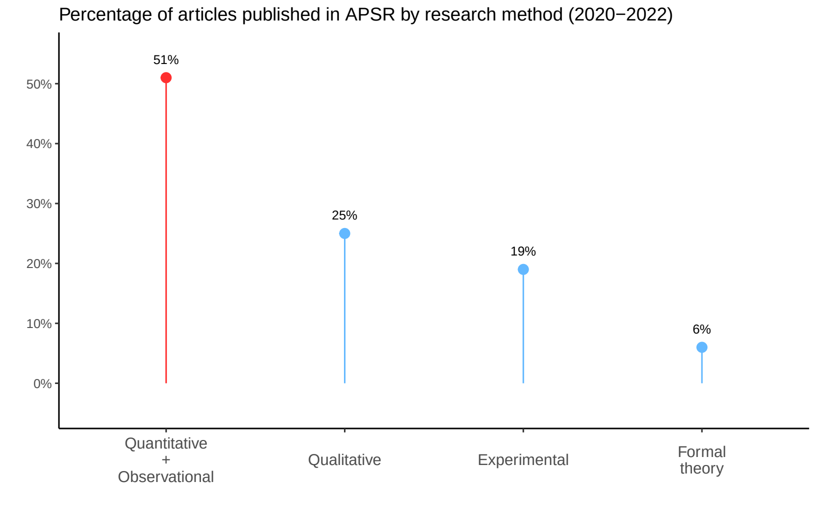
Pre-registration for Observational Analyses

Trevor Incerti
University of Amsterdam

Research Transparency and Reproducibility Training (RT2)
Berkeley Initiative for Transparency in the Social Sciences

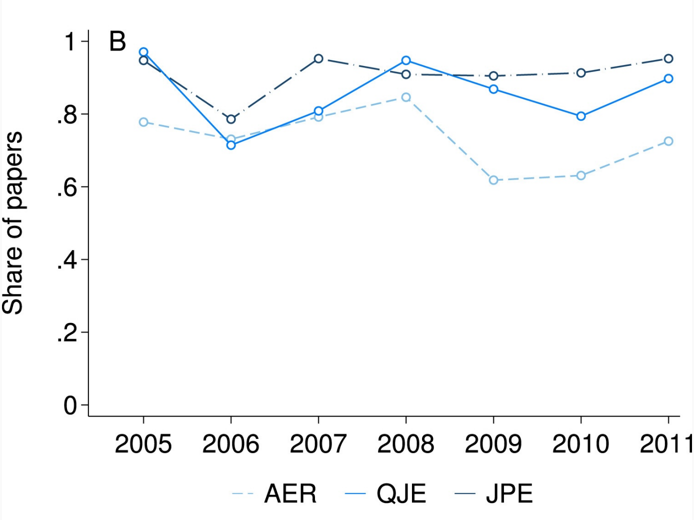
5 June 2024

Majority of quantitative studies are still observational



Source: APSR Editorial Report 2022

Majority of quantitative studies are still observational



Source: Burlig (2018)

Journals and (observational) pre-registration

	Required for RCTs	Required for Observational Studies
American Economic Review	✓	X
Econometrica	X	X
Journal of Political Economy	X	X
Quarterly Journal of Economics	X	X
Review of Economic Studies	X	X
American Journal of Political Science	X	X
American Political Science Review	X	X
Journal of Politics	✓	X

What about the literature?

- General agreement RCTs should be pre-registered whenever possible.

