Society for Causal Inference
EST. 2020

ABSTRACT BOOK

2026 American Causal Inference Conference

May 11-14, 2026

Generalizability/Transportability

Meta-analysis through Low-Rank Basis Hunting Wenqi Shi* Wenqi Shi, Kosuke Imai, Yi Zhang,

A central challenge of meta-analysis is that the populations underlying existing studies often differ from the target population in unknown ways. We study the problem of predicting function-valued quantities, such as regression functions and conditional average treatment effect functions, for a new target population using only study-level covariates and estimates. Our approach assumes a shared low-rank structure, in which the true function from each study lies within the convex hull of a small set of latent basis functions. To recover these basis functions, we extend the Successive Projection Algorithm to the functional setting, incorporating a denoised basis-hunting step. We then model the relationship between study-level covariates and the corresponding mixing weights using flexible semi-parametric or non-parametric methods. This framework enables meta-analytic prediction even when individual-level data are unavailable to analysts. For uncertainty quantification, we construct prediction intervals via conformal prediction and show that, under exchangeability and mild estimation-error conditions, these intervals achieve asymptotically valid marginal coverage. We demonstrate the effectiveness of the proposed methodology through both simulation studies and empirical applications.

Generalizability/Transportability

Transporting treatment effects by calibrating large-scale observational outcomes Harrison Li* Harrison Li,

A high-quality experimental dataset is typically much smaller than a corresponding observational dataset. We propose an estimation and inference method for a transported average treatment effect in this setting when there are imperfect and possibly confounded measurements of the outcome of interest in the observational dataset. Our point estimate involves a low-dimensional calibration of a treatment-control contrast in the observational outcome to the experimental data to estimate the conditional average treatment effect (CATE), followed by a sample average of this estimated CATE over the observational data. Unlike existing methods, our approach does not require positivity (overlap). We precisely study the asymptotic behavior of our procedure even when the low-dimensional calibration is misspecified. As long as the observational dataset size grows sufficiently quickly relative to the experimental dataset size, our estimator achieves a notion of semiparametric efficiency studied in recent work on semi-supervised learning for our limiting estimand, which has close connections to well-studied projection estimands. We illustrate the stability of our methodology compared to existing proposals for transporting average treatment effects under realistic positivity violations using simulations and a data example involving field experiments and satellite imagery to estimate the average effect of crop rotation on maize (corn) yields over a large area of the Midwestern United States.

Generalizability/Transportability

Empirical Bayes Double Shrinkage for Combining Biased and Unbiased Causal Estimates
Evan Rosenman* Evan Rosenman, Luke Miratrix, Francesca Dominici,

Motivated by the proliferation of observational datasets and the need to integrate non-randomized evidence with randomized controlled trials, causal inference researchers have recently proposed several new methodologies for combining biased and unbiased estimators. We contribute to this growing literature by developing a new class of estimators for the data-combination problem: double-shrinkage estimators. Double-shrinkers first compute a data-driven convex combination of the the biased and unbiased estimators, and then apply a final, Stein-like shrinkage toward zero. Such estimators do not require hyperparameter tuning, and are targeted at multidimensional causal estimands, such as vectors of conditional average treatment effects (CATEs). We derive several workable versions of double-shrinkage estimators and propose a method for constructing valid Empirical Bayes confidence intervals. We also demonstrate the utility of our estimators using simulations on data from the Women’s Health Initiative.

Generalizability/Transportability

Data Fusion for High-Resolution Estimation Roshni Sahoo* Roshni Sahoo, Amy Guan, Marissa Reitsma, Joshua Salomon, Stefan Wager,

High-resolution estimates of population health indicators are critical for precision public health. We propose a method for high-resolution estimation that fuses distinct data sources: an unbiased, low-resolution data source (e.g. aggregated administrative data) and a potentially biased, high-resolution data source (e.g. individual-level online survey responses). We assume that the potentially biased, high-resolution data source is generated from the population under a model of sampling bias where observables can have arbitrary impact on the probability of response but the difference in the log probabilities of response between units with the same observables is linear in the difference between sufficient statistics of their observables and outcomes. Our data fusion method learns a distribution that is closest (in the sense of KL divergence) to the online survey distribution and consistent with the aggregated administrative data and our model of sampling bias. This method outperforms baselines that rely on either data source alone on a testbed that includes repeated measurements of three indicators measured by both the (online) Household Pulse Survey and ground-truth data sources at two geographic resolutions over the same time period.

Generalizability/Transportability

More efficient transportation of causal effects using mediators Xinyi Zhang* Xinyi Zhang, Kara Rudolph, Richard Liu, Michele Santacatterina, Iván Díaz,

We develop flexible semiparametric estimators of the average treatment effect transported to a target population that leverage mediators to achieve identification, and knowledge of prior population heterogeneity to achieve improved efficiency. We first propose a one-step semiparametric estimator that assumes knowledge of the variables driving the heterogeneity across populations. This approach enables transport even when not all covariates are observed in the target population, and improves efficiency by adjusting only for those covariates. We then introduce a collaborative one-step semiparametric estimator that further improves efficiency without requiring knowledge of which covariates drive study heterogeneity. Collectively, these estimators advance our ability to transport causal effects by improving efficiency in the presence of mediation and intermediate confounding through targeted dimension reduction strategies. We use simulation to examine finite-sample performance and apply our estimators to a large-scale housing mobility trial.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Efficient Difference-in-Differences Estimation when Outcomes are Missing at Random
Lorenzo Testa* Lorenzo Testa, Edward Kennedy, Matthew Reimherr,

The Difference-in-Differences (DiD) method is a fundamental tool for causal inference, yet its application is often complicated by missing data. Although recent work has developed robust DiD estimators for complex settings like staggered treatment adoption, these methods typically assume complete data and fail to address the critical challenge of outcomes that are missing at random (MAR) — a common problem that invalidates standard estimators. We develop a rigorous framework, rooted in semiparametric theory, for identifying and efficiently estimating the Average Treatment Effect on the Treated (ATT) when either pre- or post-treatment (or both) outcomes are missing at random. We first establish nonparametric identification of the ATT under two minimal sets of sufficient conditions. For each, we derive the semiparametric efficiency bound, which provides a formal benchmark for asymptotic optimality. We then propose novel estimators that are asymptotically efficient, achieving this theoretical bound. A key feature of our estimators is their multiple robustness, which ensures consistency even if some nuisance function models are misspecified. We validate the properties of our estimators and showcase their broad applicability through an extensive simulation study and a real-world example.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Difference-in-differences with stochastic policy shifts of continuous treatments Michael Jetsupphasuk* Michael Jetsupphasuk, Chenwei Fang, Didong Li, Michael Hudgens,

Treatment effects under stochastic policy shifts quantify differences in outcomes across counterfactual scenarios with varying treatment distributions. Stochastic policy shifts generalize common notions of treatment effects since they include deterministic interventions (e.g., all individuals treated versus none treated) as a special case. While stochastic policy effects have been examined under causal exchangeability, they have not been integrated into the difference-in-differences (DiD) framework, which relies on parallel trends rather than exchangeability. In this paper, nonparametric efficient estimators of stochastic intervention effects are developed under a DiD setup with continuous treatments. The proposed causal estimand is the average stochastic dose effect among the treated, where the stochastic dose effect is the contrast between potential outcomes under a counterfactual dose distribution and no treatment. A specific counterfactual dose distribution, the exponential tilt, is considered, which increments the conditional density function of the continuous dose. A nonparametric estimator is proposed that allows for data-adaptive, machine learning nuisance function estimation. Under mild convergence rate conditions, the estimator is shown to be root-n consistent and asymptotically normal with variance attaining the nonparametric efficiency bound. The proposed method is used to study the effect of hydraulic fracturing activity on employment and income.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Balancing Spatial Confounding and Spillovers in Synthetic Control Designs Giulio Grossi* Giulio Grossi,

In policy evaluation studies with administrative units such as cities, regions, or countries, policy interventions often generate spillover effects to nearby or connected units, violating the no-interference assumption. At the same time, spatially correlated unobserved confounders frequently affect both treatment assignment and outcomes. In this setting, synthetic control methods face a fundamental trade-off: using distant donors may induce spatial confounding bias, while relying on nearby donors may reduce confounding at the cost of spillover contamination.
We propose a design-based framework that makes this trade-off explicit in the context of synthetic control and provides an operational pipeline to diagnose and quantify spatial contamination. Our approach is inclusive by default: rather than assuming the existence of pure control units, we retain all potential donors and introduce contamination indices that measure the exposure of synthetic counterfactuals to spillovers.
The framework accommodates alternative exposure mappings, based on geographic proximity or network structures, treated as objects of weak identification and assessed through pre-treatment diagnostics and spatial placebo tests. Using a realistic simulation study, we characterize regimes in which spatial confounding or spillover bias dominates and show how design choices in synthetic control critically shape causal conclusions.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Causal Inference in Staggered Adoption Panels: The Correct Comparison Set Depends on Your Intervention Ian Lundberg* Ian Lundberg, Yaling Xu,

A powerful data structure for causal inference is staggered adoption: many units are observed over many time periods, during which some units adopt an irreversible treatment at various time points while others remain untreated. Popular methods that apply in this setting include difference in difference, fixed effects, matching, and synthetic control. All of these methods compare the future outcomes of treatment adopters and non-adopters to answer the question: for units who adopt treatment, what outcomes would have been realized if they had remained untreated? We show that this question actually hides two very different causal questions of interest. The first question is what would have happened if a treated unit had remained untreated over all periods until outcome measurement. The second question is what would have happened if a treated unit had remained untreated at the particular period when they in fact became treated. Popular methods such as synthetic control are often interpreted with respect to the former question (about a longitudinal treatment). We show that they actually answer the latter question (about a point-in-time treatment). The distinction is especially relevant if many units become treated at many time points. We illustrate by a sociological application in the National Longitudinal Survey of Youth where the unit is a person, time is age, treatment is becoming a parent, and the outcome is employment.

Machine Learning and Causal Inference

Demystifying Proximal Causal Inference Grace Ringlein* Grace Ringlein, Trang Nguyen, Peter Zandi, Elizabeth Stuart, Harsh Parikh,

Proximal causal inference (PCI) has emerged as a promising framework for identifying and estimating causal effects in the presence of unobserved confounders. While many traditional causal inference methods rely on the assumption of no unobserved confounding, this assumption is likely often violated. PCI mitigates this challenge by relying on an alternative set of assumptions regarding the relationships between treatment, outcome, and auxiliary variables that serve as proxies for unmeasured confounders. We review existing identification results, discuss the assumptions necessary for valid causal effect estimation via PCI, and compare different PCI estimation methods. We offer practical guidance on operationalizing PCI, with a focus on selecting and evaluating proxy variables using domain knowledge, measurement error perspectives, and negative control analogies. Through conceptual examples, we demonstrate tensions in proxy selection and discuss the importance of clearly defining the unobserved confounding mechanism. By bridging formal results with applied considerations, this work aims to demystify PCI, encourage thoughtful use in practice, and identify open directions for methodological development and empirical research.

Machine Learning and Causal Inference

Perturbed Double Machine Learning: Nonstandard Inference Beyond the Parametric Length Mengchu Zheng* Mengchu Zheng, Matteo Bonvini, Zijian Guo,

We study inference on a low dimensional functional $beta$ in the presence of possibly infinite dimensional nuisances. Classical inference typically relies on nuisance error being asymptotically negligible; for example, estimators based on the influence curve of the parameter (Double/Debiased Machine Learning estimators) are asymptotically Gaussian when the nuisance estimators converge at rates faster than $n^{-1/4}$. Although such negligibility can hold even in nonparametric classes, it can be restrictive. To relax this requirement, we propose Perturbed Double Machine Learning (Perturbed DML) to ensure valid inference even when nuisances converge at rates slower than $n^{-1/4}$. Our proposal 1) injects randomness into the nuisance estimation to generate multiple perturbed nuisance models, each yielding a $beta$ estimate and a Wald interval, and 2) filters out perturbations whose deviations from the original DML estimate exceed a threshold. For Lasso nuisance learners, we show that, with high probability, at least one perturbation produces nuisance estimates close enough to the truth that the associated $beta$ estimator is near an oracle estimator with knowledge of the true nuisances. Taking the union of retained intervals delivers valid coverage even when the DML estimator converges more slowly than $n^{-1/2}$. The framework extends to general machine learning nuisance learners, and simulations show that Perturbed DML can have coverage when state-of-art methods fail.

Machine Learning and Causal Inference

Kernel Debiased Plug-in Estimation based on the Universal Least Favorable Submodel
Ivana Malenica* Ivana Malenica, Haiyi Chen,

The one-step universal least favorable path kernel debiased plug-in estimator (ULFP-KDPE) constructs asymptotically efficient estimators of pathwise differentiable target parameters by debiasing an initial estimator using reproducing kernel Hilbert space (RKHS) geometry and universal least favorable paths. In this work, we place ULFP-KDPE on a precise functional-analytic foundation and establish results relevant to higher-order efficiency. We formulate the ULFP update as a nonlinear ordinary differential equation on densities and establish local and global existence and uniqueness of solutions. Under standard regularity conditions, the resulting estimator is unbiased and achieves the semiparametric efficiency bound. We show that ULFP-KDPE (i) simultaneously debiases all pathwise differentiable target parameters satisfying our regularity conditions, including multidimensional targets, (ii) does not require explicit knowledge of the influence function for implementation, and (iii) is computationally tractable. In finite samples, the solved score equations may be sufficiently rich to approximate higher-order efficient influence functions, yielding improved small-sample performance. Because the proposed method is efficient influence-function (EIF) free, it also offers practical advantages in settings with limited overlap or positivity challenges, or when the EIF is not computationally feasible or explicitly known. We numerically illustrate the proposed method and validate the the

Machine Learning and Causal Inference

Sharp Structure-Agnostic Lower Bounds for General Linear Functional Estimation Jikai Jin* Jikai Jin, Vasilis Syrgkanis,

We establish a general statistical optimality theory for estimation problems where the target parameter is a linear functional of an unknown nuisance component that must be estimated from data. This formulation covers many causal and predictive parameters and has applications to numerous disciplines. We adopt the structure-agnostic framework introduced by Balakrishnan et al. [2023], which poses no structural properties on the nuisance functions other than access to black-box estimators that achieve some statistical estimation rate. This framework is particularly appealing when one is only willing to consider estimation strategies that use non-parametric regression and classification oracles as black-box sub-processes. Within this framework, we pove general results for the best attainable error rates. Notably, we differentiate between two regimes in which double robustness can and cannot be achieved and in which first-order debiasing yields different error rates. We show that first-order debiasing is simultaneously optimal in both regimes. We instantiate our theory by deriving optimal error rates that recover existing results and extend to various settings of interest, including the case when the nuisance is defined by generalized regressions and when covariate shift is present.

Machine Learning and Causal Inference

Principal-Component HAL for Scalable Nonparametric Regression Mingxun Wang* Mingxun Wang, Carlos Meixide, Alejandro Schuler, Mark van der Laan,

The Highly Adaptive Lasso (HAL) is a nonparametric minimum loss estimator for many targets, including conditional means (regression) and densities, that attains minimax-optimal convergence rates under minimal smoothness assumptions (via sectional variation norm control). In practice, standard HAL can be computationally prohibitive because its indicator-basis expansion can be extremely large in moderate to high dimensions. We introduce Principal Component–based HAL (PCHAL), Principal Component–based Highly Adaptive Ridge (PCHAR), and Principal Component–based generalized HAL (PCHAgL), which perform outcome-blind dimension reduction by projecting the HAL feature space onto the leading principal components of its Gram/kernel structure. This step reduces the effective basis dimension to at most n components, so fitting becomes a (penalized) linear regression with at most n predictors (the sample size)—often far smaller than the original HAL basis. We show that PCHAL/PCHAR/PCHAgL achieve rates equivalent to their original counterparts under comparable complexity control, while delivering substantial computational gains and competitive finite-sample performance. These results provide a scalable foundation for HAL-based nuisance estimation in HAL-TMLE workflows for causal inference.

