CauSciBench: Assessing LLM Causal Reasoning for Scientific Research

Sawal Acharya[*1*] **Terry Jingchen Zhang**[*2*] **Pepijn Cobben**[2] **Andrew Kim**[3,4]
Furkan Danisman[3,4] **Yahang Qi**[3,4] **Anahita Haghighat**[5] **Xianlin Sun**[6]
Rahul Babu Shrestha[7] **Clijo Jose**[9] **Maximilian Mordig**[2,8]
Mrinmaya Sachan[2] **Bernhard Schölkopf**[8] **Zhijing Jin**[3,4,8]

[1]Stanford University [2]ETH Zürich [3]University of Toronto [4]Vector Institute [5]Independent
[6]HKU [7]TUM [8]MPI for Intelligent Systems, Tübingen, Germany [9]IP Paris

sawal386@stanford.edu zjingchen@ethz.ch zjin@cs.toronto.edu

Abstract

While large language models (LLMs) are increasingly integrated into scientific research, their capability to perform causal inference remains under-evaluated. Existing benchmarks either focus on method execution or provide open-ended tasks lacking precision in defining causal estimands, methodological choices, and variable selection. We introduce CauSciBench, a comprehensive benchmark for assessing the causal inference capabilities of LLMs, combining expert-curated problems from published research with diverse synthetic scenarios and textbook examples. Our benchmark is the first to enable assessment across the complete causal analysis pipeline, from problem formulation through variable selection and method choice to statistical model implementation and result interpretation. We evaluate the benchmark across 3 language models and 2 prompting strategies. On real datasets, the best-performing model, OpenAI-o3, attains a mean relative error (MRE) of 53.0%. Meanwhile, for synthetic and textbook datasets, the best-performing model yields MREs of 6.2% and 30.6%, respectively. This substantial performance gap underscores both the difficulty of real-world causal inference and the opportunity for advancing LLM capabilities on this front.[2]

1 Introduction

Causal inference is fundamental to scientific discovery, enabling researchers to establish cause-and-effect relationships across social science [11], public health [5], and biomedicine [15]. The integration of large language models (LLMs) into scientific workflows creates opportunities to democratize sophisticated causal analysis. Recent LLM-powered agents show promise for automating causal inference procedures [7, 27], potentially accelerating research across disciplines [14].

Evaluating LLM causal inference capabilities presents unique challenges. Causal inference deals with unobservable counterfactual outcomes [10], requiring sophisticated methodological frameworks and identification strategies. Current approaches typically assume users can appropriately select methods and specify model-specific variables [16, 3]. Whether LLMs demonstrate genuine causal reasoning or merely employ sophisticated pattern matching remains an open question [14, 29].

Existing benchmarks address different aspects of causal reasoning but leave gaps. Text-based approaches evaluate commonsense causal understanding [23, 20, 13, 4] or formal reasoning within

[*]Equal Contribution

[2]Code and datasets are available at: https://github.com/causalNLP/CauSciBench

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

Benchmark	End-to-End Causal Analysis	Intermediate Evaluation	Data + Context Understanding	Sources	Answer Format	# Queries
RealCause [19]	✗	✗	✗	3 Datasets + Semi-synthetic Scenarios	Point Estimate	1569[1]
QRData [16]	✗	✗[2]	✓	5 Datasets + 3 Textbooks	Freeform QA	411
DiscoveryBench [18]	✗	✗	✓	26 Datasets + Synthetic Scenarios	Freeform QA	239
BLADE [6]	✗	✓	✓	12 Datasets	Code + Freeform QA	12
CauSciBench	✓	✓	✓	52 Datasets + 2 Textbooks + Synthetic Scenarios	Point Estimate + Causal Components	**305**

Table 1: Comparison of CauSciBench with related benchmark datasets for causal inference. ✓ = Yes, ✗ = No. **End-to-End Causal Analysis** indicates whether the benchmark evaluates the full pipeline of causal inference; **Intermediate Evaluation** captures whether the benchmark supports evaluation of intermediate steps; **Data + Context Understanding** assesses whether the benchmark requires models to interpret the relationship between the data variables and the background information.

Pearl's SCM framework [12, 4]. Implementation-focused benchmarks like QRData [16] assess method execution on tabular data but not problem formulation from natural language descriptions. General data analysis benchmarks such as BLADE [6] and DiscoveryBench [18] provide open-ended tasks without causal inference specificity.

CauSciBench addresses these gaps by enabling systematic evaluation across the complete analysis pipeline. Our benchmark provides fine-grained assessment from problem formulation and variable selection to method choice, causal effect estimation, and interpretation. To this end, we make three key contributions:

1. End-to-End Task Reflecting Research Demand. CauSciBench is the first benchmark that requires models to perform the complete pipeline of causal inference: choosing treatments, outcomes, and confounders; selecting appropriate identification strategies and estimation methods; implementing these approaches; and finally interpreting results in the context of the research question.

2. Hybrid Design with Real-Synthetic Comparison. We combine three complementary data sources spanning real-world research problems, synthetic scenarios with user-defined ground truth, and textbook examples to balance question validity with highly diverse problem sets. This design enables diagnosis of whether failures arise from implementation errors or from difficulties in handling the complexity of research problem descriptions.

3. Vulnerability-Aware Automated Evaluation Pipeline. To identify key vulnerabilities in our evaluation pipeline, we implement a fully automated evaluation system capable of pinpointing weaknesses in the causal inference process, which typically stem from problematic method selection or implementation. We evaluate a wide range of frontier models and show that further effort is needed to reliably integrate LLM-powered agents into research-level causal inference pipelines.

2 Problem Formulation

Our goal is to assess LLMs' ability to generate answers to causal queries through sound causal analysis involving: (i) framing the causal estimation problem by selecting appropriate treatment and outcome variables and the correct estimand, (ii) assessing whether the estimand can be identified

[1] Cartesian product of 3 datasets, 4 estimators, 15 ML models, and 10 different hyperparameter settings
[2] Partial

from the provided dataset, (iii) formulating and implementing the correct statistical model, and (iv) extracting and interpreting the causal effect.

Each benchmark instance consists of five core components: **Data** (experimental or observational input), **Dataset Description** (information about data collection, variable definitions, and background context), **Query** (causal question involving the effect of one variable on another), **Causal Inference Method and Effect Estimate** (expert-validated method and corresponding effect estimate providing ground truth), and **Model Variables** (key variables including treatment, outcome, confounders, and method-specific variables).

3 Dataset Collection: CauSciBench

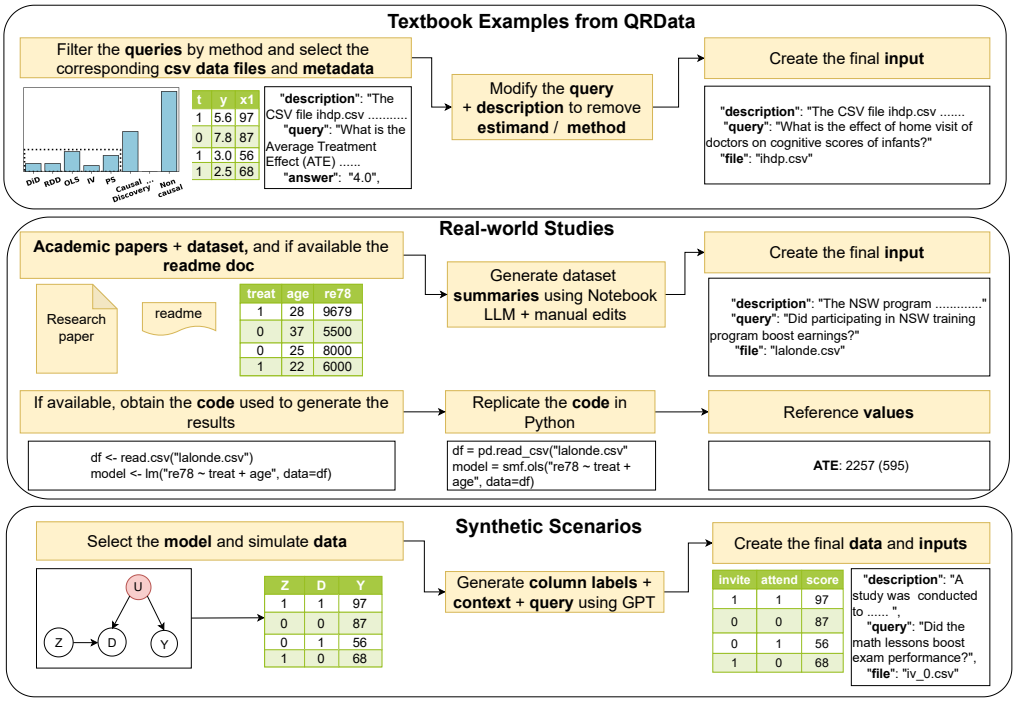


Figure 1: Dataset creation process for **Textbook Examples, Real-World Studies, and Synthetic Data**

Figure 1 details our comprehensive dataset creation process across three sources: real-world studies, synthetic scenarios, and textbook examples. We describe the steps more concretely below.