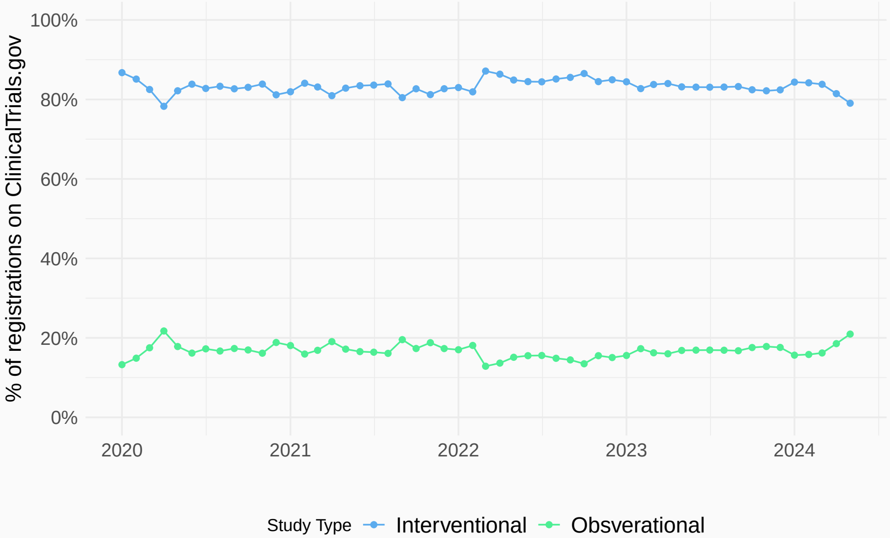
What about the literature?

- General agreement RCTs should be pre-registered whenever possible.
- Not the case with observational studies.

What about the literature?

- General agreement RCTs should be pre-registered whenever possible.
- Not the case with observational studies.
- More movement in medicine (Dal-Ré et al., 2014; Loder et al., 2010; Lancet, 2010)
  - E.g., ClinicalTrials.gov observational pre-registration

Observational pre-registration in medicine



Source: ClinicalTrials.gov

Why do it if it’s not required?

- Would likely still impress peer-reviewers.
- In the case of a registered report, could actually protect you against null result publication bias.
- Because we care about scientific best-practices.

An optimistic take

- "The goal of an observational study protocol is **not to protect against dishonest investigators** but to **aid honest investigators** to do good science." (Small, 2024)

Why is pre-registration scientific best practice?

- Forces you to think through theoretical process in advance.
- Distinguish between hypothesis-driven and exploratory analyses.
- Enhancing meta-analyses by making research traceable.

Why is pre-registration scientific best practice?

- Forces you to think through theoretical process in advance.
- Distinguish between hypothesis-driven and exploratory analyses.
- Enhancing meta-analyses by making research traceable.
- Reduce researcher degrees of freedom and risk of, e.g.,
  - p hacking
  - HARKing

Why is pre-registration scientific best practice?

- Forces you to think through theoretical process in advance.
- Distinguish between hypothesis-driven and exploratory analyses.
- Enhancing meta-analyses by making research traceable.
- Reduce researcher degrees of freedom and risk of, e.g.,
  - p hacking
  - HARKing (Hypothesizing After the Results are Known)

What observational research designs should we pre-register?

- Not necessary to pre-register descriptive or exploratory work.

What observational research designs should we pre-register?

- Not necessary to pre-register descriptive or exploratory work.
- Stronger argument for research with causal claims.
    - “We give **highest priority to studies that provide strong support for inferences** applicable to clinical practice.” (Loder et al., 2010)

What observational research designs should we pre-register?

- Not necessary to pre-register descriptive or exploratory work.
- Stronger argument for research with causal claims.
  - “We give **highest priority to studies that provide strong support for inferences** applicable to clinical practice.” (Loder et al., 2010)
- Pre-registration allows us to distinguish between the two.

Problems and tensions

- We know the data generating process in an RCT.
- In an observational study, we often need to explore our data in order to understand the DGP.
  - Where does exploratory work end and analysis begin?
  - Easier to to pre-register studies closer to experimental ideal (e.g., natural experiments, RDD), *but* we are less concerned about these designs.
- Easy to show registration occurred before analysis in an RCT.

When is pre-registration most credible?

When you have not yet collected (all) outcome data.

- "Observational studies should be designed using only background information...this activity should be conducted without any access to any outcome data." (Rubin, 2007)
- But *any* pre-registration makes your study more credible than none, so let's not get carried away.

When is pre-registration most credible?

When you have not yet collected (all) outcome data.

When is pre-registration most credible?

When you have not yet collected (all) outcome data.