Causal Inference in Networks

Causal Inference with Noisy Network Representations of Latent Interference Structure
Heejong Bong* Heejong Bong, Elizaveta Levina, Ji Zhu,

We study causal inference under interference when the observed network is only a noisy measurement of the underlying interactions that generate spillover effects. While standard network interference assumptions restrict interference to observed neighbors, many real world systems involve latent interactions that create influence even through weak or unobserved ties. Modeling interference through a latent interaction structure captures these broader dependencies but breaks key positivity conditions required for existing propensity based estimators. To address this challenge, we develop a kernel based estimator of node wise counterfactual means that smooths over egocentric features of the inferred latent interactions rather than relying on high dimensional joint propensity scores. The method generalizes recent doubly robust estimators for network interference, remains valid under diffuse or global interference, and accommodates settings where the observed network provides only partial information about underlying interactions. Simulations with noisy network measurements show that the proposed approach substantially improves accuracy in estimating node wise counterfactual means. This framework offers a practical path to causal inference when interference extends beyond the observed graph.

Causal Inference in Networks

Causal Inference under Interference without Network Information: A Partial Identification Approach JungHo Lee* JungHo Lee, Edward Kennedy, David Choi,

In many applications, especially the in social and epidemiological sciences, interference is likely present but the underlying network is unobserved. While causal effects can sometimes be estimated in randomized experiments without network information, observational studies are more challenging because treatment assignment may depend on latent neighborhood context. We propose a partial identification approach for direct effects in observational studies under network interference. Our method assumes that unobserved network context can change the odds of treatment assignment by at most $Gamma$, yielding estimable bounds that depend only on observable conditional means and variances. We also describe efficient estimation of these bounds using modern nonparametric regression tools.

Causal Inference in Networks

Hypothesis Testing for Detecting Network Interference in Randomized Experiments
Christopher Harshaw* Christopher Harshaw, Fredrik Sävje, Chao Gao, Yitan Wang,

The randomized experiment is a widely used methodological tool for estimating causal effects. In order to meaningfully estimate causal effects, experimenters have to assume structure on the potential outcome functions. The simplest assumption is that of no-interference, although a growing body of work has provided various network interference models under which causal effects can be estimated. Still, this raises the question: to what extent can we test these assumptions? For example, can we detect interference when it is present in the experiment? Can we test between stronger and weaker models of network interference? To this end, a line of previous work has focused on controlling Type I error, but Type II error of these proposed methods remains less well understood.

In this work, we further investigate hypothesis testing frameworks for detecting interference. Our main contribution is a series of negative results. Roughly speaking, we show that there is no test for detecting interference which has small Type I and Type II error, even as the sample size increases. This finding has serious consequences for the epistemological nature of the no-interference and various network interference assumptions: namely, that they are, in a rigorous sense, untestable. On the positive side, we show that there exists a uniformly consistent test of no interference when the alternative hypotheses are restricted to a linear-in-means type model and the graph is sufficiently sparse.

Causal Inference in Networks

A General Exposure-Mapping-Agnostic Framework for Causal Inference under Two-Stage Randomization Yihui He* Yihui He, Eric Tchetgen Tchetgen,

We develop a general framework for causal inference under interference in cluster experiments conducted via two-stage randomization on a network of interconnected units, without relying on exposure mapping assumptions, restrictive modeling of cross-cluster interference, or Bernoulli treatment assignments. Within this framework, we establish a complete characterization of emph{linear weighted estimators} (LW) as defined by Godambe (1955, JRSS-B) that achieve identification of various network causal effects under interference. This general class includes several new estimators with improved theoretical guarantees and superior finite-sample performance relative to existing approaches such as standard inverse-probability-of-treatment weighting. For all estimators in this class, we establish central limit theorems and conservative variance estimators, which allows us to describe the distinct asymptotic behavior exhibited by different weighting schemes potentially of interest. In particular, we study how cluster-level assignment mechanisms affect estimator asymptotics and identify a subclass of cluster-agnostic LW estimators whose convergence rates are independent of the number of clusters and attain the optimal root-N rate, where N denotes the total number of units. We complement our theoretical results with extensive simulation studies that offer practical guidance on the choice of weighting method and cluster size under a wide range of interference structures.

Causal Inference in Networks

Score-Matching and Diffusion Enable Scalable Learning of Causal Feedback Dynamics Under Full Interference Yufeng Wu* Yufeng Wu, Abhinav Kumar, Josephine Yates, Caroline Uhler,

Feedback dynamics are central to network systems, including belief propagation among friends and cell-cell interaction in tissues. Modeling them causally requires representations that accommodate both directed causal effects and symmetric dependencies induced by feedback. Recent works adopt Chain Graphs, an extension of DAGs that contain both directed and undirected edges, as a principled framework for this purpose. In our work, we focus on the full interference setting, where we observe a single network realization and wish to learn parameters of the causal system. Existing approaches for learning chain graphs under full interference typically rely on strong parametric assumptions, most notably that each node's conditional distribution given its neighbors belongs to a specified exponential family (Tchetgen Tchetgen, 2021). Such assumptions are vulnerable to misspecification, which can lead to biased causal effect estimates and is difficult to diagnose. We propose a flexible, non-parametric alternative that replaces these assumptions with neural-based score models. Our approach combines denoising score matching and Langevin diffusion, enabling scalable training and more flexible distributional modeling, while being less sensitive to misspecification. We prove that our algorithm consistently learns the chain graph as the network size grows. Empirically, we validate our method on simulated networks and demonstrate its applicability on a biological spatial perturbation dataset.

Design of Experiments

Evaluating Algorithm-assisted Human Decision-making Over Repeated Algorithm Exposure: Recommendations for Effect Estimands and Experimental Design Maggie Wang* Maggie Wang, Michael Baiocchi,

In algorithm-assisted decision-making, an algorithmic decision support tool provides a recommendation, but the human ultimately makes the decision. Determining whether algorithm assistance actually improves the quality of human decision-making is critical, and randomized experiments are one way to collect robust evidence. Historically, however, experimental designs and analyses ignore how decision-making behavior adapts with repeated algorithm exposure. In this work, we first demonstrate how using a per-decision randomized design and estimating an average effect across all decisions results in misleading effect estimates under three canonical forms of adaptation: gradual overreliance on the algorithm, gradual ignoring of the algorithm, and gradual learning of where the algorithm is right/wrong. Instead, we propose using a staggered rollout design to target three alternative estimands: the global effect, which compares outcomes when decision-makers have sustained algorithm assistance versus never being assisted; the habituation effect, which compares outcomes under sustained assistance versus first-time-assistance; and the immediate effect, which compares outcomes under first-time-assistance versus never-assisted. Finally, drawing on the concept of local average treatment effects, we show that we can identify habituated and immediate effects specifically for decisions where the decision-maker is persuaded by the algorithm to change their decision.

Design of Experiments

Longitudinal Adaptive Design for Efficient Learning of Causal Estimands: Application to Multiple Time-to-Treatment-Initiation Effects Wenxin Zhang* Wenxin Zhang, Wenxin Zhang,

Adaptive designs are increasingly used in clinical trials and digital experiments to improve estimation efficiency by updating randomization probabilities as data accumulate. While most work focuses on single-stage settings, adaptive designs for longitudinal settings with multi-stage, time-varying treatments remain under-explored. We develop a general semiparametric efficiency framework for longitudinal adaptive experiments to optimize estimation efficiency of causal estimands. An efficiency-oriented design criterion is introduced to accommodate both single-estimand targets and joint optimization across multiple estimands. We show that optimal randomization at earlier stages depends on later-stage allocations, yielding a backward-recursive strategy to obtain the oracle design. A sequential adaptive design is proposed to learn and target the oracle using accumulating data. We apply the framework to time-to-treatment-initiation effects that compare initiating treatment at different stages versus delaying to subsequent stages. We show that designs optimized for a single stage-specific effect can compromise efficiency of other effects, revealing intrinsic trade-offs in longitudinal randomization. These challenges are addressed by the proposed adaptive design. We also provide a semiparametric efficient estimation framework for post-adaptive-experiment inference. Simulations show substantial variance reduction relative to non-adaptive designs, with performance close to the oracle.

Design of Experiments

Rerandomized Saturation Designs Alessio Frosini* Alessio Frosini, Tiziano Arduini, Peng Ding, Laura Forastiere,

Interference arises when a unit’s outcome depends on other units’ treatments. Randomized saturation designs, which randomly assign clusters to saturations and individual treatments within clusters according to the assigned saturation, are used to estimate direct, indirect, total and overall effects under partial interference. To improve covariate balance, we propose rerandomized saturation designs that rerandomize saturations and individual assignments until prespecified balance criteria are met. Under randomization inference with stratified interference and without distributional or modeling assumptions on covariates or outcomes, we derive the efficiency gains as well as the joint asymptotic sampling distribution of standard weighting estimators. Rerandomization yields a symmetric, unimodal limiting distribution that is more concentrated around the true effects and supports conservative inference. We further analyze a stepwise protocol that rerandomizes saturations to balance cluster covariates and, conditional on acceptance, rerandomizes individual treatments to balance individual covariates. We compare the joint and the stepwise strategies and provide a decision-theoretic framework for choosing stage-specific acceptance thresholds under a fixed overall acceptance probability.

Design of Experiments

Benefits and Costs of Adaptive Sampling Yu-Shiou Willy Lin* Yu-Shiou Willy Lin, Dae Woong Ham, Iavor Bojinov,

Multi-armed bandits are widely used for sequential experimentation in clinical trials, online platforms, etc. While regret minimization and valid inference from adaptively collected data are well studied in isolation, a basic question remains: when does adaptivity improve estimation precision relative to uniform designs, and how should inference be traded against the short-run cost of experimentation?

We first study arm-level mean estimation under mean-squared-error (MSE) objectives. We characterize when adaptive Neyman allocation yields strict MSE improvements over uniform sampling, and provide finite-sample guarantees for the outperformance. In heterogeneous-variance settings, the improvement holds at modest sample sizes, clarifying when adaptivity is not only asymptotically beneficial but practically relevant.

To reflect operational constraints, we introduce a second objective—risk minimization—capturing short-term performance loss during experimentation. We propose pi-rho allocation that interpolates between inference and risk-optimal policies, enabling an explicit design knob for the inference–risk trade-off. We establish high-probability finite-sample guarantees that quantify its loss behavior and show that pi-rho converges to the complete-information benchmark at the optimal rate as the sampling budget grows. Simulations support the theory by demonstrating improved precision over uniform allocation while controlling performance loss.

Design of Experiments

Where to Experiment? Site Selection Under Distribution Shift via Optimal Transport and Wasserstein DRO Adam Bouyamourn* Adam Bouyamourn,

How should researchers select experimental sites when the deployment population may differ from observed data? I formulate the problem of experimental site selection as an textit{optimal transport problem}, developing methods to minimize downstream estimation error by choosing sites that minimize the Wasserstein distance between population and sample covariate distributions. I develop new theoretical upper bounds on PATE and CATE estimation errors, and show that these different objectives lead to different site selection strategies. I extend this approach by using Wasserstein Distributionally Robust Optimization to develop a site selection procedure robust to adversarial perturbations of covariate information: a specific model of distribution shift. I also propose a novel data-driven procedure for selecting the uncertainty radius the Wasserstein DRO problem, which allows the user to benchmark robustness levels against observed variation in their data. Simulation evidence, and a reanalysis of a randomized microcredit experiment in Morocco (Cr'epon et al.), show that these methods outperform random and stratified sampling of sites when covariates have prognostic $R^2 > .5$, and alternative optimization methods i) for moderate-to-large size problem instances ii) when covariates are moderately informative about treatment effects, and iii) under induced distribution shift.

Sensitivity Analysis

Partial Identification of Causal Effects for Endogenous Continuous Treatments Abhinandan Dalal* Abhinandan Dalal, Eric Tchetgen Tchetgen,

Sensitivity analysis is often employed to assess the robustness of causal conclusions to unmeasured confounding, but existing methods are predominantly designed for binary treatments. In this paper, we provide natural extensions of two extensively used sensitivity frameworks – the Rosenbaum and Marginal sensitivity models – to the setting of continuous exposures. Our generalization replaces scalar sensitivity parameters with sensitivity functions that vary with exposure level, enabling richer modeling and sharper identification bounds. We develop a unified pseudo-outcome regression formulation for bounding the counterfactual dose–response curve under both models, and propose corresponding nonparametric estimators which have second order bias, accommodate modern machine learning methods for nuisance estimation, and are shown to achieve L2-consistency, asymptotic normality and minimax rates of convergence under suitable conditions. We also offer a geometric interpretation that relates the Rosenbaum and Marginal sensitivity model and guides their practical usage in global versus targeted sensitivity analysis. The methods are validated through simulations and a real-data application on the effect of second-hand smoke exposure on blood lead levels in children.

Sensitivity Analysis

Omitted Variable Bias in Difference-in-Differences Designs Juejue Wang* Juejue Wang, Carlos Cinelli, Pedro H. C. Sant'Anna, Victor Chernozhukov,

We study the omitted-variable bias (OVB) problem in canonical difference-in-differences (DiD) designs when unobserved confounding induces departures from the parallel trends assumption. Our results provide a novel characterization of the OVB formula for the average treatment effect on the treated (ATT), which may be of independent interest. We show how the ATT bias is governed by the strength of confounding in the treatment-selection mechanism and provide alternative ways of quantifying this strength, such as (i) changes in the average odds of treatment among the treated, (ii) confounding imbalance between treated and control units, or (iii) variation explained in treatment odds among the untreated. We additionally consider DiD designs using linear regressions with two-way fixed effects and show how the OVB simplifies in such settings. Building on these results, we offer sensitivity statistics for routine reporting describing the minimum strength of confounding required to overturn the conclusions of a DiD study, as well as formal bounds on the strength of confounders based on comparisons to observed covariates. We demonstrate the utility of our approach in two empirical examples.

Partial Identification / Unmeasured Confounding

Sharp partial proximal inference: an assumption-lean approach for leveraging negative controls Alexander Levis* Alexander Levis,

Negative controls–or proxies–are variables assumed or known not to be involved in certain causal pathways, in the context of an exposure-outcome pair of interest. Proxies have historically been used to as bias detection agents, whereby a non-null result when replacing the exposure or outcome with a corresponding proxy may indicate the presence of unmeasured confounding. Recently, proximal causal inference has emerged as a promising framework for using these variables to directly identify a causal relationship of interest, even in the face of strong unmeasured confounding. Proximal methods, however, rely on untestable, often outright opaque identifying assumptions involving so-called “bridge function” or “completeness” conditions. In this work, we relax these assumptions in a commonly adopted single outcome proxy setting. We show that the defining causal assumptions satisfied by the negative control non-trivially restrict the counterfactual outcome distribution. Moreover, we derive nonparametric, robust, efficient estimators of sharp bounds for mean counterfactuals. These bounds are highly non-smooth, non-closed-form solutions to linear programs involving various potentially high-dimensional nuisance functions; our statistical approach has implications for a wide class of such challenging functionals. Practically, the proposed methodology can be used to leverage proxies for causal and missing data problems, achieving sharp, valid inference under transparent assumptions.

Sensitivity Analysis

Exploiting independence constraints for efficient estimation of bounds on causal effects in the presence of unmeasured confounding Ting-Hsuan Chang* Ting-Hsuan Chang, Eric Tchetgen Tchetgen, Ilya Shpitser, Daniel Malinsky,

Causal graphs, such as directed acyclic graphs (DAGs) and partial ancestral graphs (PAGs), can inform covariate adjustment for estimating causal effects and improve estimation efficiency by exploiting the graphical structure. In many applications, however, the target causal parameter may not be point-identified due to the presence of unmeasured cofounding. Sensitivity analysis methods address this challenge by characterizing bounds on the causal parameter under varying assumptions about the magnitude or form of unmeasured confounding. We focus on semiparametric efficient estimation of causal effects in non-identifiable settings, assuming a known (or hypothesized) causal graph. We propose an influence function projection approach that exploits the (conditional) independence constraints implied by the causal graph to improve the efficiency of semiparametric estimators of upper and lower bounds on the average causal effect within a given sensitivity analysis framework. We show that our approach applies across sensitivity analysis frameworks and causal estimands in general, thereby connecting knowledge of graphical structure with the sensitivity analysis literature. We illustrate our approach through examples thought to be affected by unmeasured confounding, including the effect of labor training program on post-intervention earnings, and the effect of low ejection fraction on heart failure death.