Source 1: Research Paper Curation We compile papers from various disciplines including economics, criminology, and political science (see Figure 2). For each paper, we create a comprehensive summary capturing key dataset information including variable descriptions, data collection procedures, and research purpose. We formulate causal queries by systematically examining empirical methodology and conclusions from causal effects, selecting methods authors cite to justify their findings and choosing the most expressive model specifications for completeness. This curation process ensures methodological rigor reflecting real-world research standards while providing authentic complexity requiring navigation of confounders, identification validity assessment, and result interpretation.

Source 2: Automated Synthesis We automatically synthesize datasets by randomly selecting true causal effects τ in the range $(1, 10)$ with continuous covariates drawn from normal distributions and binary covariates from binomial distributions. As an example, for randomized trials, the outcome Y in terms of the covariates $\mathbf{X}$ and treatment variable T is $Y = \alpha + \mathbf{X}\vec{\theta} + \tau T + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$

is the error term, $\vec{\theta}$ is the coefficient vector chosen from a normal distribution, and α is the randomly chosen intercept term. Likewise, we use GPT-4o to synthesize diverse contexts for each synthetic dataset, creating plausible scenarios explaining data collection with comprehensive dataset metadata including headings and descriptions. This approach improves dataset diversity while testing model performance consistency in scenarios mirroring real-world research contexts.

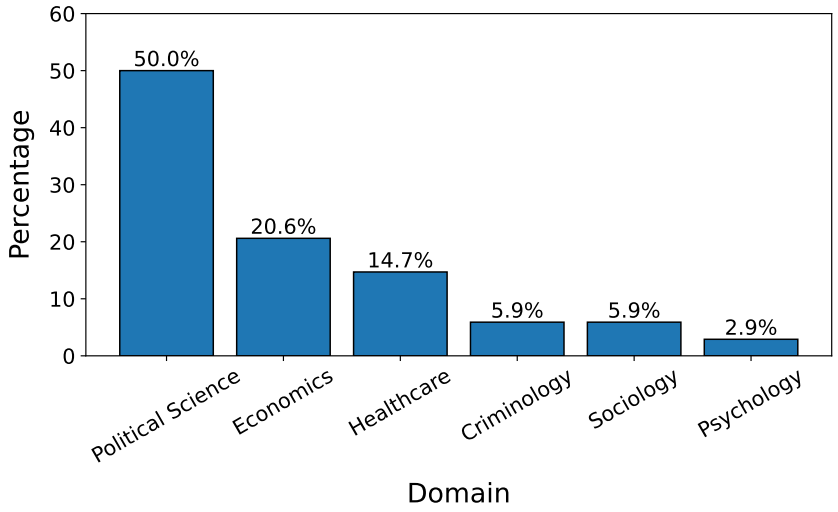


Figure 2: Distribution of domains in real-world publications

Source 3: Refined QRData Since QRData tasks specify inference methods or estimands and our focus is end-to-end causal inference including automatic method and variable selection, we systematically modify queries by removing explicit references to estimation techniques. For example, "What is the Average Treatment Effect (ATE) of the dataset?" becomes "What is the effect of home visits by doctors on cognitive scores of infants?" We retain original dataset descriptions and numerical causal effect estimates, restricting evaluation to queries with numerical answers to enable precise quantitative assessment.

4 Experimental Setup

Prompting Strategies We investigate two prompting strategies: (i) **Direct prompting** [2] provides comprehensive dataset information and the causal question of interest. The LLM must then select a causal inference method and produce executable Python code. (ii) **Chain of Thought** (CoT) [28] maintains the same input but explicitly breaks down the workflow: variable selection $\rightarrow$ inference method selection $\rightarrow$ statistical estimation model $\rightarrow$ implementation. At each step, the model must justify its decision choices.

Metrics We evaluate all models using the following two metrics: (1) **Method Selection Accuracy (MSA)**: Percentage of queries where the selected method $\hat{m}_i$ matches the reference method m_i: $\text{MSA} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{m}_i = m_i] \times 100\%$. (2) **Mean Relative Error (MRE)**: Average relative error between predicted causal effects $\hat{\tau}_i$ and reference values τ_i: $\text{MRE} = \frac{1}{N}\sum_{i=1}^{N} \min\left(\frac{|\hat{\tau}_i - \tau_i|}{|\tau_i|}, 1\right) \times 100\%$. To reduce the impact of outliers, relative error is capped at 100% per query. In the above formulas, N denotes the total number of queries in the evaluation set.

Python Libraries For causal effect estimation, we use the `DoWhy` [25, 1], `linearmodels` [26], `rdd` [17], and `statsmodels` [24] libraries. Similarly, for pre-processing and intermediate computations, we use `numpy` [8], `pandas` [21], and `scikit-learn` [22].

Dataset	Prompt	Method Accuracy (↑)			Mean Rel. Error (↓)		
		4o-mini	4.1	o3	4o-mini	4.1	o3
Real	Basic	34.57	47.78	**71.76**	71.45	58.43	53.82
	CoT	40.23	55.56	67.74	62.62	53.59	**53.02**
Synthetic	Basic	15.38	59.43	72.41	22.58	**6.16**	6.30
	CoT	24.56	**77.14**	69.23	17.25	10.99	17.24
Textbook	Basic	60.00	64.10	69.23	42.03	40.05	46.41
	CoT	53.85	**71.79**	66.67	41.29	33.68	**30.59**

Table 2: Performance comparison across datasets and prompting methods.

5 Results and Discussion

Table 2 shows the method-selection accuracy and relative errors (MRE) of causal effect estimates under pass@1.

Causal estimation from real data is challenging. Performance for real datasets in terms of causal effect estimation is lower than that for synthetic and textbook datasets. While synthetic datasets benefit from controlled generation and textbook datasets from extensive preprocessing for pedagogy, real-world data presents greater complexity through more variables, higher noise levels, and a lack of preprocessing. These factors complicate both method and variable selection, with methodological errors cascading through the causal inference pipeline to amplify estimation errors.

Wrong methods directly amplify estimation errors. Table 3 in the appendix shows that incorrect method selection is a major drive of causal inference failures, yielding substantially higher MRE across nearly all settings. This effect intensifies with dataset complexity, particularly for real-world data. The textbook-based dataset is an exception. This is because most misclassifications involve choosing propensity score methods over regression/difference-in-means in the IHDP dataset [9], a randomized experiment where both methods yield similar results.

Implementation failures persist despite correct method choice. Even with appropriate method identification, substantial errors remain due to execution failures, as evidenced by persistently high relative errors in Table 2. A major reason for this is the incorrect selection of model variables. While the treatment and outcome variables are correctly selected, LLM models tend to over-select the **control variables** in the estimation model. The likelihood of this is particularly high when the raw datasets have a large number of variables, where many of them are redundant in terms of estimation.

Models systematically default to OLS. The confusion matrices in Figure 3 reveal that LLMs exhibit a pronounced bias toward Ordinary Least Squares (OLS) across all causal inference scenarios, regardless of the appropriate method. This tendency is particularly pronounced for smaller models, such as GPT-4o-mini. The overwhelming selection of OLS stems from several factors. OLS is simpler and easier to implement. Likewise, for most empirical papers, OLS is the baseline model. However, this bias is highly problematic for causal inference. Naive OLS often fails to address the effect of unobserved confounders and the estimates have low precision.

6 Conclusion

CauSciBench establishes a comprehensive framework for evaluating causal inference capabilities in large language models, revealing critical limitations that must be addressed before these systems can reliably support scientific research. Current LLMs show systematic biases toward overly simple methods, with a concerning tendency to default to OLS estimation even when it is not appropriate. Additionally, they struggle with correct implementation even when they select the right method. The substantial performance gap between synthetic and real-world scenarios suggests that improving LLM causal reasoning requires two key developments: more robust frameworks for handling the complexity of real-world datasets, and better mechanisms to bridge the gap between theoretical understanding and practical implementation.

References

[1] Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024. URL http://jmlr.org/papers/v25/22-1258.html.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

[3] Qiang Chen, Tianyang Han, Jin Li, Ye Luo, Yuxiao Wu, Xiaowei Zhang, and Tuo Zhou. Can ai master econometrics? evidence from econometrics ai agent on expert-level tasks, 2025. URL https://arxiv.org/abs/2506.00856.

[4] Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. Causal evaluation of language models, 2024. URL https://arxiv.org/abs/2405.00622.

[5] Thomas A. Glass, Steven N. Goodman, Miguel A. Hernán, and Jonathan M. Samet. Causal inference in public health. *Annual review of public health*, 34:61–75, March 2013. ISSN 0163-7525. doi: 10.1146/annurev-publhealth-031811-124606.

[6] Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. Blade: Benchmarking language model agents for data-driven science, 2024. URL https://arxiv.org/abs/2408.09667.