- When events have not yet occurred.
  - DiD where policy change not yet implemented.
  - RDD where threshold not yet implemented.

When is pre-registration most credible?

When you have not yet collected (all) outcome data.

- When events have not yet occurred.
  - DiD where policy change not yet implemented.
  - RDD where threshold not yet implemented.
- When (all) data is not yet released.
  - Public data that is released on a rolling basis.
  - Confidential or third party controlled data

When is pre-registration most credible?

When you have not yet collected (all) outcome data.

- When events have not yet occurred.
  - DiD where policy change not yet implemented.
  - RDD where threshold not yet implemented.
- When (all) data is not yet released.
  - Public data that is released on a rolling basis.
  - Confidential or third party controlled data
- Middle ground: when you are collecting original data.
  - Exploratory analysis on baseline data → pre-register identification strategy and analysis on endline data.

Example: confidential data request

- “I would like data on your customers’ income levels in order to estimate the effect of the tax rebate program on demand for electric vehicles using a regression discontinuity design.”

Example: confidential data request

- "I would like data on your customers' income levels in order to estimate the effect of the tax rebate program on demand for electric vehicles using a regression discontinuity design."

- In the above sentence, we identified our IV, DV, and estimation strategy, before seeing the proprietary data.

Example: confidential data request

- “I would like data on your customers’ income levels in order to estimate the effect of the tax rebate program on demand for electric vehicles using a regression discontinuity design.”

- In the above sentence, we identified our IV, DV, and estimation strategy, before seeing the proprietary data.

- Now tell the world before they (hopefully) give you the data and it’s a credible PAP!

Example: confidential data request

- "I would like data on your customers' income levels in order to estimate the effect of the tax rebate program on demand for electric vehicles using a regression discontinuity design."
- In the above sentence, we identified our IV, DV, and estimation strategy, before seeing the proprietary data.
- Now tell the world before they (hopefully) give you the data and it's a credible PAP!
- Confidential data makes pre-registration easier... **but** also makes replication harder.

Example: confidential data request

- "I would like data on your customers' income levels in order to estimate the effect of the tax rebate program on demand for electric vehicles using a regression discontinuity design."
- In the above sentence, we identified our IV, DV, and estimation strategy, before seeing the proprietary data.
- Now tell the world before they (hopefully) give you the data and it's a credible PAP!
- Confidential data makes pre-registration easier... **but** also makes replication harder.

We'll return to this.

What to pre-register

Largely the same things you would in an RCT:

What to pre-register

Largely the same things you would in an RCT:

- Sample.
- Hypotheses.
- Outcome(s).
- Estimator(s).
- Primary estimand(s).
- Covariates.
- Subgroups

What to pre-register

Largely the same things you would in an RCT:

- Sample.
- Hypotheses.
- Outcome(s).
- Estimator(s).
- Primary estimand(s).
- Covariates.
- Subgroups
- Try to plan for difficulties and propose solutions in advance.
  - This may be more difficult than an RCT where noncompliance, attrition, weighting, etc. are foreseeable.

What to pre-register

Small (2024) calls for registration of:

- The study population.
- The treatment + which subjects will be considered treated.
- Primary and secondary outcomes
- The time period of measurement and analysis.
- Covariates that will be adjusted for.
- Statistical methods for adjustment and analysis.
- Robustness and sensitivity analyses.

What are we concerned about?

Some examples:

- Always: choice of subgroup(s).
- Always: choice of covariates for adjustment.

What are we concerned about?

Some examples:

- Always: choice of subgroup(s).
- Always: choice of covariates for adjustment.

- RDD: choice of bandwidth.
- Interrupted time series: choice of estimator and event window.
- DiD: choice of estimator.
- Matching: choice of matching algorithm and estimator.
- IV: selection of instrument.

What are we concerned about?

Some examples:

- Always: choice of subgroup(s).
- Always: choice of covariates for adjustment.

- RDD: choice of bandwidth.
- Interrupted time series: choice of estimator and event window.
- DiD: choice of estimator.
- Matching: choice of matching algorithm and estimator.
- IV: selection of instrument.

Are robustness checks enough? Role of replications?