Causal Fairness, and Bias/Discrimination

Monitoring Racial Bias in Police Traffic Enforcement with Imperfect Proxies of Driver Behavior Dean Knox* Dean Knox, Kai Cooper, Gregory Lanzalotto, Jacob Kaplan, Haosen Ge, Jonathan Mummolo,

Decades of work has sought to assess whether police traffic enforcement is discriminatory by comparing the racial composition of those who are stopped to some external benchmark, such as the composition of local residents in a police jurisdiction. Such “benchmark analyses” have been criticized by statisticians and law-enforcement officials alike for yielding misleading results that fail to accurately represent the population at risk of police traffic stops—namely, individuals actually engaged in dangerous driving behavior. In this work, we present a causal framework for benchmark analysis that clarifies the distinction between valid and invalid benchmarks. We use this framework to formalize implicit assumptions in prior work and demonstrate how these assumptions can be empirically falsified. By drawing connections between benchmark analysis and a recent literature on negative control outcomes, we derive new estimators and partial identification results, with extensions to address a host of common statistical challenges that arise in benchmark analyses. Among other challenges, we consider scenarios where (1) false positives induce distortions in benchmark composition; (2) civilian race cannot be directly observed and must be inferred from surnames on citations; and (3) officer decisions lead drivers to modify the behavior that is captured by benchmarks. We demonstrate the generality of the approach with applications across multiple jurisdictions and benchmark data sources.

Dynamic Treatment Regimes

A Covariate Balancing Approach for Dynamic Treatment Regimes Yige Li* Yige Li, Benjamin Buzzee, Marcela Horvitz-Lennon, Sharon-Lise Normand,

Typical methods for identifying optimal dynamic treatment regimes (DTRs), such as Q-learning and outcome-weighted learning, often perform suboptimally in chronic disease settings with long observation periods and sparse data. Small sample sizes and uncertainty make it challenging to determine optimal treatment strategies and estimate individual effects. We find that limited covariate overlap and accumulating imbalances in treatment assignment over time frequently cause instability, preventing convergence to the optimal policy. To address this, we propose a covariate balancing approach that constructs a set of synthetic controls for identifying optimal treatment rules and predicting optimal DTRs in the testing set, without explicitly modeling the outcome or treatment mechanism. Using the pessimism principle in treatment selection, our method reduces decision-making driven by uncertainty, leading to improved stability and efficiency compared to standard approaches. Numerical studies show that balancing learning outperforms standard methods under conditions of time-varying imbalance or data sparsity. We illustrate our approach using a cohort of Medicare beneficiaries with schizophrenia who initiated Clozapine and were followed for up to 18 months, demonstrating optimal prescription-changing strategies for two outcomes: acute psychiatric events and medication adherence. This work is funded by Grant R01-MH130213 from the National Institute of Mental Health.

Matching, Weighting

Forest Kernel Balancing Weights: Outcome-Guided Features for Causal Inference Andy Shen* Andy Shen, Eli Ben-Michael, Avi Feller, Luke Keele, Jared Murray,

While balancing covariates between groups is central for observational causal inference, selecting which features to balance remains a challenging problem. Kernel balancing is a promising approach that first estimates a kernel that captures similarity across units and then balances a (possibly low-dimensional) summary of that kernel, indirectly learning important features to balance. In this paper, we propose forest kernel balancing, which leverages the underappreciated fact that tree-based machine learning models, namely random forests and Bayesian additive regression trees (BART), implicitly estimate a kernel based on the co-occurrence of observations in the same terminal leaf node. Thus, even though the resulting kernel is solely a function of baseline features, the selected nonlinearities and other interactions are important for predicting the outcome—and therefore are important for addressing confounding. Through simulations and applied illustrations, we show that forest kernel balancing leads to meaningful computational and statistical improvement relative to standard kernel methods, which do not incorporate outcome information when learning features.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Prognostic score matching in difference-in-differences studies Yunshu Zhang* Yunshu Zhang, Dylan Small, Ting Ye, Laura Hatfield,

The parallel trends assumption (PTA) plays a central role in difference-in-differences (DID) studies, yet it is often violated when an arbitrary control group is used. A common remedy is to apply matching to select a subset of units for which the PTA is more plausible, typically by matching on covariates or the propensity score. However, practical guidance on how to implement matching in DID settings remains limited.

In this paper, we first characterize the set of variables that must be balanced for the PTA to hold. We define these variables as DID confounders, namely covariates that jointly affect treatment assignment and the trend of the potential outcome under control. These confounders may differ from those defined in cross-sectional studies, although simulation results suggest that balancing both types of confounders can further reduce bias. To address the curse of dimensionality, we propose a novel prognostic score matching approach for DID. The prognostic score is defined as a sufficient statistic for the trend of the potential outcome under control. Matching on this prognostic score makes the PTA more plausible. Simulation studies demonstrate that prognostic score matching yields smaller bias and variance in estimating causal effects compared with existing matching approaches.

Matching, Weighting

Bias Mitigation in Matched Observational Studies with Continuous Treatments: Calipered Non-Bipartite Matching and Bias-Corrected Estimation and Inference Anthony Frazier*
Anthony Frazier, Siyu Heng, Wen Zhou,

In matched observational studies with continuous treatments, individuals with different treatment doses but the same or similar covariate values are paired for causal inference. While inexact covariate matching (i.e., covariate imbalance after matching) is common in practice, previous matched studies with continuous treatments have often overlooked this issue as long as post-matching covariate balance meets certain criteria. Through re-analyzing a matched observational study on the social distancing effect on COVID-19 case counts, we show that this routine practice can introduce severe bias for causal inference. Motivated by this finding, we propose a general framework for mitigating bias due to inexact matching in matched observational studies with continuous treatments, covering the matching, estimation, and inference stages. In the matching stage, we propose a carefully designed caliper that incorporates both covariate and treatment dose information to improve matching for downstream treatment effect estimation and inference. For the estimation and inference, we introduce a bias-corrected Neyman estimator paired with a corresponding bias-corrected variance estimator. The effectiveness of our proposed framework is demonstrated through a re-analysis of the aforementioned observational study on the effect of social distancing on COVID-19 case counts.

Matching, Weighting

Distributional Balancing for Causal Inference: A Unified Framework via Characteristic Function Distance Diptanil Santra* Diptanil Santra, Chan Park, Guanhua Chen,

Weighting methods are essential tools for estimating causal effects in observational studies, with the goal of balancing pre-treatment covariates across treatment groups. Traditional approaches rely on propensity score modeling or matching a finite number of covariate moments, and therefore do not guarantee balance of the full covariate distributions. Recently developed distributional balancing methods overcome this limitation by directly aligning entire covariate distributions, but existing approaches lack a unified framework, formal theoretical guarantees, and valid inferential procedures. In this paper, we introduce a novel, unified framework for nonparametric distributional balancing based on the Characteristic Function Distance (CFD). We show that several commonly used discrepancy measures, including maximum mean discrepancy and energy distance, are special cases of the CFD. Our theoretical analysis establishes conditions under which the resulting weighting estimator achieves root-n-consistency. Since the standard bootstrap may fail for this estimator, we propose subsampling as a valid alternative for inference. We further extend our approach to an instrumental variable setting to address potential unmeasured confounding. Finally, we illustrate the performance of our method through simulation studies and a real-world application, where the CFD-based weighting estimator exhibits results consistent with our theoretical predictions.

Design-Based Causal Inference

Towards the most powerful randomization tests Zijun Gao* Zijun Gao,

The validity of randomization tests for sharp null hypotheses is well understood, whereas their power properties remain less explored. In this paper, we study three notions of most powerful randomization tests: (1) for alternatives that specify the full potential-outcome schedule, we construct the most powerful test statistic that typically achieves full power; (2) for alternatives in which units follow a super-population model, the most powerful test statistic corresponds to the log-likelihood of the observed outcome conditional on the assigned treatment and covariates; (3) for the super-population alternatives in (2), a studentized AIPW-type statistic yields the most powerful randomization test among those that are also asymptotically valid for the zero average treatment effect null.

Guided by the power properties, we propose estimating a super-population model and adopting the associated most powerful test statistic in (2) or (3). Learning super-population models typically involves realized treatment assignments, and sample splitting is required to preserve the validity of the subsequent randomization test. In contrast, we construct an estimator that does not use realized assignments by treating the treatment as a latent variable and recovering the model using the EM algorithm. When applied to simulated and real data, our procedure is often more powerful than both the non-adaptive and the sample-splitting–based adaptive counterparts.

Dynamic Treatment Regimes

Sigmoid-FTRL: Design-Based Adaptive Neyman Allocation for AIPW Estimators Fangyi Chen* Fangyi Chen, Shu Ge, Jian Qian, Christopher Harshaw,

We consider the problem of Adaptive Neyman Allocation for the class of AIPW estimators in a design-based setting, where potential outcomes and covariates are deterministic. As each subject arrives, an adaptive procedure must select both a treatment assignment probability and a linear predictor to be used in the AIPW estimator. Our goal is to construct an adaptive procedure that minimizes the Neyman Regret, which is the difference between the variance of the adaptive procedure and an oracle variance which uses the optimal non-adaptive choice of assignment probability and linear predictors. While previous work has drawn insightful connections between Neyman Regret and online convex optimization for the Horvitz-Thompson estimator, one of the central challenges for AIPW estimator is that the underlying optimization is non-convex. In this paper, we propose Sigmoid-FTRL, an adaptive experimental design which addresses the non-convexity via simultaneous minimization of two convex regrets. We prove that under standard regularity conditions, the Neyman Regret of Sigmoid-FTRL converges at a $T^{-1/2}R$ rate, where T is the number of subjects in the experiment and R is the maximum norm of covariate vectors. We further prove that this rate is minimax optimal. Finally, we establish a central limit theorem and a consistently conservative variance estimator which facilitate the construction of asymptotically valid Wald-type confidence intervals.

Randomized Designs and Analyses

Randomization inference for stepped-wedge designs with noncompliance with application to a palliative care pragmatic trial Jeffrey Zhang* Jeffrey Zhang, Zhe Chen, Katherine Courtright, Scott Halpern, Michael Harhay, Dylan Small, Fan Li,

While palliative care is increasingly commonly delivered to hospitalized patients with serious illnesses, few studies have estimated its causal effects. Courtright et al. (2016) adopted a cluster-randomized stepped-wedge design to assess the effect of palliative care on a patient-centered outcome. The randomized intervention was a nudge to administer palliative care butdid not guarantee receipt of palliative care, resulting in noncompliance (compliance rate ≈ 30%). A subsequent analysis using methods suited for standard trial designs produced statistically anomalous results, as an intention-to-treat analysis found no effect while an instrumental variable analysis did (Courtright et al., 2024). This highlights the need for a more principled approach to address noncompliance in stepped-wedge designs. We provide a formal causal inference framework for the stepped-wedge design with noncompliance by introducing a relevant causal estimand and corresponding estimators and inferential procedures. Through simulation, we compare an array of estimators across a range of stepped-wedge designs and provide practical guidance in choosing an analysis method. Finally, we apply our recommended methods to reanalyze the trial of Courtright et al. (2016), producing point estimates suggesting a larger effect than the original analysis of (Courtright et al., 2024), but
intervals that did not reach statistical significance.

Design-Based Causal Inference

Design-Based Inference for Attribution to Causal Interaction Zion Lee* Zion Lee, Kwonsang Lee,

Understanding causal interaction between multiple treatments is central to many scientific questions, yet standard practice typically relies on regression-based interaction terms whose interpretation depends on modeling assumptions. We propose a design-based framework for attributing outcomes specifically to causal interaction, without imposing sharp null hypotheses or outcome regression models.
Our estimand, the interaction-attributable effect, counts the number of jointly treated units whose observed outcomes are attributable to the interaction between treatments, extending Rosenbaum's attributable-effect perspective. We develop novel randomization-based tests that account for uncertainty arising from unobserved potential outcomes.
The proposed tests combine multiple contingency-table–based statistics using Wald-type procedures with analytically derived covariance structures, yielding valid inference under complete randomization and matched designs. Simulation studies demonstrate that our approach achieves higher power than conventional regression-based interaction tests. Our framework provides a principled design-based alternative for causal interaction analysis in studies with binary outcomes.

Causal Inference and SUTVA/Consistencies Violations

Unconditional Randomization Tests for Interference Liang Zhong* Liang Zhong,

Researchers are often interested in the existence and extent of interference between units when conducting causal inference or designing policy. However, testing for interference presents significant econometric challenges, particularly due to complex clustering patterns and dependencies that can invalidate standard methods. This paper introduces the pairwise imputation-based randomization test (PIRT), a general and robust framework for assessing the existence and extent of interference in experimental settings. PIRT employs unconditional randomization testing and pairwise comparisons, enabling straightforward implementation and ensuring finite-sample validity under minimal assumptions about network structure. The method's practical value is demonstrated through an application to a large-scale policing experiment in Bogota, Colombia (Blattman et al., 2021), which evaluates the effects of hotspot policing on crime at the street segment level. The analysis reveals that increased police patrolling in hotspots significantly displaces violent crime, but not property crime. Simulations calibrated to this context further underscore the power and robustness of PIRT.

Heterogeneous Treatment Effects

Debiased Front-Door Learners for Heterogeneous Effects Yonghan Jung* Yonghan Jung,

In observational settings where treatment and outcome share unmeasured confounders but an observed mediator remains unconfounded, the front-door (FD) adjustment identifies causal effects through the mediator. We study the textit{heterogeneous treatment effect} (HTE) under FD identification and introduce two debiased learners: FD-DR-Learner and FD-R-Learner. Both attain fast, quasi-oracle rates (i.e., performance comparable to an oracle that knows the nuisances) even when nuisance functions converge as slowly as n^{-1/4}. We provide error analyses establishing debiasedness and demonstrate robust empirical performance in synthetic studies and a real-world case study of primary seat-belt laws using Fatality Analysis Reporting System (FARS) dataset. Together, these results indicate that the proposed learners deliver reliable and sample-efficient HTE estimates in FD scenarios. The implementation is available at https://github.com/yonghanjung/FD-CATE.

Heterogeneous Treatment Effects

Heterogeneous Treatment Effects with Heterogeneous Treatment Exposure Hui Lan* Hui Lan, Vasilis Syrgkanis,

We study heterogeneous treatment effect estimation in settings where treatment assignment is binary but the realized treatment among treated units is heterogeneous, such as differences in intensity, composition, or even high dimension representations. In these environments, the untreated state does not generally correspond to any particular value of the treatment variable, making standard scalar or dose–response formulations ill-suited. We motivate a causal estimand that captures how treatment effects vary both across covariates and along dimensions of treatment heterogeneity by defining effects as the difference between potential outcomes under counterfactual treatment realizations and the untreated potential outcome. We propose a doubly robust estimator for this estimand that accommodates flexible modeling of outcome and treatment propensity, and show that the proposed estimator is insensitive to nuisance estimation errors.

Heterogeneous Treatment Effects

Assumption-Lean Differential Variance Inference for Heterogeneous Treatment Effect Detection Philippe Boileau* Philippe Boileau, Philippe Boileau, Hani Zaki, Gabriele Lileikyte, Niklas Nielsen, Patrick Lawler, Mireille Schnitzer,

The conditional average treatment effect (CATE) is frequently estimated to refute the homogeneous treatment effect assumption. Under this assumption, all units making up the population under study experience identical benefit from a given treatment. Uncovering heterogeneous treatment effects through inference about the CATE, however, requires that covariates truly modifying the treatment effect be reliably collected at baseline. CATE-based techniques will necessarily fail to detect violations when effect modifiers are omitted from the data due to, for example, resource constraints. Severe measurement error has a similar impact. To address these limitations, we prove that the homogeneous treatment effect assumption can be gauged through inference about contrasts of the potential outcomes' variances. We derive causal machine learning estimators of these contrasts and study their asymptotic properties. We establish that these estimators are doubly robust and asymptotically linear under mild conditions, permitting formal hypothesis testing about the homogeneous treatment effect assumption even when effect modifiers are missing or mismeasured. Numerical experiments demonstrate that these estimators' asymptotic guarantees are approximately achieved in experimental and observational data alike. These inference procedures are then used to detect heterogeneous treatment effects in the re-analysis of randomized controlled trials investigating targeted temperature management in cardiac

Heterogeneous Treatment Effects

On Non-Parametric Regression with Estimated Covariates Jiaqi Wu* Jiaqi Wu, Matteo Bonvini, Edward Kennedy, Jennie Brand, Yu Xie,

Motivated by the study of heterogeneous returns to education in Brand and Xie (2010), which considers how the effect of completing college on unemployment varies with the (unknown) probability of completing college, we analyze the problem of estimating a nonparametric regression function when certain covariates are estimated in a first step. Naive plug-in estimators that treat the estimated covariates as known generally suffer from first-stage estimation error. To mitigate this issue, we analyze two debiasing approaches within a framework that is agnostic to the choice of first-stage estimation method and consider both local- and sieves-based methods for the second-stage regression. The methods considered are: (i) influence function-based estimators of pathwise differentiable parameters that approximate the target estimand, (ii) a variant of plug-in estimators that direct aim to correct the bias. For each method, we upper bound the estimation risk and characterize conditions under which oracle rates can be approached, highlighting the possible gains in terms of convergence rates relative to the plug-in. Simulation results confirm the theoretical findings. We apply our methodology to data from the National Longitudinal Survey of Youth 1997 and find evidence that completing college yields the largest benefits for individuals least likely to complete college, consistent with earlier findings in the literature (Brand and Xie, 2010; Brand, 2019).