[7] Kairong Han, Kun Kuang, Ziyu Zhao, Junjian Ye, and Fei Wu. Causal agent based on large language model, 2024. URL https://arxiv.org/abs/2408.06849.

[8] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[9] Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162. URL https://doi.org/10.1198/jcgs.2010.08162.

[10] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. doi: 10.1080/01621459.1986.10478354. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354.

[11] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[12] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=e2wtjx0Yqu.

[13] Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vqIH0ObdqL.

[14] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=mqoxLkX210. Featured Certification.

[15] Samantha Kleinberg and George Hripcsak. Methodological review: A review of causal inference for biomedical informatics. *J. of Biomedical Informatics*, 44(6):1102–1112, December 2011. ISSN 1532-0464. doi: 10.1016/j.jbi.2011.07.001. URL https://doi.org/10.1016/j.jbi.2011.07.001.

[16] Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. Are LLMs capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9215–9235, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.548. URL https://aclanthology.org/2024.findings-acl.548.

[17] Evan Magnusson. rdd. https://pypi.org/project/rdd/, 2019. Version 0.0.3, MIT License.

[18] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models, 2024. URL https://arxiv.org/abs/2407.01725.

[19] Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking, 2021. URL https://arxiv.org/abs/2011.15007.

[20] Allen Nie, Yuhui Zhang, Atharva Amdekar, Chris Piech, Tatsunori Hashimoto, and Tobias Gerstenberg. Moca: measuring human-language model alignment on causal and moral judgment tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[21] The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. CRAB: Assessing the strength of causal relationships between real-world events. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.940. URL https://aclanthology.org/2023.emnlp-main.940/.

[24] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[25] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[26] Kevin Sheppard, Joon Ro, Snyk bot, Brian Lewis, Christian Clauss, Guangyi, Jeff, Jerry Qinghui Yu, Jiageng, Kevin Wilson, LGTM Migrator, Thrasibule, William Roy Nelson, Xavier RENE-CORAIL, and vikjam. linearmodels: Linear (regression) models for python. https://github.com/bashtage/linearmodels, 2024. Version 6.1, University of Illinois/NCSA Open Source License.

[27] Xinyue Wang, Kun Zhou, Wenyi Wu, Har Simrat Singh, Fang Nan, Songyao Jin, Aryan Philip, Saloni Patnaik, Hou Zhu, Shivam Singh, Parjanya Prashant, Qian Shen, and Biwei Huang. Causal-copilot: An autonomous causal analysis agent, 2025. URL https://arxiv.org/abs/2504.13263.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[29] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal, 2023. URL https://arxiv.org/abs/2308.13067.

A Acknowledgment

This material is based in part upon work supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by the Schmidt Sciences SAFE-AI Grant; by the Frontier Model Forum and AI Safety Fund; by Open Philanthropy; by the Cooperative AI Foundation; and by the Survival and Flourishing Fund. The use of OpenAI credits is largely supported by the Tübingen AI Center. Resources used in preparing this research project were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

We would also like to thank Samuel Simko, Vishal Verma, and Devansh Bhardwaj for helping us engineer the implementation of the baseline models. Finally, we thank all causal inference researchers for making their data available for research, and enabling this study.

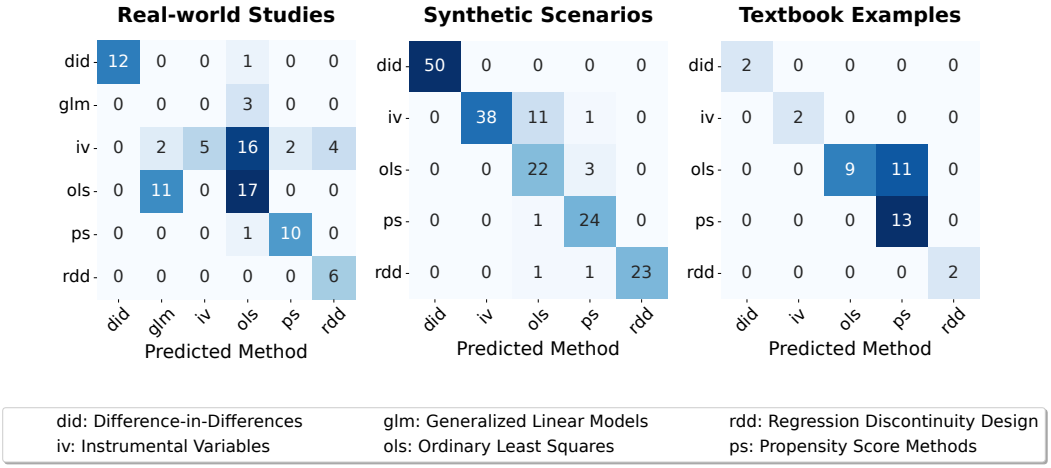
B Limitations

Our work has several limitations. The expert-curated subset requires extensive manual curation creating scalability constraints and potential annotation inconsistencies. Results are based on pass@1 evaluation to balance budgetary constraints with broad model coverage, though pass@k would strengthen findings generalizability. Our benchmark focuses primarily on potential outcomes framework with limited Pearl's structural causal model coverage. Synthetic data generation may not fully capture real-world dataset complexity including missing data patterns, measurement error, and domain-specific confounding structures. Moreover, we have not covered more recent methods, such as Double Machine Learning, which are becoming popular for high-dimensional data. Finally, emphasis on tabular data overlooks emerging applications to images and text.

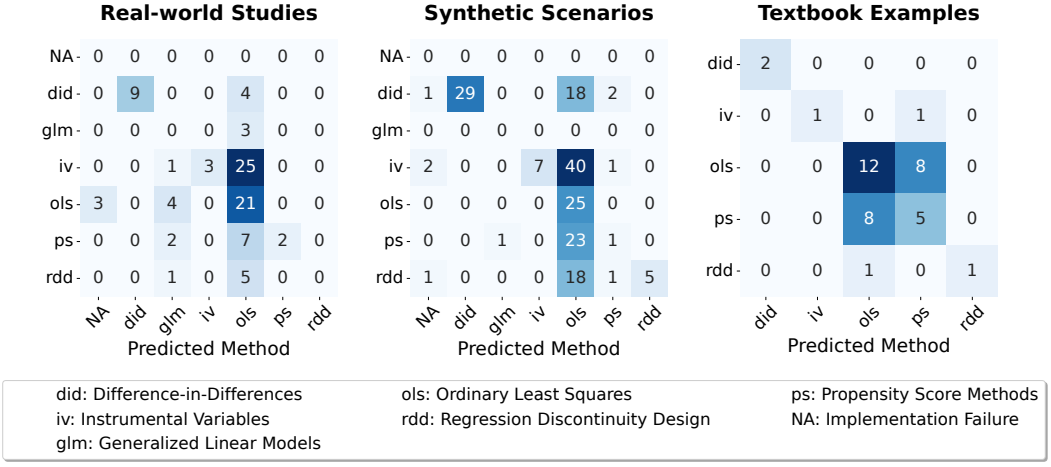
C Detailed Results and Failure Analysis

Model	**Real** (C / W)	**Synth** (C / W)	**Textbook** (C / W)
4o-mini	51.56 / 70.48	13.83 / 19.48	40.09 / 43.33
GPT-4.1	43.31 / 67.40	10.66 / 13.92	42.27 / 11.81
o3	44.51 / 71.27	14.67 / 33.75	35.35 / 15.34

Table 3: Relative error of causal effect estimation: **Correct (C)** vs. **Wrong (W)** method selection across LLMs and datasets for **CoT** prompting



(a) **GPT-4.1:** Confusion matrix for method selection across the three datasets



(b) **GPT-4o-mini:** Confusion matrix for method selection across the three datasets

Figure 3: Confusion matrix for method selection under **CoT-based baseline** with (a) GPT-4.1; (b) GPT-4o-mini.

D Dataset Curation Process

The dataset curation process of our work follows a three-stage methodology designed to ensure high-quality benchmarks through rigorous, expert-curated papers.

- **Paper Selection** focuses on finding articles from diverse fields such as healthcare and economics that utilize established estimation methods, including OLS, DiD, RDD, IV, and propensity score methods. The selection criteria emphasize reproducibility and dataset complexity, where we prioritize papers with simpler and more explicit approaches to causal estimation to work with current LLMs' preprocessing limitations. Furthermore, as we go through the replication process in future steps, we exclude papers that do not include a publicly accessible dataset with adequate data-sharing licensing.
- **Core Information Extraction** follows paper selection, focusing on extracting the core information that practitioners require for a causal analysis, including treatment variables, outcomes, and non-causal natural language queries to avoid any methodological hints. Multiple questions per paper are permitted when the controls or outcomes differ meaningfully, maximizing the scientific value while preventing analytical redundancy.
- **Quality Filtering** implements multi-layered expert inspection throughout the entire curation process. All curated datasets undergo replication verification, where experts replicate the estimation process in Python and exclude all papers that fail to reproduce the original estimates within 10% error in around 50 lines of code. This process validates that the estimates in the paper are truly replicable with the given dataset and methods, so that should the LLM fail to replicate the results, the cause lies in the LLM's approach and not the dataset or the paper's approach.