Where to pre-register

- For social science: OSF (Open Science Framework)
- Medicine and public health: clinicaltrials.gov

Pre-registering with OSF

Lets walk through how to pre-register an observational study on OSF.

▸ Link

Example: pre-registering an RDD

Cattaneo and Titiunik (2024) propose a framework for pre-registering RDDs:

- Each of the features proposed by Small (2024). Plus:
- The score that all units receive
- The cutoff value of that score
- The treatment
- The rule that determines treatment status (for units above and below the cutoff)

Example: pre-registering an RDD

- Study population = all units that receive a score.
- The treatment = intervention given to units with scores above (below) the cutoff.
- Outcomes = Few based on scientific theories or many + multiple hypothesis testing approach.
- Methods = Continuity or local randomization. If both, which is primary. Assuming continuity, details of the estimator:
  - Bandwidth selection method (e.g., minimization of MSE)
  - Polynomial order
  - Kernel function
  - Uncertainty estimates.
  - Misspecification contingencies

Example: pre-registering an RDD

- Covariates and how they will be used.
  - Similarity on pre-treatment covariates above and below cutoff.
  - What will be done in the case of imbalance?
  - Which will be included for efficiency gains?
  - Which will be conditioned on in subgroup analyses?

Example: pre-registering an RDD

- Covariates and how they will be used.
  - Similarity on pre-treatment covariates above and below cutoff.
  - What will be done in the case of imbalance?
  - Which will be included for efficiency gains?
  - Which will be conditioned on in subgroup analyses?
- Robustness and sensitivity analyses
  - e.g., local randomization if continuity primary estimator.
  - Alternative bandwidths to be tested and selection mechanisms.

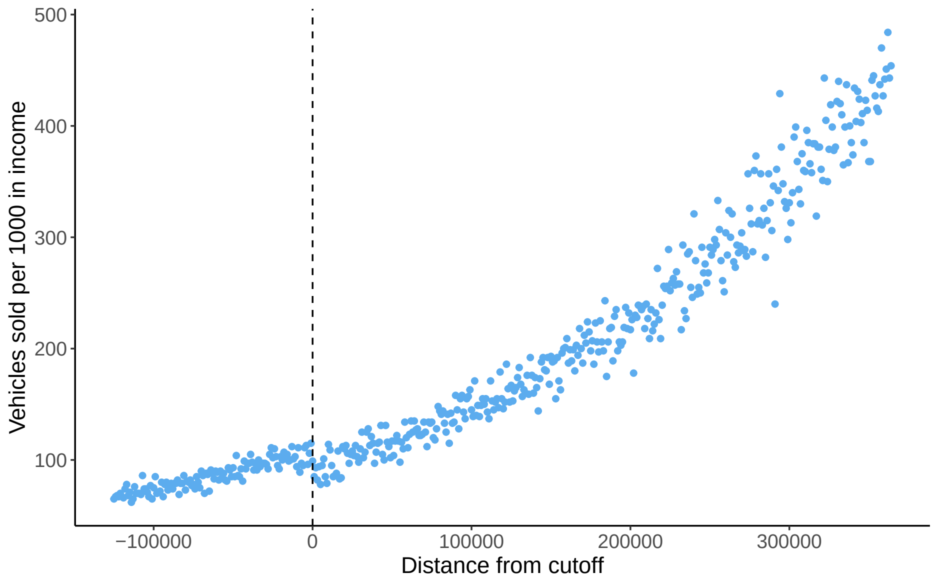
Example: pre-registering an RDD

- Ideally, simulate your data (if you have pre-intervention data, even better)
- Write code for your estimation strategy and run it on the simulated data.
- Test sensitivity to different treatment effect sizes.

