slides]

Avi Feller (Berkeley): @AviFeller

Eli Ben-Michael (Berkeley): @EliBenMichael

Funding:
Bloomberg American Health Initiative
National Institutes of Health funded RAND Opioid Policy Tools and Information Center

BTW!!! Post-doc opening on these topics!
https://www.jhsph.edu/research/centers-and-institutes/center-for-mental-health-and-addiction-policy/training-opportunities/index.html

To learn more…

Ben-Michael, Feller, and Stuart (forthcoming??): A trial emulation approach for policy evaluations with group-level longitudinal data.

French and Stuart (2020). Study designs and statistical methods for studies of child and adolescent health policies. *JAMA Pediatrics.*

Goodman-Bacon and Marcus (2020): Using difference-in-differences to identify causal effects of COVID-19 policies. https://cdn.vanderbilt.edu/vu-my/wp-content/uploads/sites/2318/2020/05/11154933/Covid-DD_v2.pdf

Haber, Clarke-Deelder, Salomon, Feller, and Stuart (2020): Policy evaluation in COVID-10: A graphical guide to common design issues. https://arxiv.org/abs/2009.01940

https://diff.healthpolicydatascience.org/

https://coronavirus.jhu.edu/from-our-experts/evaluating-the-effectiveness-of-covid-19-policies-a-q-and-a-with-dr-elizabeth-stuart

https://ncrc.jhsph.edu/non-pharmaceutical-interventions/

https://github.com/ebenmichael/policy-trial-emulation