Applicants in Social Sciences

A Win Ratio Approach to Nonparametric, Scale-Free Measure of Causal Effects on Validated Latent Traits Beom Kwon* Beom Kwon, Hyunseung Kang,

Many studies in education, political science, and psychology aim to estimate treatment effects on validated latent traits, such as ability, attitudes, and disease severity, measured through multiple item responses. Most existing approaches are fully parametric, requiring correct specification of both the latent-trait measurement model and the structural model for causal effects. These approaches can lead to biased estimates when either model is mis-specified. Worse, the causal estimand may depend on the scale of the latent trait, which often lacks an intrinsic scale. In this work, we propose a win ratio approach for measuring treatment effects on validated latent traits. We show that the win ratio is nonparametric, agnostic to the underlying measurement model, and scale-free. We also establish necessary and sufficient conditions under which it can be related to existing causal effect measures. For estimation, we apply an influence function based estimator that accommodates flexible machine-learning methods for nuisance functions without requiring a correctly specified item-response model. Under mild assumptions, the estimator is doubly robust, asymptotically normal, and semiparametrically efficient. Extensive simulation studies under various measurement models, including Rasch, 2PL, 3PL, and graded response models, along with a real data analysis of the effect of game-based learning on math ability, demonstrate that our approach is more robust and stable than existing method.

Applicants in Social Sciences

Causal Approach for Careless Responding Jiwoo Kim* Jiwoo Kim, Felix Thoemmes,

Careless responding has long been recognized as a threat to data quality, and methods have been developed to detect and remove potentially careless responses. However, existing approaches rarely incorporate a formal causal perspective, limiting their ability to account for the data-generating mechanisms underlying careless responses. When causal assumptions are made, they are often adopted unintentionally and may be unrealistically strong, making it implausible to identify true careless responses in practice. In addition, current definitions of careless responding do not align well with the methods used to detect it, as detection approaches classify responses based on observable patterns without linkage to the underlying causal process, creating conceptual inconsistencies. In this study, we redefine careless responding using a causal framework and reevaluate prior definitions in light of this definition. We examine how different causal mechanisms of careless responding influence bias and identifiability, using simulations that reflect distinct causal structures generating careless responses. We also propose criteria for determining when existing careless-responding methods should—or should not—be applied. By introducing a causal definition of careless responding and demonstrating its implications, this study offers a new perspective on the problem and provides a foundation for developing more principled approaches to handling careless responding in empirical research.

Applicants in Social Sciences

Scalable Causal Inference in Marketing Mix Modeling: An Automated Double Machine Learning Pipeline with Continuous Treatments and Domain-Informed Constraints Hirotoshi Nakahara* Hirotoshi Nakahara, Martin Spindler,

In marketing, Marketing Mix Modeling (MMM) is essential for privacy-compliant media measurement. However, traditional approaches often suffer from functional misspecification and subjective tuning. This study presents an automated causal inference pipeline for MMM, leveraging Double Machine Learning (DML) to estimate the impact of continuous media treatments with Neyman orthogonality to mitigate regularization bias.

Our framework introduces an AutoML-based selection process for nuisance parameters (outcome and treatment models), exploring various algorithms such as XGBoost, LightGBM, and Random Forest. To identify the optimal DML architecture, we utilize a “combined loss” metric, defined as the product of the treatment model RMSE and the sum of the treatment and outcome model RMSEs. This criterion prioritizes models that achieve high-quality identification of the treatment assignment mechanism alongside predictive accuracy.

Furthermore, we incorporate domain-informed constraints: (1) a non-negativity filter on estimated causal effects to align with marketing priors, and (2) a stability-based selection that minimizes the total variance of Conditional Average Treatment Effect (CATE) estimates across iterations. Validated on semi-synthetic data with complex multi-collinearity and a real-world retail case study, our pipeline demonstrates superior accuracy and stability over conventional MMMs.

Applicants in Social Sciences

Uncovering Treatment Effect Heterogeneity in New York City's Gifted and Talented Program using BART Katherine Strickland* Katherine Strickland, Wei Li, Jennifer Hill,

This study examines the heterogeneous effects of New York City's Gifted and Talented (G&T) program on student achievement. Using administrative data from the 2010-2023 school years, we use Bayesian Additive Regression Trees (BART) to estimate treatment effects at the individual level, then aggregate these to examine heterogeneity across ethnic and socioeconomic groups. Our findings reveal small but consistent impacts of G&T participation on Grade 6-8 Mathematics and English Language Arts (ELA) performance across all student groups. We explore treatment heterogeneity by student demographics, program type, entry timing, program duration, cohort, and school district, finding effect sizes ranging from 0.09 to 0.28 standard deviations. These results suggest targeted expansion of G&T programs could potentially reduce achievement gaps while supporting high-achieving students from underrepresented groups.

Applicants in Social Sciences

Do Test Scores Help Teachers Give Better Track Advice to Students? A Principal Stratification Analysis Fabrizia Mealli* Fabrizia Mealli, Javier Viviens, Andrea Ichino,

We study whether access to standardized test scores improves the quality of teachers' secondary school track recommendations. We frame teachers' advice as a decision problem and evaluate efficiency and fairness by building on literature on algorithm-assisted human decisions and using a principal stratification framework. We target Average Principal Causal Effects (APCEs), which measure how access to test scores affects student graduation outcomes within latent strata defined by students' potential graduation under each teacher's decision. APCEs quantify the gains from improved targeting and the losses arising from misallocation. Approaches in the literature identify APCEs under strong assumptions (e.g., unconfoundedness, exclusion restriction). We extend this literature by identifying the APCEs without these assumptions. We also provide a comprehensive welfare evaluation that allows for differential weighting of short-term and long-term gains and losses. Finally, we study principal fairness, assessing whether access to test scores improves or worsens equity with respect to certain protected attributes. We find that providing test results to teachers induces fairer recommendations for immigrant and low-SES students; it also increases the share of students successfully placed in more demanding tracks, but misplaces some of the weaker students. However, only implausibly high weights on the short-term losses of weaker students would justify prohibiting test-score-based upgrades.

Applications in Health and Biology

Microbiome data integration via shared dictionary learning Shulei Wang* Shulei Wang, Bo Yuan,

Data integration is a powerful tool for facilitating a comprehensive and generalizable understanding of microbial communities and their association with outcomes of interest. However, integrating data sets from different studies remains a challenging problem because of severe batch effects, unobserved confounding variables, and high heterogeneity across data sets. We propose a new data integration method called MetaDICT, which initially estimates the batch effects by weighting methods in causal inference literature and then refines the estimation via novel shared dictionary learning. Compared with existing methods, MetaDICT can better avoid the overcorrection of batch effects and preserve biological variation when there exist unobserved confounding variables, data sets are highly heterogeneous across studies, or the batch is completely confounded with some covariates. Furthermore, MetaDICT can generate comparable embedding at both taxa and sample levels that can be used to unravel the hidden structure of the integrated data and improve the integrative analysis. Applications to synthetic and real microbiome data sets demonstrate the robustness and effectiveness of MetaDICT in integrative analysis. Using MetaDICT, we characterize microbial interaction, identify generalizable microbial signatures, and enhance the accuracy of outcome prediction in two real integrative studies, including an integrative analysis of colorectal cancer metagenomics studies and a meta-analysis of immun

Applications in Health and Biology

Environmental Noise Exposure and Annual Healthcare Cost and Utilization in Medicare Patients with Alzheimer's Disease and Related Dementias: A Retrospective Cohort Study

Tien Tran* Tien Tran, Shirley Huang, Edmund Seto, Zafar Zafari,

Environmental noise may accelerate cognitive decline and increase healthcare costs and utilization in individuals with Alzheimer's disease and related dementias (AD/ADRD). However, its impact on Medicare beneficiaries with AD/ADRD remains unclear. This study evaluates the association between chronic environmental noise exposure and healthcare costs and utilization (inpatient admissions and emergency department visits) in Medicare beneficiaries with AD/ADRD. We conducted a retrospective cohort study using a 25% random sample of Medicare fee-for-service beneficiaries (2017–2021). The sample included 1,064,686 individuals with AD/ADRD and continuous enrollment in Medicare Parts A and B. Chronic noise exposure was estimated at the ZIP-code level. Healthcare costs and utilization were measured annually. We used double negative controls method to address confounding: a negative control outcome (preceding-year utilization and costs) and a negative control exposure (subsequent-year noise exposure). Models were further adjusted for individual demographics, comorbidities, medication use, Medicare plan type, and county-level socioeconomic factors using the CDC Social Vulnerability Index.

Applications in Health and Biology

Beyond Proportional Hazards: Double Machine Learning of the Causal Average Hazard
Xiang Meng* Xiang Meng, Hajime Uno, Kenneth Kehl, Lu Tian,

The Cox model and its hazard ratio (HR) are standard for treatment effects, yet face limitations like non-collapsibility and sensitivity to censoring. This paper develops a semiparametric framework for the average hazard with survival weight (AH), a model-free, population-level person-time event rate. We contrast the AH with pairwise win ratios, assumption-lean regression projections, and fixed-horizon risks, highlighting its utility as a marginal, rate-based estimand robust to non-proportional hazards and invariant to independent censoring.

We contribute to causal inference in three ways. First, we formalize the AH as a causal estimand using potential outcomes, identifying it under standard assumptions while clarifying its distinction from marginalized conditional rates. Second, we establish the pathwise differentiability of the AH and derive its efficient influence function (EIF). Third, we propose a cross-fitted, doubly robust estimator leveraging machine learning for nuisance estimation while maintaining sqrt{n}-consistency and asymptotic normality. Simulations show our estimator maintains near-nominal coverage and minimal bias, even with crossing hazards. Finally, we apply our method to the SEER-Medicare database, a massive health record system that links official cancer registries with long-term insurance data to show how well different treatments work for thousands of patients in real-world settings.

Applications in Health and Biology

The Clone-Censor-Weight Method Applied to Continued Colorectal Cancer Screening Participation: A Target Trial Emulation Buket Öztürk Esen* Buket Öztürk Esen, Lars Pedersen,

Background: The effectiveness of the fecal immunochemical test (FIT) in reducing colorectal cancer (CRC) mortality is well-known. However, studies evaluating the effect of attending a second FIT screening remain limited.

Objective: By utilizing the target trial framework and the clone-censor-weight (CCW) method, we aim to evaluate the effect of attending a second FIT screening while mitigating immortal time bias and selection bias.

Method: We applied the CCW method in three steps: 1) Cloning: We created two copies of each eligible individual (N=622,265) and assigned each copy to one of two screening strategies: participating within three years of an initial negative result vs. not participating. 2) Censoring: At monthly intervals, we assessed adherence. If a copy deviated from the assigned strategy, it was censored. 3) Weighting: To adjust for selection bias introduced by censoring, we estimated stabilized inverse probability weights. This created a pseudo-population in which censoring was independent of baseline factors.

Result: At 6 years, the risk of CRC mortality was 0.07% (95%CI: 0.06%; 0.08%) for continued participation vs. 0.17% (95%CI: 0.14%; 0.21%) for non-participation, resulting in a relative risk of 0.39 (95%CI: 0.30; 0.52) and a risk difference of -0.10% (95%CI: -0.14%; -0.07%).

Conclusion: CCW is a useful method for evaluating the effectiveness of attending a second FIT screening; however, selection bias due to unmeasured factors cannot be ruled out.

Applications in Health and Biology

Disentangling Misreporting from Genuine Adaptation in Strategic Settings: A Causal Approach Dylan Zapzalka* Dylan Zapzalka, Trenton Chang, Lindsay Warrenburg, Sae-Hwan Park, Daniel Shenfeld, Ravi Parikh, Jenna Wiens, Maggie Makar,

In settings where ML models are used to inform the allocation of resources, agents affected by the allocation decisions might have an incentive to strategically change their features to secure better outcomes. While prior work has studied strategic responses broadly, disentangling misreporting from genuine adaptation remains a fundamental challenge. In this work, we propose a causally-motivated approach to identify and quantify how much an agent misreports on average by distinguishing deceptive changes in their features from genuine adaptation. Our key insight is that, unlike genuine adaptation, misreported features do not causally affect downstream variables (i.e., causal descendants). We exploit this asymmetry by comparing the causal effect of misreported features on their causal descendants as derived from manipulated datasets against those from unmanipulated datasets. We formally prove identifiability of the misreporting rate and characterize the variance of our estimator. We empirically validate our theoretical results using a semi-synthetic and real Medicare dataset with misreported data, demonstrating that our approach can be employed to identify misreporting in real-world scenarios.

Applications in Health and Biology

Causal Evaluation of Larvicide Intervention on West Nile Virus Vectors Jian Yang* Jian Yang, Trevor Harris, Chan Park, Kristina Lopez, Rebecca Smith,

Effective larval control is a cornerstone of Integrated Vector Management for West Nile virus mitigation. In 2021, the North Shore Mosquito Abatement District in the Chicago area implemented a strategic switch in larvicide products. Evaluating the causal impact of such operational interventions is challenging due to strong seasonality and meteorological confounders (e.g., temperature, precipitation). This study applies quasi-experimental frameworks—specifically Difference-in-Differences and Synthetic Control—to estimate the intervention's causal effect on Culex mosquito abundance and the West Nile Virus Vector Index using weekly surveillance data (2017–2023). While preliminary Interrupted Time Series models suggested an 11.6% reduction in abundance, they rely solely on temporal controls. To improve identification, we leverage data from neighboring districts as control units. We implement Difference-in-Difference and Synthetic Control methods to construct a data-driven counterfactual that minimizes pre-intervention trend divergence. This research demonstrates the utility of causal inference methods in validating public health interventions where randomized control trials are infeasible, offering robust evidence of larvicide efficacy while addressing complex environmental noise.

Applications in Health and Biology

Quantifying the Causal Impact of Political Polarization on COVID-19 Dynamics: A Longitudinal Proximal Causal Inference Approach Jian Yang* Jian Yang, Rebecca Smith,

The politicization of public health has been a defining feature of the COVID-19 pandemic. However, estimating the causal effect of political affiliation on infection rates is challenged by unmeasured, time-varying confounders, specifically community-level behavioral compliance and risk attitudes. Standard regression models often fail to account for these dynamic latent factors, leading to biased estimates. This study proposes a longitudinal Proximal Causal Inference framework to identify the causal effect of county-level political leaning on weekly COVID-19 case growth rates from 2021 to 2023. We address the challenge of unobserved confounding by employing a Two-Stage Least Squares strategy with distinct proxy variables. Crucially, to overcome the 2022 discontinuity of commercial mobility datasets, we leverage the Bureau of Transportation Statistics‘ data as a continuous, outcome-inducing proxy. Conversely, we utilize the historical presidential election results as a static, treatment-inducing proxy. Our model incorporates a specific lag structure and time fixed effects to control for incubation periods and temporal dependencies. By formally separating static political treatments from dynamic behavioral mechanisms, this research provides a robust identification strategy for evaluating how polarization structurally impacted pandemic trajectories over an extended period.