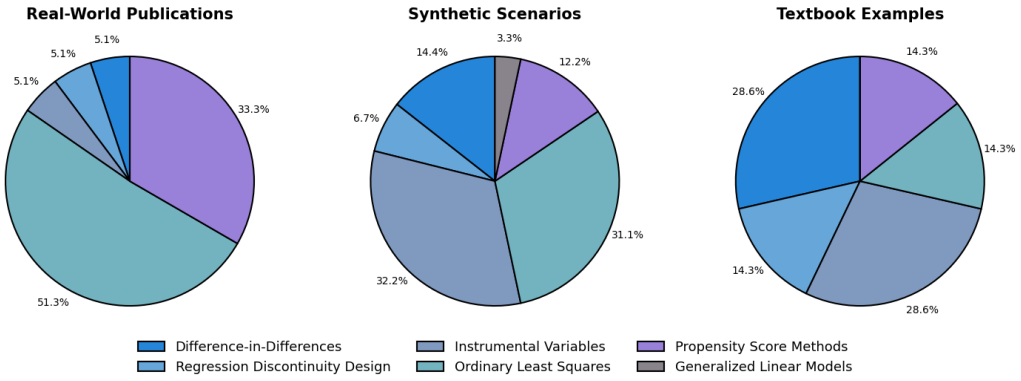


Figure 4: Distribution of estimation methods across the three dataset collections

E Sample Questions From Each Source

Real-World Publications
Source: Cities as Lobbyists [42]
Domain: Economics
Natural Language Query: How much does the money spent on lobbying increase the number of earmarks received?
Method: Instrumental Variables **Treatment:** `ln_citylobby` (log of city lobbying spending) **Instrument:** `direct_flight_dc` (1=direct flight to DC in 2007, 0=otherwise) **Outcome:** `ln_earmark` (log of total earmarks 2008-2009) **Other Variables:** `state`, `pop_e`, `land_e`, `water_e`, `senior_e`, `student_e`, `ethnic_e`, `mincome_e`, `unemp_e`, `poverty_e`, `gini_e`, `city_propertytaxshare_e`, `city_intgovrevenueshare_e`, `city_airexp_e`, `houdem_e`, `ln_countylobby` **Data:** Cities with population over 25,000, 2007-2009 panel
Synthetic Dataset
Source: Cardiovascular Rehabilitation Program Effectiveness Study
Domain: Healthcare
Natural Language Query: Does the new rehabilitation program help patients with cardiovascular diseases recover faster?
Method: Regression Discontinuity Design **Treatment:** `treatment_received` (1=new program, 0=standard care) **Running Variable:** `income_level` (threshold at 12 for eligibility) **Outcome:** `recovery_time` (days to recovery) **Other Variables:** `patient_age`, `health_index`, `smoking_status`, `obesity_status` **Data:** Regional health department evaluation study
Textbook Examples
Source: Effect of Cigarette Taxation on Consumption [16]
Domain: Healthcare, Political Science
Natural Language Query: Did Proposition 99 help reduce cigarette sales?
Method: Difference-in-Differences **Treatment:** `california` (1=CA with Prop 99, 0=other states) **Time:** `after_treatment` (1=post-1988, 0=pre-1988) **Outcome:** `cigsale` (total cigarette sales) **Other Variables:** `state`, `year`, `lnincome`, `beer`, `age15to24`, `retprice` **Data:** 39 US states, 1970-2000 panel

Table 4: Sample questions from each source pillar with the information regarding the paper that the LLM uses as context.

F Prompt Templates

In this section, we present the templates for two of the baseline prompting strategies: **Direct Prompt** and **Chain of Thought (CoT) prompt**

Direct Prompt

You are an expert in statistics and causal reasoning. You will answer a causal question on a tabular dataset.
The dataset is located at `{self.dataset_path}`.
The dataset has the following description: `{self.dataset_description}`
To help you understand it, here is the result of `df.describe()`:

`{df_info}`

Here are the columns and their types:

`{columns_and_types}`

Here are the first 5 rows of the dataset:

`{df.head()}`

If there are fewer than 10 columns, here is the result of `df.cov()`:

```
{(df.cov(numeric_only=True) if len(df.columns) < 10
else "Too many columns to compute covariance")}
```

Here is the output of `df.isnull().sum(axis=0)`: `{nan_per_column}`
The causal question I would like you to answer is: `{self.query}`
Using the descriptions and information from the dataset, write Python code to build the causal inference model based on the method and variables you have selected, and compute the causal effect to answer the query. If you need to preprocess the data, please do so in the code.
Important: Only use these approved packages: pandas, numpy, scipy, scikit-learn (sklearn), statsmodels, dowhy, rdd (for regression discontinuity design), linearmodels, econml.
Here are some example methods; you can choose one from them:

- IPW (Inverse Probability Weighting): choose the right estimand (ATE/ATT/ATC), and compute the causal effect
- Linear regression with control variables: build a regression model with the treatment, outcome, and confounders/control variables, and compute the causal effects
- Instrumental variable: build an instrumental variable model, and compute the causal effects associated with the treatment variable
- Matching: choose the correct estimand (ATE/ATT/ATC), match accordingly, and then compute the causal effects
- Difference-in-differences: build a difference-in-differences model, and output the coefficient of the variable of interest
- Regression discontinuity design: build a regression discontinuity design model, and output the coefficient of the variable of interest
- Linear regression/difference-in-means: either build a regression model consisting of the treatment and outcome variables, and compute the coefficient associated with the treatment variable or compute the difference in means across treatment and control groups
- Generalized linear models/GLM: build a GLM model, and output the coefficient of the variable of interest
- Frontdoor adjustment: build a causal graph, identify a mediator variable between the treatment and outcome, check for the frontdoor criterion, and compute the causal effect using the frontdoor adjustment formula

Make sure the code prints the final results, including:

1. The causal effect (the value only)
2. The standard deviation (the value only)
3. The causal inference method that was used to compute the effect (the method name only)
4. The treatment variable (the variable name only)
5. The outcome variable (the variable name only)

6. The mediator variable (the variable name only if frontdoor adjustment was used)
7. RCT: True/False (NA if not sure; whether the data is from a randomized controlled trial or not)
8. The covariates/control variables that were used in the causal inference model (the variable names only)
9. Instrumental variable, if the instrumental variable method was used (the variable name only)
10. Running variable, if regression discontinuity design was used (the variable name only)
11. Temporal variable, if difference-in-differences was used (the variable name only)
12. Results of statistical tests, if applicable
13. Brief explanation for model choice
14. The regression formula, if applicable.

If a variable is not applicable, print "NA" for it.
The code you output will be executed, and you will receive the output. Please make sure to output only one block of code, and make sure the code prints the result you are looking for at the end. Everything between your first code block: ```` ```python ```` and ```` ``` ```` will be executed. If there is an error, you will have several attempts to correct the code.

Chain of Thought Prompt

You are an expert in causal inference. You will use a chain-of-thought approach to answer a causal question on a tabular dataset.
The dataset is located at `{self.dataset_path}`
The dataset has the following description: `{self.dataset_description}`
Here are the columns and their types: `columns_and_types`
Here is the statistical summary of the dataset: `df.describe()`
Here are the first 5 rows of the dataset: `{df.head()}`
If there are fewer than 10 columns, here is the result of df.cov():

```
{(df.cov(numeric_only=True) if len(df.columns) < 10
else "Too many columns to compute covariance")}
```

Here is the output of df.isnull().sum(axis=0): `{nan_per_column}`
The causal question I would like you to answer is: `{self.query}`
Let us approach this problem step by step. Step 1. First, go through the dataset description and the columns and their types. Then, identify the treatment variable, the outcome variable, and the potential confounders. Explain your reasoning for choosing these variables. Remember, the dataset is located at: `{self.dataset_path}`.
Step 2. What would be the right estimand to consider for this problem? Then, choose the most appropriate method that can be used to estimate the causal effect. The available methods are:

- IPW (Inverse Probability Weighting): Choose the right estimand (ATE/ATT/ATC), and compute the causal effect
- Linear regression with control variables: Build a regression model with the treatment, outcome, and confounders/control variables, and compute the causal effects
- Instrumental variable: Build an instrumental variable model, and compute the causal effects associated with the treatment variable
- Matching: Choose the correct estimand (ATE/ATT/ATC), match accordingly, and then compute the causal effects
- Difference-in-differences: Build a difference-in-differences model, and output the coefficient of the variable of interest
- Regression discontinuity design: Build a regression discontinuity design model, and output the coefficient of the variable of interest
- Linear regression/difference-in-means: Either build a regression model consisting of the treatment and outcome variables, and compute the coefficient associated with the treatment variable or compute the difference in means across treatment and control groups
- Generalized linear models/GLM: Build a GLM model, and output the coefficient of the variable of interest