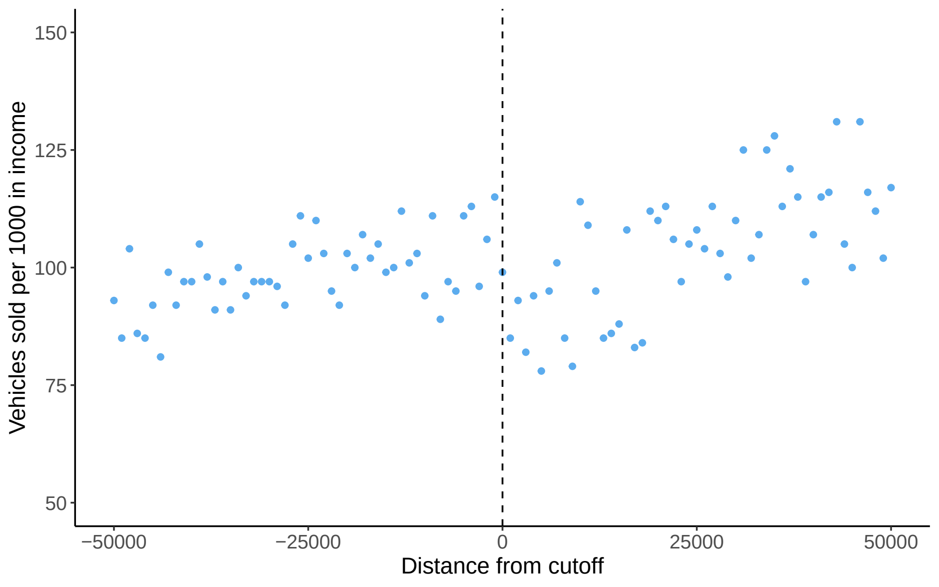
Example: CA income threshold for EV rebates

- California has a \$135,000 income limit for receiving tax rebates for sales of electric vehicles.
- What if we wanted to know if this income limit is pushing high income earners away from purchasing EVs?
- I don't have this data, but maybe we can guess at what it might look like *before* getting access to private purchase data.

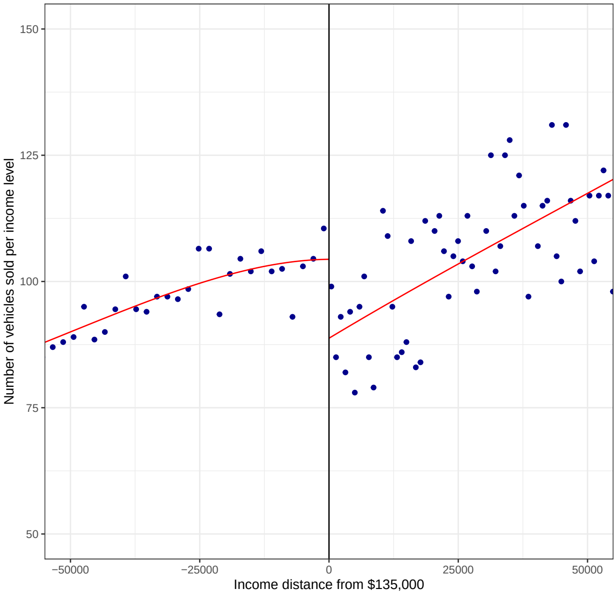
Example: CA income threshold for EV rebates



Example: CA income threshold for EV rebates



Example: CA income threshold for EV rebates

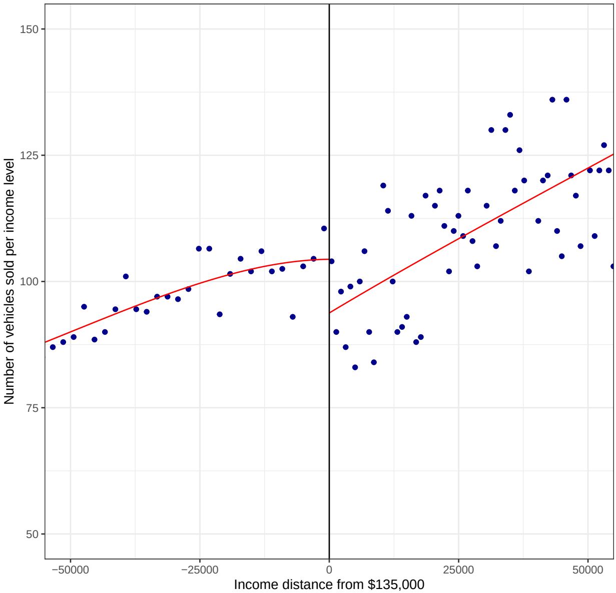


Example: CA income threshold for EV rebates

Table 1: Simulated estimates from RD Robust

Estimator	Estimate	SE	Kernel	Bandwidth
Conventional	-17.72	3.82	Triangular	MSE
Bias-Corrected	-17.79	3.82	Triangular	MSE
Robust	-17.79	4.59	Triangular	MSE