Applications in Health and Biology

Causal Modeling and Discovery of Brain Network Interactions Ang Li* Ang Li,

Understanding the causal organization of brain networks is a central problem in cognitive neuroscience, as disruptions in these networks are linked to disorders such as Parkinson’s disease, Alzheimer’s disease, and epilepsy. Describing how communication between brain regions breaks down can help explain disease-related symptoms and inform markers of progression and potential intervention strategies. More broadly, identifying causal relationships among brain regions is essential for understanding brain function and cognition.
Most analyses of functional brain imaging data rely on correlation-based measures, which cannot distinguish true causal influence from spurious associations. We address this limitation by applying the Structural Causal Model framework and Probabilities of Causation to large-scale fMRI data collected without an explicit task. This allows interventional and counterfactual reasoning about brain network interactions, including counterfactual questions related to disease-associated changes.
We also introduce an expert-guided causal discovery strategy that incorporates expert knowledge into discovery based on conditional independence tests. When statistical tests leave edge directions unresolved, plausible orientations are propagated through the graph and evaluated for contradictions with expert knowledge, allowing implausible causal structures to be ruled out.

Applications in Health and Biology

Provider-Level Heterogeneity in Inpatient Efficiency Under Resource Strain: A Multilevel Quasi-Experimental Analysis Eliot Weinstein* Eliot Weinstein, Matthew Cerasale,

Length of stay (LOS) is a central measure of inpatient efficiency, yet isolating provider-level contributions is challenging because encounters are nested within complex hospital systems and confounded by fluctuations in patient mix and resource availability. Using encounter-level data from adult inpatient admissions at an urban tertiary medical center, we leverage a hospital-defined "Jeopardy" status (periods of heightened resource strain requiring additional provider coverage) as a quasi-experimental shock to examine how LOS responds to system stress and, novelly, how this response varies across physicians within a hospital. We estimate multilevel generalized linear mixed models to characterize population-level and physician-specific effects of Jeopardy on LOS. These models let us quantify heterogeneity in provider efficiency under strain, identifying clinicians whose LOS patterns diverge from peers. Motivated by consistent associations between ICU use and LOS, we also outline a multilevel two-stage residual inclusion (2SRI) strategy that treats Jeopardy as an instrument for encounter-level ICU admission to estimate the causal effect of ICU use on LOS among encounters whose ICU status is shifted by resource strain. Together, these approaches allow us to characterize system-level impacts of resource strain and physician-level variation in responsiveness to that strain, offering a rigorous foundation for studying provider efficiency under high-demand conditions.

Applications in Health and Biology

Causal dimension reduction for multiple continuous exposures with an application to environmental mixtures analysis Thomas Hsiao* Thomas Hsiao, Howard Chang, Razieh Nabi,

Evaluating the health consequences of environmental and chemical mixtures has become a central focus in environmental epidemiology. Although substantial progress has been made in methods development, balancing flexible estimation with interpretable mixture effects remains challenging, and existing tools for drawing causal conclusions are limited. Many of these challenges stem from the difficulty of defining practical causal estimands for multidimensional, continuous exposures. We propose a sufficient dimension reduction approach that identifies low-dimensional representations preserving the causal exposure–response surface of the original mixture. We define the statistical objectives, establish theoretical properties of the resulting estimator, and evaluate its finite-sample performance through simulation. We also discuss visual and analytical strategies for interpreting the reduced dimensions. The proposed methods are compared with association-based dimension reduction techniques in simulations and illustrated through an analysis of maternal pro-inflammatory cytokine exposures and fetal inflammation during pregnancy.

Applications in Health and Biology

ENDS and Cigarette Reduction: A Causal Bayesian Additive Regression Tree Analysis Shu Xu* Shu Xu, Jennifer Hill, Luchang Cui, Yang Feng, Raymond Niaura,

There is ongoing debate about whether ENDS can help adult cigarette smokers quit smoking. One way to address this question is to examine the unbiased association between ENDS use and the frequency of cigarette smoking. However, this analysis faces challenges, including potential confounding in observational data, nonlinear and interactive relationships among study variables, and the difficulty of modeling the skewed distribution of cigarette frequency outcome.
We analyzed longitudinal data from Waves 4 through 6 of the Population Assessment of Tobacco and Health (PATH) Study. The analytic sample included 4193 adults who were current established cigarette smokers identified at Wave 4. We assessed the association between current ENDS use at Wave 5 and the number of days of cigarette smoking in the past 30 days at Wave 6 using Bayesian Additive Regression Tree analyses for causal inference. Models adjusted for a range of Wave 4 covariates (e.g., sociodemographic, nicotine dependence, and other relevant factors).
Assuming we have adjusted for all confounders, results indicate that Daily use of ENDS among established cigarette smokers is associated with a subsequent decline in days of cigarette smoking, compared to occasional or no use. These findings highlight the potential of ENDS as a harm reduction tool for adult smokers.

Applications in Health and Biology

A CV-TMLE global test Approach to Multiple-Component Endpoints in Rare Disease Clinical Trials Tianyue Zhou* Tianyue Zhou, Susan Gruber, Mark van der Laan, Hana Lee, Wonyul Lee, Lei Nie,

Rare disease trials face unique statistical challenges due to limited patient populations and heterogeneous clinical manifestations among patients. Multiple endpoints are often necessary to comprehensively capture treatment benefits. A global test is an approach for evaluating whether a treatment has any beneficial effect across multiple endpoints. We propose a new global test based on a weighted composite endpoint. The proposed global test employs shrinkage-based cross-validated targeted maximum likelihood estimation (CV TMLE) to learn data-adaptive weights that maximize power while maintaining Type I error control. Shrinkage can be tailored to incorporate existing domain knowledge, such as anticipated relative effect sizes. In simulation studies designed to reflect real rare disease trial settings, the proposed procedure demonstrated improved power over standard multiplicity adjustments and classical global tests (e.g., O'Brien test), while maintaining nominal type I error, when effects are heterogeneous across endpoints. The proposed method simultaneously learns an optimal weighted composite outcome and provides an efficient and unbiased TMLE for the average treatment effect (ATE) on that weighted outcome, with valid inference taking into account that the ATE is data dependent.

Applications in Health and Biology

Three-sided Testing with Two Control Groups with an Application to an Observational Study of the Impact of High School Football Participation on Subsequent Cognitive Functioning Iris Horng* Iris Horng, Da Wu, Sameer Deshpande, Dylan Small,

American football, the most popular high school sport for American boys, has sparked concern about its long-term cognitive effects. Using the Wisconsin Longitudinal Study dataset, which tracks individuals who graduated from Wisconsin high schools in 1957, we focus on their outcomes of letter fluency and delayed word recall at ages 65 and 72. While prior studies have used binary hypothesis tests for cognitive decline, such approaches cannot determine whether an effect is small enough to be considered negligible. To address this, we use a three-sided testing framework that simultaneously assesses superiority, inferiority, and equivalence. We further extend this framework to incorporate two distinct control groups – non-contact sport participants and non-athletes — to improve robustness to hidden bias. Our main contribution is a rigorous methodology for integrating two control groups within three-sided testing, while ensuring proper control of familywise error. Assuming limited unmeasured confounding, we find no meaningful impact of playing high school football on later life cognitive functioning. An R package, sen2controlgroups, implements the method.

Applications in Health and Biology

Ensemble Causal Structure Learning for Actionable Insights into Repeated Healthcare Utilization Shishir Adhikari* Shishir Adhikari, Guido Muscioni, Mark Shapiro, Plamen Petrov, Elena Zheleva,

Understanding the factors that trigger or prevent repeated undesirable health outcomes, such as emergency department (ED) visits and hospital readmissions, is critical for improving quality of care and reducing costs. When randomized controlled trials are infeasible, causal structure learning (CSL) provides an alternative for generating causal hypotheses from observational data, but its reliability is limited by strong assumptions and model uncertainty. We hypothesize that an ensemble of CSL algorithms improves robustness by identifying causes that persist across different assumptions. We propose an end-to-end framework that integrates an ensemble of CSL algorithms with causal effect estimation to identify and rank causal risk and preventive factors, and to quantify heterogeneous effects across subpopulations. The framework outputs candidate causes and effect modifiers with confidence scores based on agreement across methods. Experiments on synthetic and semi-synthetic data show that a majority-voting ensemble improves recall of causal factors while maintaining precision. Application to real-world healthcare data yields clinically plausible hypotheses aligned with existing knowledge and identifies subpopulations with differential susceptibility. This approach enables data-driven identification of actionable interventions, supporting targeted strategies to improve patient outcomes while reducing avoidable healthcare utilization and associated costs.

Applications in Health and Biology

CausalRAG: An Overview And Its Application to Precision Oncology Brennan Kelley* Brennan Kelley,

LLMs, when deployed in clinical support roles, have routinely hallucinated recommendations that are unvalidated. To combat this, RAG, or Retrieval-Augmented Generation, was integrated, which has partially addressed the hallucination issue by grounding the LLM outputs in retrieved literature. However, this approach can still be contextually inappropriate, and treatments that do not have enough data to evaluate their efficacy can slip through. CausalRAG adds an additional layer of validation, where the recommendations for treatment go through a causal inference pipeline, helping to ensure that the treatments recommended by the LLM produce the intended outcomes. The pipeline involves an Empirical Bounding Box, Abstract extraction, and the DoWhy and EconML packages. The poster presentation introduces CausalRAG and display the results of some trials of CausalRAG in the context of the TCGA-BRCA breast cancer dataset containing records for 1,050 patients. Along with a breakdown of the Causal Pipeline.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

Is Cross Country a Team Sport? A Bayesian Hierarchical Analysis of Attached vs. Unattached Runners Jared Fisher* Jared Fisher, Nathan Sandholtz, Sam Lee, Brylee Wilcox, Garritt Page,

Cross country races are scored as team competitions yet experienced primarily as individual efforts, raising the question of whether team affiliation measurably affects individual performance. We investigate this question by comparing collegiate runners' race times when they compete "attached" to a team versus "unattached." Using more than a decade of results from the Track and Field Results Reporting System (TFRRS), we develop a directed acyclic graph (DAG) to formalize the causal structure governing performance and to identify covariates required to block relevant backdoor paths. Guided by this DAG, we fit Bayesian hierarchical models that adjust for runner ability, race distance, course, year in school, and time-in-season, i.e. factors that jointly confound team status and individual performance. Across specifications, we find consistent evidence that runners perform better when racing as part of a team, improving by roughly seven to thirteen seconds depending on race distance. These findings provide quantitative, causally interpretable evidence that team affiliation enhances individual performance, suggesting that psychological or motivational mechanisms associated with shared identity play a meaningful role even in a minimally interactive sport.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

Challenges and Opportunities in Causal Inference with Complex Treatments Michael Valancius* Michael Valancius,

Most causal inference research has centered on estimating average effects of simple, well-defined interventions. While these methods have been highly successful, further innovation is needed to tackle the challenges posed when treatments are complex, high-dimensional objects such as images, audio, or text. The rapid proliferation of AI systems has only amplified this challenge: their outputs are rarely directly manipulable, defy straightforward characterization, and pose significant obstacles for causal discovery and intervention.
We argue that the next frontier for causal inference lies in developing methods that can reason about nuanced, individualized, and counterfactual questions in these complex settings. For example: Would a different dub of a specific movie have improved member satisfaction? Or, how can we define the relevant quality components of a video to understand their causal impact? In many real-world applications, particularly in R&D and creative support, actionable feedback and structured discovery are as important as traditional effect estimation, and often require “reverse” causal reasoning.
We illustrate these challenges and opportunities with examples from Netflix, where treatments like dubs and artwork are inherently complex, and understanding their causal impact is essential for supporting creative innovation. We discuss limitations of current methodologies, highlight open questions, and share early modeling directions.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

Causal Inference for AIOps: Root Cause Analysis in Microservices Incidents Leandro Siqueira* Leandro Siqueira, Victor Medeiros, Willian Honorato, Kauê França, Alvino Júnior,

Microservice architectures propagate latency in complex, non-linear ways, making root cause analysis of tail latency degradation difficult when relying solely on correlations. In banking production environments, P95 latency is a key indicator of user-perceived impact during incidents. This work presents a causal AI approach based on graphical causal models (GCM/DAGs) applied to real distributed trace data to explain increases in P95 (ΔP95) in an interpretable and operationally actionable manner. The method is explicitly route aware: instead of relying on an aggregated service topology, it constructs just-in-time causal DAGs conditioned on the exact customer journeys actually traversed.
From 72 observed routes in production, only four exhibited significant degradation and were selected for detailed causal analysis, ensuring the model focuses on real, impactful user behavior. For each impacted route, the operational call graph is inverted to model upstream services as candidate causes of end-to-end latency at the BFF. Exclusive service latency defines each node’s causal mechanism, learned from baseline traces and contrasted against anomalous traces isolated via Isolation Forest. ΔP95 is attributed to services using Shapley values over mechanism changes, producing responsibility scores that disentangle true drivers from propagated effects.
Across the four impacted routes in a 52-service system, ΔP95 consistently concentrates in a small subset of services -predominantly Core and

Bayesian Causal Inference

Bay-PIE: Correcting Attenuation Bias in Predictive Incrementality by Experimentation Ran Wang* Ran Wang,

Predictive Incrementality by Experimentation (PIE; Gordon et al., 2023) calibrates attribution by relating total lift to attributed lift across many randomized experiments. In practice, both lifts are often noisy finite-sample estimates with heterogeneous standard errors, and they may share sampling noise within an experiment. Conventional OLS/WLS PIE can therefore exhibit attenuation bias, inefficient weighting, and poorly calibrated intervals.

We propose Bayesian PIE (Bay-PIE), a Hierarchical Bayesian Model for the latent PIE relationship. Both observed lifts are treated as error-contaminated measurements (e.g., hat{tau}_i^{total}=tau_i^{total}+v_i^{total}, hat{tau}_i^{attr}=tau_i^{attr}+v_i^{attr}); Bay-PIE effectively deconvolves measurement noise to recover latent lifts, with within-experiment dependence reduced via sample splitting. In simulations, Bay-PIE reduces bias and improves interval coverage versus WLS PIE, with larger gains when attribution is noisier. Using the public Criteo Attribution Modeling for Bidding dataset, Bay-PIE corrects slope shrinkage and helps separate low-signal from low-precision regimes. The key advantage is modularity for predictive-incrementality calibration: Bay-PIE casts PIE as a single probabilistic measurement-error model where weighting and extensions are built into the Bayesian framework and estimated via unified MCMC posterior inference, rather than using ad-hoc weighting rules or re-derived OLS/MLE estimators.

Bayesian Causal Inference

Improving Generative Methods for Causal Evaluation via Simulation-Based Inference
Pracheta Amaranath* Pracheta Amaranath, Vinitra Muralikrishnan, Amit Sharma, David Jensen,

Generating synthetic datasets that accurately reflect real-world observational data is critical for evaluating causal estimators, but remains a challenging task. Existing generative methods offer a solution by producing synthetic datasets anchored in the observed data (source data) while allowing variation in key parameters such as the treatment effect and amount of confounding bias. However, it is often unclear which generative methods to use and which values of parameters to choose when generating synthetic datasets. Moreover, existing methods typically require users to provide point estimates of such parameters (rather than distributions) and fixed estimates (rather than estimates that can be improved with reference to the source data). This denies users the ability to express uncertainty over both generative methods and parameter values and removes the potential for posterior inference, potentially leading to unreliable estimator comparisons. We introduce simulation-based inference for causal evaluation (SBICE), a framework that models the generative method and its corresponding generative parameters as uncertain and infers their posterior distribution given a source dataset. Leveraging techniques in simulation-based inference, SBICE identifies suitable generative methods and infers distributions over its parameter configurations to produce synthetic datasets closely aligned with the source data distribution, improving the reliability of estimator evaluations.

Causal Bounds, Partial Identification

Information-Theoretic Causal Bounds Yonghan Jung* Yonghan Jung,

We develop an information-theoretic framework for partial identification of causal effects under unmeasured confounding. Existing partial identification approaches suffer from one or more of the following limitations: restricting outcomes to be bounded or discrete, requiring auxiliary inputs (instruments, proxies, or user-specified sensitivity parameters), necessitating full structural modeling, or neglecting effect heterogeneity. We universally address these limitations through a novel information-theoretic divergence bound. Our key insight is that the f-divergences of the observational distribution P(Y | A=a, X=x) and the interventional distribution P(Y | do(A=a), X=x) are upper bounded by a function of the propensity score Pr(A=a|X=x). We translate these upper bounds of f-divergences into sharp lower and upper bounds of conditional causal effects without requiring boundedness assumptions, auxiliary variables, or full data-generating process specification. We develop a semiparametric estimator for the proposed causal bounds that attains fast convergence rates even when nuisance components converge slowly. Simulation studies and real data applications demonstrate the practical utility of our bounds.