- Frontdoor adjustment: Build a causal graph, identify a mediator variable between the treatment and outcome, check for the frontdoor criterion, and compute the causal effect using the frontdoor adjustment formula

Explain why you chose the selected method, and how the data and its description support your choice. This means you should explain why the identification assumptions of the method are satisfied.
Step 3. Next, we will plan the implementation. Before writing the code, describe your implementation process. This includes:

1. Describing the necessary preprocessing steps.
2. How will we select the variables to use in the model?

Step 4. Finally, reflecting on the previous steps, write Python code to answer the causal question: `{self.query}`. Feel free to preprocess the data.
Important: Only use these approved packages: pandas, numpy, scipy, scikit-learn, statsmodels, dowhy, rdd, linearmodels, econml.
Use the methods from the above libraries to implement the method you chose. Be careful about implementation.
Make sure the code prints the final results, including:

1. The causal effect (the value only)
2. The standard deviation (the value only)
3. The causal inference method used (the method name only)
4. RCT: True/False/NA
5. The treatment variable
6. The outcome variable
7. The mediator variable (if applicable)
8. The covariates/control variables
9. Instrumental variable (if applicable)
10. Running variable (if applicable)
11. Temporal variable (if applicable)
12. Results of statistical tests, if applicable
13. Brief explanation for model choice
14. The regression formula, if applicable

If a variable is not applicable, print "NA" for it.
The code you write will be executed, and you will next analyze the output. To ease the process, please output one block of code, and make sure the code prints the key results and values. Everything between your first code block: ```` ```python ```` and ```` ``` ```` will be executed. If there is an error, you will have several attempts to correct the code. Hence, if there is an error, please fix it and rerun.

G Annotation Details

For each article we curate the following information:

- **Paper Name:** Name of the study
- **Description:** The description of the dataset that includes the collection process, purpose, and brief explanation of the variable names
- **Query:** Causal question associated with the dataset
- **Answer:** Causal effect derived in the paper
- **Standard Error:** Standard error associated with the causal effect estimate
- **Significant:** Binary variable indicating if the effect is statistically significant
- **Method:** The causal inference method
- **Treatment:** The name of the treatment variable in the dataset
- **Outcome:** The name of the outcome variable in the dataset
- **Control Covariates:** The control variables/confounders used in the estimation model
- **Interaction Variable:** The name of the variable that interacts with the treatment. This is used for measuring heterogeneous treatment effects
- **Instrument:** The variable used as an instrument. If an instrumental variable is not used, this is set to null
- **Running Variable:** The running variable for Regression Discontinuity Design (RDD). If RDD is not used, we set this to null
- **Temporal Variable:** The variable denoting the timing of treatments. This is used for difference-in-differences
- **State Variable:** The variable denoting the different participating entities. This is used for two-way fixed effects versions of difference-in-differences
- **Multi-RCT Treatment Variable:** The treatment type of interest. This is used in RCTs with multiple treatments
- **Data File:** The name of the csv file containing the data
- **Reference:** Reference to the original paper, where the result is found
- **Publication Year:** The year the original study was published

G.1 Annotation Example

Below we provide a sample annotation for a query based on the paper by Gerber et al. [39]

Annotation Example

- **Paper Name:** Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment
- **Description:** The randomized experiment aims to analyze the effect of different types of social pressure on voter behavior. A field experiment was conducted in Michigan ahead of the August 2006 primary election. Households were randomly assigned to a control group or one of four treatment groups: Civic Duty, Hawthorne, Self, Neighbors. Eleven days before the election, each treatment group received a different mailing:
  - Civic Duty: Emphasized the recipient's responsibility as a citizen to vote.
  - Hawthorne: Notified recipients that their voting behavior would be studied using public records, introducing mild social pressure.
  - Self: Listed the voting history of all registered voters in the household and noted that an updated chart would be mailed after the election.
  - Neighbors: Included both the household's and neighbors' voting records, implying public exposure of voting behavior.
  - Control Group: Received no mailing.

  Variables in the dataset:
  - sex: Participant's sex (male or female)
  - g2000, g2002, g2004: Voted in the 2000, 2002, and 2004 gubernatorial elections
  - p2000, p2002, p2004: Voted in the 2000, 2002, and 2004 primary elections
  - treatment: Assigned group (Civic Duty, Hawthorne, Neighbors, Self, or Control)
  - cluster: Cluster identifier for the unit
  - voted: Indicator for voting in the 2006 primary election
  - hh_id: Household ID
  - hh_size: Number of individuals in the household
  - yob: Year of birth of the participant
- **Query:** Does the Hawthorne scheme lead to an increase in voter turnout?
- **Answer:** 0.026
- **Standard Error:** 0.003
- **Significant:** Yes
- **Method:** OLS
- **Treatment:** treatment
- **Outcome:** voted
- **Control Covariates:** g2000, g2002, p2000, p2002, p2004
- **Interaction Variable:** null
- **Instrument:** null
- **Running Variable:** null
- **Temporal Variable:** null
- **State Variable:** null
- **Multi-RCT Treatment Variable:** Hawthorne
- **Data File:** voter_turnout_data.csv
- **Reference:** Table 3a
- **Publication Year:** 2008

H Data Source

The datasets used in this benchmark come from the following publications

Table 5: Dataset Sources

Publication Name	**Publication Reference**	**Data Reference**
Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment	[39]	[38]
Propensity Score Matching for methods for non-experimental causal studies	[31]	[30, 31, 53]
Can immigrants counteract employer discrimination? A factorial field experiment reveals the immutability of ethnic hierarchies	[95]	[94]
Using geographic variation in college proximity to estimate the return to schooling	[22]	[49]
Randomized experiments from non-random selection in U.S. House elections	[54]	[66]
The Long-run Effect of Abortion on Sexually Transmitted Infections	[27]	[28, 49]
Black Politicians Are More Intrinsically Motivated to Advance Blacks' Interests: A Field Experiment Manipulating Political Incentives	[14]	[49]
Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Castle Doctrine	[26]	[49]
Government Transfers and Political Support	[65]	[49]
Don't Take 'No' For An Answer: An Experiment With Actual Organ Donor Registrations	[50]	[49]
The Demand for, and Impact of, Learning HIV Status	[86]	[49]
Do Voters Affect or Elect Policies: Evidence from the U.S. House	[55]	[49, 28]
The effects of rural electrification in India: An instrumental variable approach at the household level	[83]	[82]

Publication Name	Publication Reference	Data Reference
Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania	[23]	[76]
Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference	[47]	[48]
Markets: The Fulton Fish Market	[43]	[29]
Punishment and Deterrence: Evidence from Drunk Driving	[46]	[29]
The causal effect of economic sanctions on political stability: A two-stage difference-in-differences analysis	[80]	[79]
Public Trust and Collaborative Governance: An Instrumental Variable Approach	[59]	[58]
Does Compulsory School Attendance Affect Schooling and Earnings?	[5]	[6]
Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement	[7]	[8]
The China Syndrome: Local Labor Market Effects of Import Competition in the United States	[9]	[10]
Do Wall Street Landlords Undermine Renters' Welfare?	[45]	[44]
Privatization and Quality of Carceral Healthcare: A difference-in-differences analysis of jails in the United States, 2008-2019	[103]	[102]
Estimating the impact of gubernatorial partisanship on policy settings and economic outcomes: A regression discontinuity approach	[56]	[57]
Do Congressional Candidates Have Reverse Coattails? Evidence from a Regression Discontinuity Design	[15]	[13]
Early Medicaid Expansions and Drug Overdose Mortality in the USA: a Quasi-experimental Analysis	[93]	[92]
Metrics Management and Bureaucratic Accountability: Evidence from Policing	[36]	[35]

Publication Name	Publication Reference	Data Reference
Double-Shift Schooling and Student Success: Quasi-experimental Evidence from Europe	[60]	[99]
Are Minimum Wages a Silent Killer? New Evidence on Drunk Driving Fatalities	[75]	[74]
Understanding the Impact of the 2018 Voter ID Pilots on Turnout at the London Local Elections: A Synthetic Difference-in-Difference Approach	[11]	[12]
The Minimum Legal Drinking Age and Morbidity in the US	[25]	[24]
How Partisan Is Local Law Enforcement? Evidence from Sheriff Cooperation with Immigration Authorities	[85]	[84]
Immigration Policies and Access to the Justice System: The Effect of Enforcement Escalations on Undocumented Immigrants and Their Communities	[33]	[32]
The effect of number of siblings and birth order on educational attainment: Empirical Evidence from Chinese General Social Survey	[96]	[37]
Do Televised Presidential Ads Increase Voter Turnout? Evidence from a Natural Experiment	[52]	[51]
Voting Made Safe and Easy: The Impact of e-voting on Citizen Perceptions	[2]	[3]
Can Learning Constituency Opinion Affect How Legislators Vote? Results from a Field Experiment	[16]	[17]
Doing Well by Doing Good: The Impact of Foreign Aid on Foreign Public Opinion	[41]	[40]
Were Newspapers More Interested in pro-Obama Letters to the Editor in 2008? Evidence from a Field Experiment	[18]	[19]
Foreign Direct Investors as Agents of Economic Transition: An Instrumental Variables Analysis	[64]	[61]