Example: CA income threshold for EV rebates

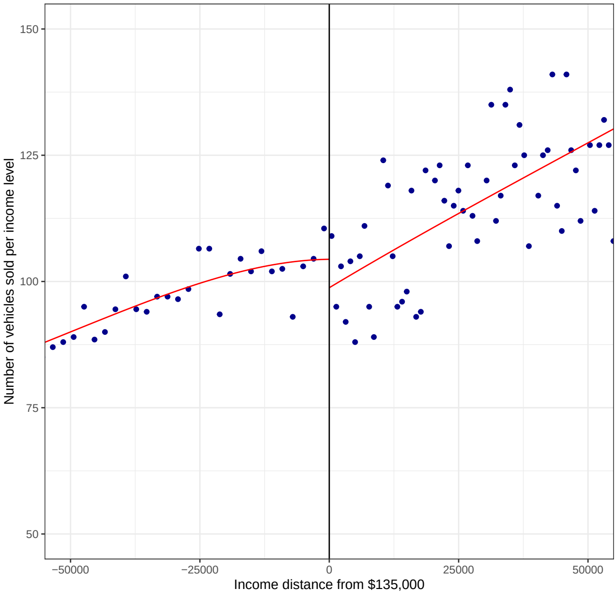


Example: CA income threshold for EV rebates

Table 2: Simulated estimates from RD Robust

	Estimator	Estimate	SE	Kernel	Bandwidth
1	Conventional	-12.719	3.822	Triangular	MSE
2	Bias-Corrected	-12.794	3.822	Triangular	MSE
3	Robust	-12.794	4.594	Triangular	MSE

Example: CA income threshold for EV rebates



Example: CA income threshold for EV rebates

Table 3: Simulated estimates from RD Robust

	Estimator	Estimate	SE	Kernel	Bandwidth
1	Conventional	-7.719	3.822	Triangular	MSE
2	Bias-Corrected	-7.794	3.822	Triangular	MSE
3	Robust	-7.794	4.594	Triangular	MSE

Takeaways

- Observational pre-registration still not the norm or expected.
- But, will likely impress reviewers.
- Good practice for studies with causal claims.

Takeaways

- Best practices from RCT pre-registration largely follow.
    - Most credible before outcome data collected.
    - Should pre-register similar items as RCT protocols.
    - Forces researchers to pre-define theory, hypotheses, and estimation.
    - Helps clarify causal vs. exploratory vs descriptive.
    - Simulate study data and perform power analyses.

References

Fiona Burlig. Improving transparency in observational social science research: A pre-analysis plan approach. *Economics Letters*, 168:56–60, 2018.

Matias D Cattaneo and Rocio Titiunik. Protocols for observational studies: An application to regression discontinuity designs. *arXiv preprint arXiv:2402.11640*, 2024.

Rafael Dal-Ré, John P Ioannidis, Michael B Bracken, Patricia A Buffler, An-Wen Chan, Eduardo L Franco, Carlo La Vecchia, and Elisabete Weiderpass. Making prospective registration of observational research a reality. *Science translational medicine*, 6(224):224cm1–224cm1, 2014.

The Lancet. Should protocols for observational research be registered? *The Lancet*, 375(9712):348, 2010.

Elizabeth Loder, Trish Groves, and Domhnall MacAuley. Registration of observational studies. *BMJ*, 340, 2010.

Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36, 2007.

Dylan S Small. Protocols for observational studies: Methods and open problems. *arXiv preprint arXiv:2403.19807*, 2024.