Causal Discovery

Causal Discovery for High-Dimensional Functional Data with Latent Confounders Filippo Fiocchi* Filippo Fiocchi, Samuel Wang,

Constraint-based causal discovery methods such as the PC algorithm are widely used to infer causal structure from observational data, but their application to high-dimensional functional data has only recently been explored. Moreover, existing approaches assume that all variables of interest are observed, excluding the presence of latent confounders. In this work, we develop a framework for causal discovery in multivariate functional data that extends the PC algorithm to settings with unobserved confounding. Our approach estimates conditional independence relations of the underlying unconfounded processes via low-rank–plus–signal decompositions of covariance operators, enabling separation of latent confounding effects from intrinsic dependencies. Under partial separability and more general covariance structures, we show that conditional independences between functional variables can be characterized through partial correlations of projected scores, leading to a modified PC procedure that remains valid in high-dimensional regimes. We establish consistency guarantees for skeleton recovery and CPDAG estimation, and demonstrate through simulations that the proposed method reliably recovers causal structure even when classical functional PC methods fail in the presence of hidden confounders.

Causal Discovery

Convex Mixed-Integer Programming for Causal Additive Models with Optimization and Statistical Guarantees Xiaozhu Zhang* Xiaozhu Zhang, Nir Keret, Ali Shojaie, Armeen Taeb,

We study the problem of learning a directed acyclic graph from data generated according to an additive, non-linear structural equation model with Gaussian noise. We express each non-linear function through a basis expansion, and derive a maximum likelihood estimator with a group l0-regularization that penalizes the number of edges in the graph. The resulting estimator is formulated through a convex mixed-integer program, enabling the use of branch-and-bound methods to obtain a solution that is guaranteed to be accurate up to a pre-specified optimality gap. Our formulation can naturally encode background knowledge, such as the presence or absence of edges and partial ordering constraints among the variables. We establish consistency guarantees for our estimator in terms of graph recovery, even when the number of variables grows with the sample size. Additionally, by connecting the optimality guarantees with our statistical error bounds, we derive an early stopping criterion that allows terminating the branch-and-bound procedure while preserving consistency. Compared with existing approaches that either assume equal error variances, restrict to linear structural equation models, or rely on heuristic procedures, our method enjoys both optimization and statistical guarantees. Extensive simulations and real-data analysis show that the proposed method achieves markedly better graph recovery performance.

Causal Discovery

Causal Discovery in Game Telemetry Minh Nguyen* Minh Nguyen,

Causal discovery is increasingly applied to large-scale telemetry data to estimate the effects of user-facing interventions, yet its reliability for decision-making in feedback-driven systems with strong self-selection remains unclear. In this paper, we propose an effect-centric, admissibility-first framework that treats discovered graphs as structural hypotheses and evaluates them by identifiability, stability, and falsification rather than by graph recovery accuracy alone. Empirically, we study the effect of early exposure to competitive gameplay on short-term retention using real-world game telemetry.We find that many statistically plausible discovery outputs do not admit point-identified causal queries once minimal temporal and semantic constraints are enforced, highlighting identifiability as a critical bottleneck for decision support. When identification is possible, several algorithm families converge to similar, decision-consistent effect estimates despite producing substantially different graph structures, including cases where the direct treatment–outcome edge is absent and the effect is preserved through indirect causal pathways. These converging estimates survive placebo, subsampling, and sensitivity refutation. In contrast, other methods exhibit sporadic admissibility and threshold-sensitive or attenuated effects due to endpoint ambiguity. These results suggest that graph-level metrics alone are inadequate proxies for causal reliability for a given target query.

Causal Fairness, and Bias/Discrimination

Fairness-Constrained Individualized Treatment Rules and Medication Decisions for Opioid Use Disorder Safiya Sirota* Safiya Sirota, Rachael Ross, Kara Rudolph, Daniel Malinsky,

There are multiple evidence-based medications that are commonly used for the treatment of opioid use disorder (OUD). It is not well-established which drug most effectively prevents negative health outcomes, given a patient's individual characteristics. One may endeavor to estimate an individualized treatment rule (ITR) from observational data to address this question. However, observational datasets may encode social or structural biases (e.g., provider bias) that reflect disadvantages faced by marginalized groups. Imposing fairness constraints may help mitigate these pre-existing biases in the training data. We outline a framework for estimating an optimal fair ITR, using a causal path-specific approach to fairness. Our procedure enables estimation of the constrained ITR's value and valid quantification of uncertainty in the value. We demonstrate the methodology with simulated data and an application to Medicaid data on medication decisions for patients with OUD. The estimation of a valid confidence interval around the fair ITR value provides a means of comparing alternative candidate rules, including the optimal unconstrained ITR. Assessing differences in expected outcomes under candidate ITRs is crucial to knowledgeably balance the dual objectives of curbing retrospective bias and suggesting treatment regimes that perform well.

Causal Fairness, and Bias/Discrimination

A Formal Causal Perspective on Outcome Tests for Discrimination Kai Cooper* Kai Cooper, Dean Knox,

Outcome tests detect discrimination by comparing success rates across groups: differing standards should produce different success rates. But inframarginality bias (ignorance of individual heterogeneity) undermines this—success rates mix obvious and borderline cases, obscuring the differential nature of the decision standard. Recent work shows even comparing marginal cases requires strong structural assumptions. Their remedy, via an econometric Roy Model, improves on earlier game-theoretic approaches but needs restrictive assumptions which are difficult to justify in practice. We propose probabilistic epistemic causal models (ECMs), extending structural causal models by giving a mathematical representation of decision makers' beliefs about counterfactual outcomes across possible worlds via graphical models defined by Single World Intervention Graphs. Our approach permits the analyst to target a broader range of estimands of interest in empirical studies of discrimination without requiring game-theoretic notions of equilibrium or parametric specifications of decision rules. Additionally, since decisions depend on both group membership and case characteristics, the decision node acts as a collider. In turn, we offer an avenue to understand outcome tests through collider bias and causal interaction. We demonstrate our findings across a range of domains including: college admissions, bail decisions, and traffic stops and searches.

Causal Fairness, and Bias/Discrimination

The IEEE P3591 Standard for Fair Decision Making Through Causal Analysis Christopher Lam* Christopher Lam,

Causal inference is increasingly used to study algorithmic bias, fairness, and discrimination in high-stakes decision systems, yet the field lacks a shared, operational standard for representing these concepts in causal terms. This gap is often filled by statistical fairness metrics that are incompatible with one another and poorly aligned with legal and regulatory requirements.

IEEE P3591 is an emerging international standard that defines fair decision making as a property of causal structure rather than statistical outcomes. The standard introduces a causal ontology for protected attributes, mediators, decisions, and outcomes, and specifies how permissible and impermissible causal pathways can be identified and documented. IEEE P3591 is designed to bridge causal inference research and real-world AI deployment while preserving methodological flexibility, providing a common causal language for policymakers, lawyers, economists, and data scientists.

Causal Inference in Networks

Design and analysis for valid causal inference with Network-dependent data Zhejia Dong*
Zhejia Dong, Youjin Lee,

Matching is widely used to mimic randomized experiments by forming matched sets in which treated and control units differ only randomly with respect to important confounding variables. However, when the study population consists of interconnected units from a single network or a small number of networks, matching solely on confounding variables may produce matched units that are not randomly different with respect to their network distance, but instead are more likely to be closely connected after matching. Such increased network closeness within matched sets may induce spurious associations between treatment and outcome, when both variables exhibit shared autocorrelation patterns on the network. To reduce spurious associations within matched sets while preserving the validity of within-matched-set causal comparisons, we propose a new matching method that matches units with similar covariates while reducing within-matched-set dependence by imposing additional constraints on network proximity. Furthermore, at the analysis stage, to account for residual dependence across matched sets, we propose a valid randomization inference procedure for testing the sharp null hypothesis of no causal effect that accommodates across-matched-set dependence without explicit assumptions on the underlying dependence structure. We demonstrate the validity and utility of the proposed methods through simulation studies and an application to real-world HIV transmission network data.

Causal Inference in Networks

AIPW Estimation in Network Experiments Gernot Zöcklein* Christopher Harshaw, Gernot Zöcklein,

In classical experiments without interference, AIPW estimators use covariate information to achieve improved efficiency over unadjusted estimators such as Horvitz—Thompson. Considerably less is known about the efficiency of similarly adjusted estimators for network experiments, where treatment given to one unit can affect the outcomes of neighboring units. While AIPW-style estimators for network experiments have been proposed and analyzed throughout the literature, they may — in some cases — incur a larger variance than their unadjusted counterparts. A central challenge which prevents efficiency gains is the irregular correlations of exposures which are necessarily induced by the underlying network.

In this paper, we propose an AIPW estimator for arbitrary contrastive effects in network experiments. Our AIPW estimator is constructed to perform well under the recently introduced Conflict Graph Design of Kandiros et al (2024), which achieves the currently best known rates of estimation. We show that the AIPW estimator achieves improved efficiency over the Horvitz–Thompson estimator when the dimension of the covariates is appropriately asymptotically bounded. An important aspect of our AIPW estimator is a “pruning away” of high degree nodes as a means to resolve a bias-variance trade-off, which may be of independent interest. We provide conservative variance estimators which facilitate asymptotically valid inference.

Causal Inference in Networks

Improving causal inference controls using network theory in discrete choice data Bernardo Modenesi* Bernardo Modenesi, Sima Najafzadehkhoei,

Many datasets in health and social sciences result from agents making repeated choices over time, each choice leading to an observable outcome. Researchers often aim to model the causal impact of covariates on the outcome variable using various estimation strategies (e.g. fixed effects regression, difference-in-differences, instrumental variables, etc). We propose a new way to increase control in these estimation procedures by applying network theory models motivated by a discrete choice framework. We suggest representing these datasets as a bipartite network, where agents are nodes on one side and choices are nodes on the other. Edges in this network represent a choice made by an agent at a certain time, stemming from a discrete choice problem. We argue that the structure of connections in this choice-network allows the researcher to further improve controls when modeling the outcome variable. For instance, we use the choice-network to project agents into a multidimensional latent space that captures each agent’s choice-profile. Distances between agents in this latent space represent a metric of similarity between them. By exploring the high-dimensional choice-profile of agents, we propose several ways to enhance causal inference exercises, as well as to compute heterogeneous treatment effects.

Causal Inference in Networks

Optimal Design under Interference, Homophily, and Robustness Trade-offs Vydhourie Thiyageswaran* Vydhourie Thiyageswaran, Alex Kokot, Jennifer Brennan, Marina Meila, Christina Yu, Maryam Fazel,

To minimize the mean squared error (MSE) in global average treatment effect (GATE) estimation under network interference, a popular approach is to use a cluster-randomized design. However, in the presence of homophily, which is common in social networks, cluster randomization can instead increase the MSE. We develop a novel potential outcomes model that accounts for interference, homophily, and heterogeneous variation. In this setting, we establish a framework for optimizing designs for worst-case MSE under the Horvitz-Thompson estimator. This leads to an optimization problem over the covariance matrices of the treatment assignment, trading off interference, homophily, and robustness. We frame and solve this problem using two complementary approaches. The first involves formulating a semidefinite program (SDP) and employing Gaussian rounding, in the spirit of the Goemans-Williamson approximation algorithm for MAXCUT. The second is an adaptation of the Gram-Schmidt Walk, a vector-balancing algorithm which has recently received much attention. Finally, we evaluate the performance of our designs through various experiments on simulated network data and a real village network dataset.

Design of Experiments

Does Rerandomization Help Beyond Covariate Adjustment? A Review and Guide for Theory and Practice Antônio Carlos Ribeiro Junior* Antônio Carlos Ribeiro Junior, Zach Branson,

Rerandomization is a modern experimental design technique that repeatedly randomizes treatment assignments until covariates are deemed balanced between treatment groups. This enhances the precision and coherence of causal effect estimators, mitigates false discoveries from p-hacking, and increases statistical power. Recent work suggests that balancing covariates via rerandomization does not alter the asymptotic precision of covariate-adjusted estimators, thereby making it unclear whether rerandomization is worthwhile if adjusted estimators are used. However, these results have two key caveats. First, these results are asymptotic, leaving finite sample performance unknown. Second, these results focus on precision, while other potential benefits, such as increased coherence among flexible estimators, remain understudied. Hence, in this paper we provide three main contributions: (i) a comprehensive review of the rerandomization literature, covering historical foundations, theoretical developments, and recent methodological advancements, (ii) an extensive simulation study examining finite-sample performance, and (iii) a practical guide for practitioners. Our study compares precision, coherence, power, and coverage of various estimators under rerandomization versus complete randomization. We find rerandomization to be a complementary design strategy that enhances the precision, robustness, and reliability of causal effect estimators, especially for smaller sample sizes.

Design of Experiments

Logging Policy Design for Efficient Off-Policy Evaluation Connor Douglas* Connor Douglas, Joel Persson, Foster Provost,

Off-policy evaluation (OPE) estimates the value of a candidate “target” policy, such as a recommender system, using data logged by a different “logging” policy, enabling safe assessment without deploying changes live. While prior work emphasizes estimator guarantees under strong assumptions, in practice, the logging policy is a first-order driver of OPE quality. We demonstrate this and study how to optimally design logging policies to efficiently evaluate target policies. We characterize a reward–coverage tradeoff in choosing which actions to log and provide a sufficient condition for when an item should enter the logging support. We introduce a unifying framework that characterizes logging design settings by (i) what is known about the target policy and (ii) what is known about reward distributions. Within this space, we derive optimal logging policies in extremes where target policies and rewards are fully known or fully unknown, and show that OPE can actually improve policy value estimates compared to on-policy estimates. We extend to intermediate cases with probabilistic knowledge over target policies and noisy reward estimates, yielding optimal designs under each information regime. Our results provide actionable guidance for firms that must collect data to compare multiple candidate recommendation policies when the optimal logging policy may be infeasible.

Design of Experiments

Emulating Factorial Designs for Multiple Concurrent Binary Interventions Nicholas Bakewell* Nicholas Bakewell,

The causal inference literature has focused on univariate binary interventions. However, many interventions are implemented concurrently, and ignoring between-intervention dependence or treating concurrent interventions as nuisance parameters may bias estimates. Recent multiple causes literature uses linear factor models, copulas, or other approaches (e.g. Bahadur expansion, saturated models) to capture dependence among interventions; however, these may ignore residual dependence, lack uniqueness, use inefficient parameterizations or cannot disentangle main and interaction effects. Focusing on settings where few binary interventions are selected based on subject-matter knowledge, Target Trial Emulation and potential outcomes frameworks are used to conceptualize this as a 2^k factorial design. Appealing to sparsity-of-effects, main effects and two-way interaction estimands are of interest. Integrating Empirical Likelihood (EL), calibrated covariate balancing weights, and joint marginal structural models (MSMs), estimation approaches are outlined for point and time-varying interventions. Simulations compare the proposed approach to existing joint MSMs estimation approaches via logistic regression-based inverse probability weighting and covariate balancing propensity score. Existing approaches scale poorly with the number of interventions/time periods, are more sensitive to misspecification of the joint distribution of interventions and do not account for simultaneous inference.

Design of Experiments

Flexible inference with split samples via data turnover William Bekerman* William Bekerman, Dylan Small,

We introduce data turnover, a general framework enabling a single group of statisticians and domain experts to assess the strength of evidence gathered across multiple data splits, effectively integrating both qualitative and quantitative findings from data exploration. Data turnover can accommodate a wide range of statistical tasks, including inference and estimation, while ensuring the validity of certain data-driven decisions and providing a straightforward approach to evaluating replicability. As a motivating example, we study the effects of growing up with a father with an alcohol use disorder on later life outcomes. Data turnover allows us to augment our analysis with exploratory insights while leveraging the full dataset for confirmatory testing, avoiding stringent adjustments for post-selection inference that can erode power. We also apply our new technique to evaluate variable importance in a clinical prediction model of mortality in premature babies. We prove the theoretical validity of our procedure and examine its power in extensive simulation studies.