Publication Name	Publication Reference	Data Reference
Timing Is Everything? Primacy and Recency Effects in Voter Mobilization Campaigns	[70]	[69]
Computational and Robustness Reproducibility of "UN Peacekeeping and Democratization in Conflict-Affected Countries"	[68]	[67]
Incumbents Beware: The Impact of Offshoring on Elections	[73]	[72]
Does mislabeling COVID-19 elicit the perception of threat and reduce blame?	[97]	[98]
Undermining U.S. reputation: Chinese vaccines and aid and the alternative provision of public goods during COVID-19	[91]	[90]
China's Foreign Aid Political Drivers: Lessons from a Novel Dataset of Mask Diplomacy in Latin America during the COVID-19 Pandemic	[81]	[89]
Regression Discontinuity Designs Using Covariates	[21]	[20]
The influence of waiting times and sociopolitical variables on public trust in healthcare: A cross-sectional study of the NHS in England	[34]	[71]
Europeanisation beyond the EU: Tobacco Advertisement Restrictions in Swiss Cantons	[88]	[87]
Does 'right to work' imperil the right to health? The effect of labour unions on workplace fatalities	[101]	[100]
The Effect of Property Assessment Reductions on Tax Delinquency and Tax Foreclosure	[1]	[77]
The Impact of Recentralization on Public Services: A Difference-in-Differences Analysis of the Abolition of Elected Councils in Vietnam	[63]	[62]
Zika Epidemic and Birth Rates in Brazil	[78]	[4]

References

[1] Fernanda Alfaro, Dusan Paredes, and Mark Skidmore. The effect of property assessment reductions on tax delinquency and tax foreclosure. *National Tax Journal*, 78(2):415–433, None

2025. doi: 10.1086/735110. URL https://ideas.repec.org/a/ucp/nattax/doi10.1086-735110.html.

[2] R. Michael Alvarez, Ines Levin, Julia Pomares, and Marcelo Leiras. Voting made safe and easy: The impact of e-voting on citizen perceptions. *Political Science Research and Methods*, 1(1):117–137, 2013. doi: 10.1017/psrm.2013.2.

[3] R. Michael Alvarez, Ines Levin, Julia Pomares, and Marcelo Leiras. Replication data for: Voting Made Safe and Easy: The Impact of e-voting on Citizen Perceptions, 2015. URL https://doi.org/10.7910/DVN/24896.

[4] Leila Amorim. Replication Data for: Zika Epidemic and Birth Rates in Brazil, 2022. URL https://doi.org/10.7910/DVN/ENG0IY.

[5] Joshua D Angrist and Alan B Krueger. Does compulsory school attendance affect schooling and earnings? Working Paper 3572, National Bureau of Economic Research, December 1990.

[6] Joshua D. Angrist and Alan B. Krueger. Replication data for: Does Compulsory School Attendance Affect Schooling and Earnings?, 2009. URL https://doi.org/10.7910/DVN/ENLGZX.

[7] Joshua D Angrist and Victor Lavy. Using maimonides' rule to estimate the effect of class size on student achievement. Working Paper 5888, National Bureau of Economic Research, January 1997. URL http://www.nber.org/papers/w5888.

[8] Joshua D. Angrist and Victor Lavy. Replication data for: Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement, 2009. URL https://doi.org/10.7910/DVN/XRSUJU.

[9] David H. Autor, David Dorn, , and Gordon H. Hanson. The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review*, 103(6):2121–2168, October 2013. doi: 10.1257/aer.103.6.2121.

[10] David H. Autor, David Dorn, and Gordon H. Hanson. Replication data for: The china syndrome: Local labor market effects of import competition in the united states, 2019. URL https://doi.org/10.3886/E112670V1. Originally published 2013; replication dataset released 2019-10-11.

[11] Tom Barton. Understanding the impact of the 2018 voter id pilots on turnout at the london local elections: A synthetic difference-in-difference approach. *Political Science Research and Methods*, page 1–18, 2025. doi: 10.1017/psrm.2025.7.

[12] Tom Barton. Replication Data for: Understanding the Impact of the 2018 Voter ID Pilots on Turnout at the London Local Elections: A Synthetic Difference-in-Difference Approach, 2025. URL https://doi.org/10.7910/DVN/KULZYU.

[13] David Broockman. Replication Data for: Do Congressional Candidates Have Reverse Coattails? Evidence from a Regression Discontinuity Design, 2016. URL https://doi.org/10.7910/DVN/DCKUNX.

[14] David E. Broockman. Black politicians are more intrinsically motivated to advance blacks' interests: A field experiment manipulating political incentives. *American Journal of Political Science*, 57(3):521–536, 2013. doi: 10.1111/ajps.12018.

[15] David E. Broockman. Do congressional candidates have reverse coattails? evidence from a regression discontinuity design. *Political Analysis*, 17(4):418–434, 2017. doi: 10.1093/pan/mpp013.

[16] Daniel M. Butler and David W. Nickerson. Can learning constituency opinion affect how legislators vote? results from a field experiment. *Quarterly Journal of Political Science*, 6(1):55–83, August 2011. doi: 10.1561/100.00011019. URL https://ideas.repec.org/a/now/jlqjps/100.00011019.html.

[17] Daniel M. Butler and David W. Nickerson. Replication materials for: 'can learning constituency opinion affect how legislators vote? results from a field experiment', 2011. URL http://hdl.handle.net/10079/47d7x05.

[18] Daniel M. Butler and Emily Schofield. Were newspapers more interested in pro-obama letters to the editor in 2008? evidence from a field experiment. *American Politics Research*, 38 (2):356–371, 2010. doi: 10.1177/1532673X09349912. URL https://doi.org/10.1177/1532673X09349912.

[19] Daniel M. Butler and Emily Schofield. Replication materials for: 'were newspapers more interested in pro-obama letters to the editor in 2008? evidence from a field experiment.', 2010. URL http://hdl.handle.net/10079/bzkh1nf.

[20] Sebastian Calonico, Matias Cattaneo, Max Farrell, and Rocio Titiunik. Replication Data for: "Regression Discontinuity Designs Using Covariates", 2018. URL https://doi.org/10.7910/DVN/LPZLBF.

[21] Sebastian Calonico, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. Regression discontinuity designs using covariates. *The Review of Economics and Statistics*, 101(3):442–451, 07 2019. ISSN 0034-6535. doi: 10.1162/rest_a_00760. URL https://doi.org/10.1162/rest_a_00760.

[22] David Card. Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, National Bureau of Economic Research, October 1993. URL http://www.nber.org/papers/w4483.

[23] David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. Working Paper 4509, National Bureau of Economic Research, October 1993. URL http://www.nber.org/papers/w4509.

[24] Christopher Carpenter and Carlos Dobkin. Replication Data for: "The Minimum Legal Drinking Age and Morbidity in the US", 2016. URL https://doi.org/10.7910/DVN/Q9VQIU.

[25] Christopher Carpenter and Carlos Dobkin. The minimum legal drinking age and morbidity in the united states. *The Review of Economics and Statistics*, 99(1):95–104, 03 2017. ISSN 0034-6535. doi: 10.1162/REST_a_00615. URL https://doi.org/10.1162/REST_a_00615.

[26] Cheng Cheng and Mark Hoekstra. Does strengthening self-defense law deter crime or escalate violence? evidence from castle doctrine. Working Paper 18134, National Bureau of Economic Research, June 2012. URL http://www.nber.org/papers/w18134.

[27] Christopher Cornwell and Scott Cunningham. The long-run effect of abortion on sexually transmitted infections. *American Law and Economics Review*, 15(1):381–407, 01 2013. ISSN 1465-7252. doi: 10.1093/aler/ahs019. URL https://doi.org/10.1093/aler/ahs019.

[28] Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, 2021. ISBN 9780300251685. URL http://www.jstor.org/stable/j.ctv1c29t27.

[29] Scott Cunningham. Data and program files for causal inference: The mixtape. GitHub repository, 2025. URL https://github.com/scunning1975/mixtape.

[30] Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999. doi: 10.1080/01621459.1999.10473858. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473858.

[31] Rajeev H. Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161, 02 2002.

[32] Reva Dhingra, Mitchell Kilborn, and Olivia Woldemikael. Replication Data for: Immigration Policies and Access to the Justice System: The Effect of Enforcement Escalations on Undocumented Immigrants and Their Communities, 2020. URL https://doi.org/10.7910/DVN/RGZWNJ.

[33] Reva Dhingra, Mitchell Kilborn, and Olivia Woldemikael. Immigration policies and access to the justice system: The effect of enforcement escalations on undocumented immigrants and their communities. *Political Behavior*, 44(3):1359–1387, 2022. doi: 10.1007/s11109-020-09663-w. URL https://doi.org/10.1007/s11109-020-09663-w.