Design of Experiments

Semiparametric Manski-Style Inference and Optimal Experimental Design in Demand Modeling Jia Wan* Jia Wan, Antoine Scheid, Guy Aridor, Nathan Kallus, Aurélien Bibaut,

A key challenge for online platforms is measuring demand in environments with limited price variation. In this paper we study what platforms can learn about demand via experiments that vary the set of recommendations presented and apply it to counterfactuals of interest. We consider a canonical model of choice – the mixed multinomial logit model – and develop a semiparametric framework for sharp Manski-style bounds on linear functionals of counterfactual choice probabilities with no restrictions on preference heterogeneity. This provides identification bounds of counterfactual shares of particular goods (or overall engagement) under a range of counterfactuals, including procurement of new goods and alternative recommendation algorithms. Our main characterization shows that the first-order sensitivity of these bounds admits a Riesz representer on the constraint range – equivalently, the dual multipliers of an infinite-dimensional linear program – which yields an efficient influence function. For inference, we compute identification bound endpoints by scanning a KL distance-to-feasibility functional, evaluated via an EM-type alternating KL projections for the mixture nuisance, and then apply a cross-fitted one-step correction to obtain asymptotically normal inference. Finally, we study optimal experimental design – selecting a set of recommendation slates – to minimize the worst-case bound width for a target counterfactual and support our findings with simulation results.

Design of Experiments

AI-assisted design and analysis of experiments with unstructured treatments Eli Ben-Michael* Eli Ben-Michael, Zach Branson,

Randomized experiments with unstructured treatments—such as text or images—are common in social science research. However, isolating the causal effect of a focal attribute (e.g., the style of text or facial features in images) is challenging because the attribute is typically correlated with other, non-focal attributes of the treatments. While AI technology could be used in an attempt to change focal attributes of a treatment while keeping all non-focal attributes identical, it offers no guarantees that non-focal attributes are not inadvertently changed in the process, such that confounding can still be an issue. We develop a framework for designing and analyzing experiments that target the isolated effect of a binary attribute of unstructured treatments. We consider designs where treatments are drawn from arbitrary distributions—including hand-crafted treatments, existing databases, or AI systems—and we map the bias of the difference-in-means estimator to the discrepancy in non-focal attributes across treatment arms. We develop a procedure that minimizes this bias via a second-stage rejection sampler that adjusts for observable imbalances in non-focal attributes, without assuming the original distributions correctly isolate the focal attribute. For analysis, we show how to conduct asymptotic inference for the difference-in-means estimator in a finite population setting, where inference is justified by the randomization of treatment. We also develop a calibrated model

Design-Based Causal Inference

Estimating within-cluster and between-cluster spillover effects in randomized saturation designs Sizhu Lu* Sizhu Lu, Lei Shi, Peng Ding,

Randomized saturation designs are two-stage experiments: they first randomly assign treatment probabilities over the clusters and then randomly assign the treatment to the units within the clusters. The existing literature on randomized saturation designs focuses on estimating within-cluster spillover effects by assuming away between-cluster spillover effects. However, the units may interact across clusters in many practical randomized saturation designs. A leading example is that some units are geographically close to each other, so spillover effects arise across clusters. Based on the potential outcomes framework, we formulate the causal inference problem of estimating within-cluster and between-cluster spillover effects in randomized saturation designs. We clarify the causal estimands and establish the statistical theory for estimation and inference. We also apply our method to analyze a recent randomized saturation design of cash transfer on household expenditure in Kenya.

Design-Based Causal Inference

Data-driven hypotheses and designs for theory testing and theory building Molly Offer-Westort* Molly Offer-Westort,

This project develops a design-based framework for theory testing when a theory admits multiple empirical realizations, and theory building when multiple alternative mechanisms are under consideration. Experiments in the social sciences often test a single treatment arm against a control, or test the effects of multiple treatments independently. However, many social-scientific mechanisms can be implemented through a family of related interventions. This is formalized by indexing a treatment space W and defining data-driven hypotheses such as an EXISTENCE ($\max_{w \in W} \theta(w) \leq c$) or UNIVERSAL ($\min_{w \in W} \theta(w) \leq c$) hypothesis. For these types of hypotheses, data-driven explore-confirm designs that separate learning from inference can improve power. A simple power factorization is used to compare uniform, cross-fit, and data adaptive designs. Simple two-stage data-adaptive designs improve power by concentrating confirmatory allocation to the treatment(s) most likely to yield rejection under the alternative. Adaptive first stage exploration improves selection quality. The framework links selective inference, adaptive experimentation, and mechanism testing, providing a general approach to data-driven design.

Design-Based Causal Inference

Design-based Inference with the Estimated Propensity Score Shunzhuang Huang*
Shunzhuang Huang, Jiangchuan Du, Azeem Shaikh, Panos Toulis,

In this paper, we study the properties of design-based (or randomization) inference for treatment effects when analyzing observational data under ignorability. In such settings, we interpret the common ignorability assumption as defining an artificial randomized experiment and study approximate randomization tests that use the estimated propensity score, i.e., the distribution of treatment status given the covariates, as a foundation for design-based inference. Under the sharp null hypothesis of no treatment effect in distribution, we derive non-asymptotic bounds on the size distortion of such tests that depend only on the error in estimating the propensity score. Under the weak null hypothesis of no average treatment effect, we show that the proposed tests are asymptotically valid for common estimators, including inverse-propensity-weighted and doubly-robust estimators. We further compare our tests with conventional tests based on the asymptotic normality for the weak null hypothesis and, since these tests are shown to be first-order equivalent, develop higher-order comparisons using novel Edgeworth expansions. Our analysis reveals that, from this perspective, neither approach uniformly dominates the other. However, the randomization test for the weak null achieves higher-order accuracy when the sharp null “nearly holds”; for example, when treatment effects are small or rare.

Design-Based Causal Inference

Causal Effects of Health-Related Social Needs on Mammography Screening: Propensity-Score Weighting in Complex Survey Data Fode Tounkara* Fode Tounkara,

Screening mammography lowers breast cancer mortality, yet many women in the United States remain overdue for recommended screening. Uptake is especially low among women facing social and economic hardship. Health-related social needs (HRSNs), such as housing instability, food insecurity, and transportation barriers, can interfere with preventive care by limiting access, increasing stress, and competing with daily survival needs. Although these factors are widely recognized, their cumulative impact on mammography use has rarely been evaluated using causal methods in nationally representative data.
We analyzed data from the 2023 National Health Interview Survey to examine whether unmet HRSNs are associated with being up to date on screening mammography among women aged 40 to 74 years. Both individual social needs and a cumulative burden index (0, 1, or ≥2 unmet needs) were assessed. To address confounding while accounting for the complex survey design, we applied inverse probability of treatment weighting (IPW) integrated with survey weights, stratification, and clustering. Covariate balance was evaluated using standardized mean differences. Weighted logistic regression models estimated marginal associations overall and within age groups (40–49, 50–64, and 65–74 years).
After weighting, covariate balance improved substantially across exposure groups. Unmet social needs were associated with lower odds of guideline-concordant mammography, with stronger effects observed as the num

Design-Based Causal Inference

Stratified Sampling for Model-Assisted Estimation with Surrogate Outcomes Reagan Mozer* Reagan Mozer,

In many randomized trials, outcomes such as essays or open-ended responses must be manually scored before impact analysis, a process that is costly and limiting. Model-assisted estimation combines surrogate outcomes from machine learning or large language models with a human-coded subset to obtain unbiased estimates, but existing approaches rely on simple random sampling and ignore systematic structure in prediction errors. We extend this framework by incorporating stratified sampling to more efficiently allocate human coding effort. We derive the exact variance of the stratified estimator, characterize conditions under which stratification improves precision, and identify a Neyman-type optimal allocation rule that oversamples strata with larger residual variance. Comprehensive simulation studies confirm that stratification consistently improves efficiency when surrogate prediction errors exhibit structured bias or heteroskedasticity. We present two empirical applications, including an education RCT and a large observational corpus, to illustrate practical implementation using ChatGPT-generated surrogate outcomes. Overall, this framework provides a practical design-based approach for leveraging surrogate outcomes and strategically allocating human coding effort to obtain unbiased estimates with greater efficiency.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Small-Sample Performance of Synthetic Difference-in-Differences Luke Stewart* Luke Stewart, Nandita Mitra, Youjin Lee, Gary Hettinger,

Methods for estimating causal effects in longitudinal, quasi-experimental settings are widely used in economics, public health, and other fields. The recently proposed synthetic difference-in-differences (SDiD) method improves on difference-in-differences (DiD) and synthetic controls (SC) by weakening the conditions necessary to identify and estimate a causal effect in settings with longitudinal panel data or repeated cross sectional data. However, these studies often have limited sample sizes, both in terms of units and periods analyzed. Although SDiD generally relies on weaker assumptions than both DiD and SC, the required asymptotics may restrict the use of the method.

To evaluate the performance of SDiD in realistic, small-sample settings, we employ two complementary simulation strategies. First, we reanalyze existing placebo control simulation studies while artificially limiting the number of units and periods. Second, we introduce a novel simulation framework for panel data that allows the injection of a known treatment effect into existing data in a setting of interest. We simulate data under small samples and assess the robustness of SDiD to violations of its assumptions about the error term in the latent factor model. Drawing on findings from these investigations, we offer guidance for researchers on when SDiD is likely to yield reliable estimates. Finally, we apply SDiD to an investigation of the effect of the Philadelphia Beverage Tax on youth soda consumption.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Proximal Learning for Trials With External Controls: A Case Study in HIV Prevention Yilin Song* Yilin Song, Yinxiang Wu, Ting Ye,

With the advent of effective pre-exposure prophylaxis agents, active-controlled HIV prevention trials have become a common study design. Nevertheless, estimating absolute efficacy relative to a placebo remains important. In this paper, we introduce a novel application of proximal causal inference methods to estimate the counterfactual cumulative HIV incidence under placebo for participants in an active-controlled trial of cabotegravir, using external control data from a placebo-controlled trial with similar eligibility criteria. We leverage baseline sexually transmitted infection status and geographic region as negative control outcome and exposure variables, respectively. We address two key challenges: unmeasured differences in HIV risk between trials and statistical difficulties arising from low HIV incidence rates in both studies. To overcome these challenges, we develop two proximal inference approaches: (1) a semiparametric inverse probability of censoring weighting estimator and (2) a two-stage regression-based strategy tailored to low-event-rate settings. Our theoretical and numerical investigations demonstrate that these methods yield reliable estimates of the counterfactual one-year cumulative HIV incidence under placebo, and provide robust evidence of the superior efficacy of cabotegravir compared with placebo. These findings highlight the potential of proximal inference methods to estimate placebo-controlled effects in both single-arm and active-controlled trials.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Estimating effects of longitudinal modified treatment policies (LMTPs) at target times in studies with irregular assessment times Anja Shahu* Anja Shahu, Daniel Malinsky,

Longitudinal studies are often designed to assess participants at a common set of pre-specified times after baseline (e.g., annual visits over a fixed follow-up period). In practice, however, assessment times can vary considerably from these targets in both timing and frequency. Such irregular assessment times pose challenges for estimating causal effects at target times, as outcomes are not observed at those times for all participants and the subset of participants observed at or around those times may not be representative due to informative assessment timing. In Shahu et al. (2025), we introduced a framework for estimating and testing hypotheses about effects of complex interventions on rates of change in an outcome over time and demonstrated its utility for examining the effect of a longitudinal shift intervention on the trajectory of an outcome over time. The original framework was developed for balanced discrete-time longitudinal studies with fixed visit schedules. We extend this framework to accommodate a setting with irregular assessment times by introducing a novel estimator for the projection of the causal effect at a specified target time. The proposed approach enables analysis of longitudinal studies, in which participants are assessed irregularly within pre-defined visit windows and randomized to be observed at only one of two consecutive visits. Through a simulation study, we illustrate the performance of our proposed approach in this setting.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Inference for synthetic controls via refined placebo tests Lihua Lei* Lihua Lei, Timothy Sudijono,

The synthetic control method is often applied to problems with one treated unit and a small number of control units. A common inferential task in this setting is to test null hypotheses regarding the average treatment effect on the treated. Inference procedures that are justified asymptotically are often unsatisfactory due to (1) small sample sizes that render large-sample approximation fragile and (2) simplification of the estimation procedure that is implemented in practice. An alternative is permutation inference, which is related to a common diagnostic called the placebo test. It has provable Type-I error guarantees in finite samples without simplification of the method, when the treatment is uniformly assigned. However, it often has low power at a common level like $\alpha=0.05$ when N is small. We propose a novel leave-two-out procedure that bypasses this issue, while still maintaining the same finite-sample Type-I error guarantee under uniform assignment for a wide range of N. Unlike the placebo test whose Type-I error always equals the theoretical upper bound, our procedure often achieves a lower unconditional Type-I error than theory suggests; this enables useful inference in the challenging regime when $\alpha<1/N$. Empirically, our procedure achieves a higher power when the effect size is reasonably large and a comparable power otherwise. We generalize our procedure to non-uniform assignments and show how to conduct sensitivity analysis.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Proximal Causal Inference for Contemporaneous Treatment Effect Estimation in Time Series Data Fanyu Cui* Fanyu Cui, Charlotte Fowler, Xiaoxuan Cai, Jukka-Pekka Onnela, Justin T Baker, Linda Valeri,

Unmeasured confounding challenges causal inference in intensive longitudinal studies, potentially biasing treatment effect estimates. The proximal causal inference (PCI) framework offers a promising approach to nonparametric identification using proxies or negative control variables in the presence of hidden confounding bias. While prior literature considers the joint effect of time-varying treatments, our work extends the framework to a time series setting to estimate the contemporaneous or lagged effect of time-varying treatments. We demonstrate that under traditional PCI assumptions, we can recover unbiased effect estimation in the presence of unmeasured confounding by leveraging the intensive longitudinal nature of time series data. Specifically, we develop identifiability conditions and show that past and future observations can be employed as natural proxies for unmeasured confounders. We further develop the bridge function required for valid proximal causal inference estimation along with asymptotic variance. Simulation studies illustrate our method’s validity and robustness to violation of certain PCI identification assumptions. We demonstrate the potential for studying socio-environmental exposure health effects applying our method to a smartphone study of bipolar patients. Our work contributes to the growing literature on PCI and provides a powerful tool for analyzing longitudinal data, including in observational mobile health research prone to confounding bias.

Dynamic Treatment Regimes

Causality with aging: Estimation and inference in a dynamic directed acyclic VAR graph with aging and subaging Andrej Srakar* Andrej Srakar,

Directed acyclic graphs have seldom been studied in a dynamic context. We include stochastic aging in vector autoregression form for causal time-series. Let
(G = (mathcal{V,E)}) be a graph, and let (E = {{ E_{i}}}_{i in mathcal{V}}) be collection of i.i.d. random variables indexed by vertices of graph with exponential distribution with mean (1). We consider continuous-time Markov chain (X(t)) with state space (mathcal{V}). Transition rates are (w_{ij} = nuexpleft(– betaleft((1 – a)E_{i} – aE_{j} right) right)). Proving an aging result consists in finding two-point function
(F(t_{w},t_{w} + t)) such that nontrivial limit(lim_{t rightarrow infty,frac{t}{t_{w}} = theta}{F(t_{w},t_{w} + t)} = F(theta)) exists. We consider VAR-based dynamic DAG combined with trap aging model above and prove results on aging and subaging. We propose causal estimators in this context and study their inferential properties. In application we study causal effects of early life education on health of an older person. Our novel approach is the first to include aging phenomena in causal inference and has applications among other in economics and health. We consider extensions to more complex causal time-series structures and to elephant random walk possibilities.