[34] H. Dorussen, M.E. Hansen, S.D. Pickering, J. Reifler, T.J. Scotto, Y. Sunahara, and D. Yen. The influence of waiting times and sociopolitical variables on public trust in healthcare: A cross-sectional study of the nhs in england. *Public Health in Practice*, 7:100484, 2024. ISSN 2666-5352. doi: https://doi.org/10.1016/j.puhip.2024.100484. URL https://www.sciencedirect.com/science/article/pii/S2666535224000211.

[35] Laurel Eckhouse. Replication Data for: Metrics Management and Bureaucratic Accountability: Evidence from Policing, 2021. URL https://doi.org/10.7910/DVN/3E7JXB.

[36] Laurel Eckhouse. Metrics management and bureaucratic accountability: Evidence from policing. *American Journal of Political Science*, 66(2):385–401, 2022. doi: https://doi.org/10.1111/ajps.12661. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12661.

[37] Xiong Feng, Leizhen Zang, Ling Zhou, and Fei Liu. Replication Data for: The effect of number of siblings and birth order on educational attainment: Empirical Evidence from Chinese General Social Survey, 2020. URL https://doi.org/10.7910/DVN/HHSREH.

[38] Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. Replication materials for "social pressure and voter turnout: Evidence from a large-scale field experiment", 2008. URL http://hdl.handle.net/10079/c7507a0d-097a-4689-873a-7424564dfc82.

[39] Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102 (1):33–48, 2008. doi: 10.1017/S000305540808009X.

[40] Benjamin Goldsmith, Yusaku Horiuchi, and Terence Wood. Replication data for: Doing Well by Doing Good: The Impact of Foreign Aid on Foreign Public Opinion, 2018. URL https://doi.org/10.7910/DVN/IHMFPJ.

[41] Benjamin E. Goldsmith, Yusaku Horiuchi, and Terence Wood. Doing well by doing good: The impact of foreign aid on foreign public opinion. *Quarterly Journal of Political Science*, 9(1): 87–114, March 2014. doi: 10.1561/100.00013036. URL https://ideas.repec.org/a/now/jlqjps/100.00013036.html.

[42] Rebecca Goldstein and Hye Young You. Cities as lobbyists. *American Journal of Political Science*, 61(4):864–876, 2017. doi: https://doi.org/10.1111/ajps.12306. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12306.

[43] Kathryn Graddy. Markets: The fulton fish market. *Journal of Economic Perspectives*, 20 (2):207–220, June 2006. doi: 10.1257/jep.20.2.207. URL https://www.aeaweb.org/articles?id=10.1257/jep.20.2.207.

[44] Umit Gurun, Jiabin Wu, Steven Xiao, and Serena Xiao. Replication Data for "Do Wall Street Landlords Undermine Renters' Welfare?", 2022. URL https://doi.org/10.7910/DVN/HCWJRW.

[45] Umit G Gurun, Jiabin Wu, Steven Chong Xiao, and Serena Wenjing Xiao. Do wall street landlords undermine renters' welfare? *The Review of Financial Studies*, 36(1):70–121, 03 2022. ISSN 0893-9454. doi: 10.1093/rfs/hhac017. URL https://doi.org/10.1093/rfs/hhac017.

[46] Benjamin Hansen. Punishment and deterrence: Evidence from drunk driving. *American Economic Review*, 105(4):1581–1617, April 2015. doi: 10.1257/aer.20130189. URL https://www.aeaweb.org/articles?id=10.1257/aer.20130189.

[47] Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236, 2007.

[48] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Replication data for: Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference, 2007. URL https://doi.org/10.7910/DVN/RWUY8G.

[49] Nick Huntington-Klein and Malcolm Barrett. *causaldata: Example Data Sets for Causal Inference Textbooks*, 2024. URL https://github.com/nickch-k/causaldata. R package version 0.1.4.

[50] Judd B Kessler and Alvin E Roth. Don't take 'no' for an answer: An experiment with actual organ donor registrations. Working Paper 20378, National Bureau of Economic Research, August 2014. URL http://www.nber.org/papers/w20378.

[51] Jonathan S. Krasno and Donald P. Green. Replication materials for 'do televised presidential ads increase voter turnout? evidence from a natural experiment.', 2008. URL http://hdl.handle.net/10079/9a83be42-d671-4264-8d72-56777c7ea529.

[52] Jonathan S. Krasno and Donald P. Green. Do televised presidential ads increase voter turnout? evidence from a natural experiment. *The Journal of Politics*, 70(1):245–261, 2008. doi: 10.1017/S0022381607080176. URL https://doi.org/10.1017/S0022381607080176.

[53] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620, September 1986. doi: None. URL https://ideas.repec.org/a/aea/aecrev/v76y1986i4p604-20.html.

[54] David S. Lee. Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142:675–697, 2008. doi: 10.1016/j.jeconom.2007.05.004.

[55] David S. Lee, Enrico Moretti, and Matthew J. Butler. Do voters affect or elect policies? evidence from the u. s. house*. *The Quarterly Journal of Economics*, 119(3):807–859, 08 2004. ISSN 0033-5533. doi: 10.1162/0033553041502153. URL https://doi.org/10.1162/0033553041502153.

[56] Andrew Leigh. Estimating the impact of gubernatorial partisanship on policy settings and economic outcomes: A regression discontinuity approach. *European Journal of Political Economy*, 24(1):256–268, 2008. ISSN 0176-2680. doi: https://doi.org/10.1016/j.ejpoleco.2007.06.003. URL https://www.sciencedirect.com/science/article/pii/S0176268007000572.

[57] Andrew Leigh. Estimating the Impact of Gubernatorial Partisanship on Policy Settings and Economic Outcomes : A Regression Discontinuity Approach, 2013. URL https://doi.org/10.7910/DVN/JPDQ5P.

[58] Yixin Liu. Replication Data for: Public Trust and Collaborative Governance: An Instrumental Variable Approach, 2022. URL https://doi.org/10.7910/DVN/ATDFQN.

[59] Yixin Liu. Public trust and collaborative governance: an instrumental variable approach. *Public Management Review*, 26(2):421–442, 2024. doi: 10.1080/14719037.2022.2095003. URL https://doi.org/10.1080/14719037.2022.2095003.

[60] Lester Lusher and Vasil Yasenov. Double-shift schooling and student success: Quasi-experimental evidence from europe. *Economics Letters*, 139:36–39, 2016. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2015.12.009. URL https://www.sciencedirect.com/science/article/pii/S0165176515005182.

[61] Edmund Malesky. Replication Data for: Foreign Direct Investors: Agents of Economic Transition An Instrumental Variables Analysis, 2015. URL https://doi.org/10.7910/DVN/PDVSW5.

[62] Edmund Malesky. Replication Data for: The Impact of Recentralization on Public Services: A Difference-in-Differences Analysis of the Abolition of Elected Councils in Vietnam, 2015. URL https://doi.org/10.7910/DVN/IUG2C4.

[63] Edmund Malesky, Cuong Nguyen, and Anh Tran. The impact of recentralization on public services: A difference-in-differences analysis of the abolition of elected councils in vietnam. MPRA Paper 54187, University Library of Munich, Germany, Aug 2013. URL https://ideas.repec.org/p/pra/mprapa/54187.html.

[64] Edmund J. Malesky. Foreign direct investors as agents of economic transition: An instrumental variables analysis. *Quarterly Journal of Political Science*, 4(1):59–85, March 2009. doi: 10.1561/100.00008068. URL https://ideas.repec.org/a/now/jlqjps/100.00008068.html.

[65] Marco Manacorda, Edward Miguel, and Andrea Vigorito. Government transfers and political support. *American Economic Journal: Applied Economics*, 3(3):1–28, July 2011. doi: 10.1257/app.3.3.1. URL https://www.aeaweb.org/articles?id=10.1257/app.3.3.1.

[66] Mauricio Olivares and Ignacio Sarmiento-Barbieri. *RATest: Randomization Tests*, 2022. URL https://CRAN.R-project.org/package=RATest. R package version 0.1.10.

[67] Christian Oswald and Julian Walterskirchen. Replication Data for: Computational and robustness reproducibility of "UN Peacekeeping and Democratization in Conflict-Affected Countries", 2024. URL https://doi.org/10.7910/DVN/G37BHE.

[68] Christian Oswald and Julian Walterskirchen. Computational and robustness reproducibility of "un peacekeeping and democratization in conflict-affected countries". I4R Discussion Paper Series 138, The Institute for Replication (I4R), 2024. URL https://ideas.repec.org/p/zbw/i4rdps/138.html.

[69] Costas Panagopoulos. Replication materials for: 'timing is everything? primacy and recency effects in voter mobilization campaigns', 2010. URL http://hdl.handle.net/10079/6a4d6ad0-1f4e-47df-aa52-6a68949d9ab5.

[70] Costas Panagopoulos. Timing is everything? primacy and recency effects in voter mobilization campaigns. *Political Behavior*, 33(1):79–93, 2011. doi: 10.1007/s11109-010-9125-x. URL https://doi.org/10.1007/s11109-010-9125-x.

[71] Steve Pickering. Replication Data for: Influence of Waiting Times and Sociopolitical Variables on Public Trust in Healthcare: A Cross-Sectional Study of the NHS in England, 2023. URL https://doi.org/10.7910/DVN/AQYYNK.

[72] Stephanie Rickard. Replication Data for: Incumbents Beware: The Impact of offshoring on elections, 2021. URL https://doi.org/10.7910/DVN/X1JL2H.

[73] Stephanie Rickard. Incumbents beware: the impact of offshoring on elections. LSE Research Online Documents on Economics 107517, London School of Economics and Political Science, LSE Library, Apr 2022. URL https://ideas.repec.org/p/ehl/lserod/107517.html.

[74] Joseph Sabia, Melinda Pitts, and Laura Argys. Replication data for: "Are Minimum Wages a Silent Killer? New Evidence on Drunk Driving Fatalities", 2018. URL https://doi.org/10.7910/DVN/PYWQYU.

[75] Joseph J. Sabia, M. Melinda Pitts, and Laura M. Argys. Are minimum wages a silent killer? new evidence on drunk driving fatalities. *The Review of Economics and Statistics*, 101(1):192–199, March 2019. doi: None. URL https://ideas.repec.org/a/tpr/restat/v101y2019i1p192-199.html.

[76] Justin M Shea. *115 Data Sets from "Introductory Econometrics: A Modern Approach, 7e" by Jeffrey M. Wooldridge*, 2024. URL https://cran.r-project.org/web/packages/wooldridge/index.html. R package version 0.1.4.

[77] Mark Skidmore. Replication for The Effect of Property Assessment Reductions on Tax Delinquency and Tax Foreclosure National Tax Journal, 2025. URL https://doi.org/10.7910/DVN/YZOUI2.

[78] Marcelo Taddeo, Leila Amorim, and Rosana Aquino. Causal measures using generalized difference-in-difference approach with nonlinear models. *Statistics and Its Interface*, 15: 399–413, 01 2022. doi: 10.4310/21-SII704.

[79] Dongan Tan. Replication Data for: The causal effect of economic sanctions on political stability: A two-stage difference-in-differences analysis, 2024. URL https://doi.org/10.7910/DVN/KCSVWH.

[80] Dongan Tan. The causal effect of economic sanctions on political stability: A two-stage difference-in-differences analysis. *International Journal of Social Welfare*, 34(1):e12707, 2025. doi: https://doi.org/10.1111/ijsw.12707. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ijsw.12707.

[81] Diego Telias and Francisco Urdinez. China's foreign aid political drivers: Lessons from a novel dataset of mask diplomacy in latin america during the covid-19 pandemic. *Journal of Current Chinese Affairs*, 51(1):108–136, 2022. doi: 10.1177/18681026211020763. URL https://doi.org/10.1177/18681026211020763.

[82] Daniel Thomas, SP Harish, Ryan Kennedy, and Johannes Urpelainen. The Effects of Rural Electrification in India: An Instrumental Variable Approach at the Household Level , 2020. URL https://doi.org/10.7910/DVN/CIPVBK.

[83] Daniel Robert Thomas, S.P. Harish, Ryan Kennedy, and Johannes Urpelainen. The effects of rural electrification in india: An instrumental variable approach at the household level. *Journal of Development Economics*, 146:102520, 2020. ISSN 0304-3878. doi: https://doi.org/10.1016/j.jdeveco.2020.102520. URL https://www.sciencedirect.com/science/article/pii/S030438782030095X.

[84] Daniel M Thompson. Replication Data for: How Partisan Is Local Law Enforcement? Evidence from Sheriff Cooperation with Immigration Authorities, 2020. URL https://doi.org/10.7910/DVN/CFASH6.

[85] DANIEL M. THOMPSON. How partisan is local law enforcement? evidence from sheriff cooperation with immigration authorities. *American Political Science Review*, 114:222–236, 2020. doi: 10.1017/S0003055419000613.

[86] Rebecca L. Thornton. The demand for, and impact of, learning hiv status. *American Economic Review*, 98(5):1829–63, December 2008. doi: 10.1257/aer.98.5.1829. URL https://www.aeaweb.org/articles?id=10.1257/aer.98.5.1829.

[87] Philipp Trein. Replication Data for: Europeanisation beyond the EU: Tobacco Advertisement Restrictions in Swiss Cantons, 2016. URL https://doi.org/10.7910/DVN/0T5AN0.

[88] Philipp Trein. Europeanisation beyond the european union: tobacco advertisement restrictions in swiss cantons. *Journal of Public Policy*, 37(2):113–142, 2017. doi: 10.1017/S0143814X16000167.

[89] Francisco Urdinez. Replication Data for: China's Foreign Aid Political Drivers: Lessons from a Novel Dataset of Mask Diplomacy in Latin America During the COVID-19 Pandemic, 2021. URL https://doi.org/10.7910/DVN/EIAXSE.

[90] Francisco Urdinez. Replication Data for: Undermining US reputation: Chinese vaccines and aid and the alternative provision of public goods during COVID-19, 2023. URL https://doi.org/10.7910/DVN/KNG7CY.

[91] Francisco Urdinez. Undermining u.s. reputation: Chinese vaccines and aid and the alternative provision of public goods during covid-19. *The Review of International Organizations*, 19 (2):243–268, April 2024. doi: 10.1007/s11558-023-09508-1. URL https://ideas.repec.org/a/spr/revint/v19y2024i2d10.1007_s11558-023-09508-1.html.

[92] Atheendar Venkataramani. Replication Data for: Venkataramani and Chatterjee "Early Medicaid Expansions and Drug Overdose Mortality in the USA", 2018. URL https://doi.org/10.7910/DVN/9ZS4KR.

[93] Atheendar S. Venkataramani and Paula Chatterjee. Early medicaid expansions and drug overdose mortality in the usa: a quasi-experimental analysis. *Journal of General Internal Medicine*, 34:23–25, 2018. URL https://api.semanticscholar.org/CorpusID:52308028.

[94] Kåre Vernby. Replication data for "Can immigrants counteract employer discrimination? A factorial field experiment reveals the immutability of ethnic hierarchies", 2019. URL https://doi.org/10.7910/DVN/HVRLOS.

[95] Kåre Vernby and Rafaela Dancygier. Can immigrants counteract employer discrimination? a factorial field experiment reveals the immutability of ethnic hierarchies. *PLOS ONE*, 14:1–19, 07 2019. doi: 10.1371/journal.pone.0218044. URL https://doi.org/10.1371/journal.pone.0218044.

[96] Feng Xiong, Leizhen Zang, Ling Zhou, and Fei Liu. The effect of number of siblings and birth order on educational attainment: Empirical evidence from chinese general social survey. *International Journal of Educational Development*, 78:102270, 2020. ISSN 0738-0593. doi: https://doi.org/10.1016/j.ijedudev.2020.102270. URL https://www.sciencedirect.com/science/article/pii/S0738059320304296.

[97] Chengxin Xu and Yixin Liu. Does mislabeling covid-19 elicit the perception of threat and reduce blame? *Journal of Behavioral Public Administration*, 4(2), May 2021. doi: 10.30636/jbpa.42.225. URL https://journal-bpa.org/index.php/jbpa/article/view/225.

[98] Chengxin Xu and Yixin Liu. Replication Data for: Does Mislabeling COVID-19 Elicit the Perception of Threat and Reduce Blame?, 2021. URL https://doi.org/10.7910/DVN/CDUROD.

[99] Vasil Yasenov. Replication Data for: Double-Shift Schooling and Student Success: Quasi-experimental Evidence from Europe, 2022. URL https://doi.org/10.7910/DVN/6AYLVO.

[100] Michael Zoorob. Replication Data for: Does 'right to work' imperil the right to health? The effect of labour unions on workplace fatalities, 2018. URL https://doi.org/10.7910/DVN/UVCZ5D.

[101] Michael Zoorob. Does 'right to work' imperil the right to health? the effect of labour unions on workplace fatalities. *Occupational and Environmental Medicine*, 75(10):736–738, 2018. ISSN 1351-0711. doi: 10.1136/oemed-2017-104747. URL https://oem.bmj.com/content/75/10/736.

[102] Michael Zoorob. Replication Data for: Privatization and Quality of Carceral Healthcare: A difference-in-differences analysis of jails in the United States, 2008-2019, 2022. URL https://doi.org/10.7910/DVN/91S9AP.

[103] Michael James Zoorob. Privatization and quality of carceral healthcare: A difference-in-differences analysis of jails in the united states, 2008-2019. *International Journal of Public Administration*, 47(1):14–25, 2024. doi: 10.1080/01900692.2022.2081177. URL https://doi.org/10.1080/01900692.2022.2081177.