Dynamic Treatment Regimes

Individual Treatment Effects in Bipolar Disorder with Latent Mood Dynamics Anna Pham*
Anna Pham, Amy Cochran, Melvin Mcinnis,

Individual-level experimentation is central to personalized decision-making, particularly for chronic conditions such as bipolar disorder, where treatment responses vary widely across individuals. In an ideal “n-of-1” setting, we would learn an individual’s treatment effect by repeatedly turning an intervention on and off and observing how outcomes respond. In practice, this is complicated by the fact that outcomes evolve dynamically and depend on prior outcomes, past interventions, and unobserved psychological states, creating carryover and temporal dependence that confound simple comparisons. We study the identification and estimation of individual treatment effects on mood in people with bipolar disorder, using a simulation environment grounded in parameters estimated from real patient data via a Kalman filter. Interventions are applied repeatedly over time and influence outcomes through latent state dynamics. We examine how longitudinal design choices affect the ability to recover accurate individual treatment effects and characterize how assumptions about mood dynamics and carryover shape estimation bias and variance. Our results provide practical guidance for individualized causal analysis in bipolar disorder and more broadly inform methods for personalized causal inference in dynamic settings.

Dynamic Treatment Regimes

Non-parametric Causal Inference in Dynamic Thresholding Designs Aditya Ghosh* Aditya Ghosh, Stefan Wager,

Consider a setting where we regularly monitor patients' fasting blood sugar, and declare them to have prediabetes (and encourage preventative care) if this number crosses a pre-specified threshold. The sharp, threshold-based treatment policy suggests that we should be able to estimate the long-term benefit of this preventative care by comparing the health trajectories of patients with blood sugar measurements right above and below the threshold. A naive regression-discontinuity analysis, however, is not applicable here, as it ignores the temporal dynamics of the problem where, e.g., a patient just below the threshold on one visit may become prediabetic (and receive treatment) following their next visit. Here, we study thresholding designs in general dynamic systems, and show that simple reduced-form characterizations remain available for a relevant causal target, namely a dynamic marginal policy effect at the treatment threshold. We develop a local-linear-regression approach for estimation and inference of this estimand, and demonstrate the promise of our approach in numerical experiments.

Dynamic Treatment Regimes

High-Dimensional Doubly Robust Inverse Probability Weighting for Dynamic Treatment Effects Feiyang Yi* Feiyang Yi, Jelena Bradic,

Inverse probability weighting is a common approach for estimating treatment effects. However, this method can yield potentially biased estimation when propensity score models are incorrectly specified. When confounders change over time and their dimensionality is much larger than the sample size, correctly specifying the propensity score models becomes more challenging, which can introduce substantial bias. This paper proposes a robust inverse probability weighting estimator for dynamic treatment effects that allows high-dimensional, time-varying confounding. We prove that the proposed estimator achieves root-N consistency and asymptotic normality under a sequential model double robustness condition, where at least one of the nuisance models is correctly specified at each treatment stage. Simulation studies illustrate the advantage of the proposed estimator compared with other inverse propensity methods.

Foundations

Causal Geodesy Kyle Schindl* Larry Wasserman, Kyle Schindl,

We introduce causal geodesy, a framework for studying the landscape of stochastic interventions that lie between the two extremes of performing no intervention, and performing a sharp intervention that sets an exposure equal to a specific value. We define this framework by constructing paths of distributions that smoothly interpolate between the treatment density and a point mass at the target intervention. Thus, each path starts at a purely observational (or correlational) quantity and moves into a counterfactual world. Of particular interest are paths that correspond to geodesics in some metric, i.e. the shortest path. We then consider the interpretation and estimation of the corresponding causal effects as we move along the path from correlation toward causation.

Generalizability/Transportability

Federated Causal Survival Analysis Under Distribution Shift Yi Liu* Yi Liu, Alexsander Levis, Ke Zhu, Shu Yang, Peter Gilbert, Larry Han,

Causal inference across multiple data sources can improve the generalizability and reproducibility of scientific findings. However, for time-to-event outcomes, data integration methods remain underdeveloped, especially when populations are heterogeneous and privacy constraints prevent direct data pooling. We propose a federated learning method for estimating target site-specific causal effects in multi-source survival settings. Our approach dynamically re-weights source contributions to correct for distributional shifts, while preserving privacy. Leveraging semiparametric efficiency theory, data-adaptive weighting and flexible machine learning, the method achieves both double robustness and efficiency improvement. Through simulations and two real data applications: (i) multi-site randomized trials of monoclonal antibodies for HIV-1 prevention among cisgender men and transgender persons in the United States, Brazil, Peru, and Switzerland, as well as women in sub-Saharan Africa, and (ii) an analysis of sex disparities across biomarker groups for all-cause mortality using the “flchain” dataset, we demonstrate the validity, efficiency gains, and practical utility of the approach. Our findings highlight the promise of federated methods for efficient, privacy-preserving causal survival analysis under distribution shift.

Generalizability/Transportability

Testing Generalizability in Causal Inference Daniel Manela* Daniel Manela, Linying Yang, Robin Evans,

Ensuring robust model performance in diverse real-world scenarios requires addressing generalizability across domains with covariate shifts. However, no formal procedure exists for statistically evaluating generalizability in machine learning algorithms.
Existing predictive metrics like mean squared error (MSE) help to quantify the relative performance between models, but do not directly answer whether a model can or cannot generalize.
To address this gap in the domain of causal inference, we propose a systematic framework for statistically evaluating the generalizability of high-dimensional causal inference models. Our approach uses the frugal parameterization to flexibly simulate from fully and semi-synthetic causal benchmarks, offering a comprehensive evaluation for both mean and distributional regression methods.
Grounded in real-world data, our method ensures more realistic evaluations, which is often missing in current work relying on simplified datasets.
Furthermore, using simulations and statistical testing, our framework is robust and avoids over-reliance on conventional metrics, providing statistical safeguards for decision making.

Generalizability/Transportability

Beyond Covariate Shift: Fusing Trial Data for Treatment Comparisons” to: “Trial Data Fusion: Indirect Comparisons Beyond Covariate Shift Kuan-Hung Yeh* Kuan-Hung Yeh, Ronghui Xu, Siddharth Singh,

Data fusion methods enable indirect treatment comparisons by combining evidence from randomized trials without head-to-head comparisons. Most existing approaches assume covariate shift, meaning that cross-trial differences are fully explained by observed covariates and that covariate–outcome relationships remain invariant across populations. This assumption may be oversimplified, as it ignores conditional shift, where these relationships differ between trials, potentially leading to biased treatment effect estimates in target populations.
We develop a unified data fusion framework that accommodates both covariate and conditional shift when two randomized trials share a common arm. In this setting, we establish identifiability of average treatment effects in a target trial population under a relaxed set of assumptions that allow for changes in both covariate and conditional outcome distributions. We introduce two estimators, a weighting-based and a resampling-based estimator recently proposed in literature, that leverage discriminative learning to estimate joint density ratios for both shifts. We evaluate their finite-sample performance through simulation studies. Results show that existing estimators can exhibit substantial bias in the presence of conditional shift, whereas proposed methods remain robust. An application to inflammatory bowel disease trials illustrates the practice of this framework in comparative effectiveness research when direct comparisons are unavailable.

Generalizability/Transportability

Population-Level Causal Effect Estimation in the Presence of Noncompliance: A Bayesian Approach Integrating RCT with Observational Studies Yueying Hu* Yueying Hu, Yajuan Si, Michael Elliott,

RCTs are the gold standard for causal inference but are often challenged by treatment noncompliance and limited generalizability when trial participants do not represent the target population. This work is motivated by the REFLUX trial, which compared laparoscopic fundoplication with medical management for gastroesophageal reflux disease (GERD). Of 810 enrolled patients, 453 expressed strong treatment preferences and were excluded from randomization but followed in a parallel preference arm. Both arms collected identical covariates and outcomes, while treatment uptake followed different mechanisms.
To jointly address noncompliance and generalizability, we combine the RCT and preference arm as a proxy for the target population and develop an integrative Bayesian framework for population-level complier average causal effect (CACE) estimation. Using the RCT data, we estimate a covariate-dependent latent compliance class model via principal stratification. This model is then used to infer compliance class membership for preference-arm participants from baseline covariates. Conditional on inferred classes and observed treatment uptake, outcomes from both arms are used to estimate population-level CACE by averaging class-specific treatment effects over the target covariate distribution. We evaluate the proposed methods through simulations and an application to REFLUX, and compare them with weighting-based alternatives in settings where outcomes may not be available for benchmark.

Generalizability/Transportability

Generalized projection tests for function-valued parameters with applications to testing structural causal assumptions Rui Wang* Rui Wang, Albert Osom, Bo Zhang,

Structural assumptions are central to the causal inference literature. In practice, it is often crucial to assess their validity or to test implications that follow from them. In many settings, such tests can be framed as evaluating whether a function-valued parameter equals zero. In this paper, we propose a class of generalized projection tests based on series estimators for testing such function-valued parameters. We establish conditions under which the proposed tests are valid and illustrate their applicability through examples from the data fusion and instrumental variables literature. Our approach accommodates flexible machine learning methods for estimating nuisance parameters. In contrast to many existing approaches, the limiting distribution of the proposed test statistics is straightforward to compute under the null hypothesis. We apply our method to test the equality of conditional COVID-19 risk across vaccine arms in the COVID-19 Variant Immunologic Landscape (COVAIL) trial.

Heterogeneous Treatment Effects

Tilted Intervention Effect and its Limiting Causal Estimand for Continuous Treatments
Yikun Zhang* Yikun Zhang, Yen-Chi Chen, Andrea Rotnitzky,

There is growing interest in the causal inference literature in defining causal estimands through stochastic, rather than static, interventions, motivated by both improved identifiability without strong positivity assumptions and greater flexibility in treatment assignment. In this paper, we study a causal estimand for continuous treatments induced by a class of stochastic interventions known as the tilted intervention, in which the estimand is identifiable without positivity. Specifically, our general tilted intervention framework unifies several constructions, including exponential, kernel-smoothed, Beta, Gamma, and weighted nearest neighbors tilting. The proposed estimand depends on a sensitivity parameter $hin [0,infty)$, allowing it to interpolate between the causal estimand under a static intervention and the mean outcome under no interventions. We further characterize the limiting behavior of the estimand under static interventions when positivity fails, yielding a generalized version of the G-computation formula with both causal and geometric interpretations. Finally, we develop estimation procedures for the tilted estimand and its limiting form, and establish their asymptotic properties.

Heterogeneous Treatment Effects

The Parachute Hybrid CATE Estimator and Bootstrap Methods: Theory and Applications
Xianlin Sun* Xianlin Sun, Stephen Man Sing Lee,

We propose a parachute hybrid estimator of the conditional average treatment effect (CATE) under the potential outcome framework. Itis a fully robust approach combining parametric (Meng and Qiao 2022) and nonparametric (Abrevaya, Hsu, and Lieli 2015) methods using the hybrid mechanism of Lee and Soleymani (2015) to estimate CATE. The key innovation is “graceful degradation”: when at least one of the propensity score or outcome models is correctly specified, it achieves parametric convergence rates—the defining property of double robustness. When both are misspecified, it remains consistent, with convergencedegraded to nonparametric rates—hence “parachute.” Its asymptotic distribution is derived by adopting M-estimation (Stefanski and Boos 2002) and incorporating Lyapunov’s Central Limit Theorem.
Bootstrap-based methods for statistical inference overcome plug-in variance estimation challenges. We prove that the generalized bootstrap method provides a consistent estimator for the distribution of the parachute hybrid estimator under general assumptions, grounded in Chatterjee and Bose (2005).

Heterogeneous Treatment Effects

CEEClust: A Bayesian Heterogeneous Time-Varying Causal Effect Model for Micro-Randomized Trials Brody Erlandson* Brody Erlandson, Tianchen Qian, Matt Koslovsky, Ander Wilson,

Micro-randomized trials (MRTs) are designed for developing and optimizing mobile health interventions. A primary estimand in MRT research is the causal excursion effect (CEE), which estimates the difference in outcomes that would result from following one excursion policy versus another over a given time period. Existing research has largely focused on population-level CEEs. However individuals often have different response dynamics, including variation in response time, effect magnitude, and adherence to the intervention. Capturing this heterogeneity is essential for understanding underlying behavioral mechanisms and to help inform the design of more adaptive and personalized interventions. We propose a new CEE model in the Bayesian paradigm that incorporates nonparametric priors to learn latent subgroups of individuals with similar time-varying treatment effects. Additionally, the proposed method provides uncertainty quantification for the population-level and cluster-specific CEE. Although the primary motivation are mobile health interventions, the proposed method could be applied with minimal modification to other large decision point time-varying treatment settings in which the CEE is of interest; such as, estimating the causal effect of heat warnings on reducing heat-related hospital visits. Lastly, we conduct simulation studies to evaluate the proposed method’s ability to recover heterogeneous response patterns and apply our approach to data from the HeartSteps MRT.

Heterogeneous Treatment Effects

Quantifying Treatment Effect Heterogeneity via the CATE Distribution Function Nolan Cole*
Nolan Cole, Marco Carone, Lars van der Laan,

Understanding how treatment effects are distributed across a population is essential for moving beyond mean summaries toward a richer characterization of treatment effect heterogeneity. The conditional average treatment effect (CATE) function is a fundamental object for studying such heterogeneity. Visualizing the distribution of CATE values in the target population (i.e., the CATE distribution) can reveal meaningful patterns of heterogeneity beyond the single metrics commonly used for this purpose. However, in a nonparametric model, the corresponding distribution function is an irregular parameter, precluding standard root-n inference.

In this work, we leverage the monotonicity of the distribution function to develop principled estimators that enable valid statistical inference. Specifically, we construct a Grenander-type estimator of the CATE distribution function and establish its large-sample properties, including consistency, double robustness, and cube-root convergence of its estimation error to a Chernoff limit distribution, which enables the construction of asymptotically valid confidence intervals. Simulation studies demonstrate that the proposed procedure has favorable finite-sample performance and can be readily used to obtain valid inference for quantiles of the CATE distribution.

Instrumental Variables

Marginal Causal Effect Estimation with Continuous Instrumental Variables Mei Dong* Mei Dong, Lin Liu, Dingke Tang, Geoffrey Liu, Wei Xu, Linbo Wang,

Instrumental variables (IVs) are often continuous, arising in diverse fields such as economics, epidemiology, and the social sciences. Existing approaches for continuous IVs typically impose strong parametric models or assume homogeneous treatment effects, while fully nonparametric methods may perform poorly in moderate- to high-dimensional covariate settings. We propose a new framework for identifying the average treatment effect with continuous IVs via conditional weighted average derivative effects. Using a conditional Riesz representer, our framework unifies continuous and categorical IVs. In this framework, the average treatment effect is typically overidentified, leading to a semiparametric observed-data model with a nontrivial tangent space. Characterizing this tangent space involves a delicate construction of a second-order parametric submodel, which, to the best of our knowledge, has not been standard practice in this literature. For estimation, building on an influence function in the semiparametric model that is also locally efficient within a submodel, we develop a locally efficient, triply robust, bounded, and easy-to-implement estimator. We apply our methods to an observational clinical study from the Princess Margaret Cancer Centre to examine the so-called obesity paradox in oncology, assessing the causal effect of excess body weight on two-year mortality among patients with non-small cell lung cancer.

Instrumental Variables

Estimation of a Common Local Average Treatment Effect with Multiple Instruments
Aniruddhan Ganesaraman* Aniruddhan Ganesaraman, Patrick Lopatto, P. M. Aronow,

Recently, Ghosh and Rothenhäusler (2025) have proposed an assumption-robust approach to causal inference for the average treatment effect (ATE) in the presence of multiple plausible adjustment sets. When it is unclear which candidate set satisfies ignorability, they construct a reweighted target population, as close as possible to the original one in KL divergence, such that the estimands obtained from the different adjustment sets agree. If at least one adjustment set is valid, this common estimand equals the ATE in the reweighted population, yielding a single asymptotically valid confidence interval for the reweighted-population ATE.

Building on the work of Ghosh and Rothenhäusler, we propose an assumption-robust approach to inference with multiple instrumental variables. Different instruments may identify local average treatment effects (LATEs) for distinct complier subpopulations, and disagreement among instrument-specific Wald estimates is generically attributable to treatment effect heterogeneity, instrument invalidity, or both. We propose to estimate the LATE in a reweighted population, which is constructed to guarantee causal identifiability when at least one of the instruments is valid. The corresponding estimator is doubly robust, $sqrt{n}$-consistent, and asymptotically normal. Our theoretical results are illustrated with a simulation study.

Instrumental Variables

Categorical Instrumental Variable Models: Characterization, Inference and Computation
Richard Guo* Richard Guo, Yilin Song, K. C. Gary Chan, Thomas Richardson,

The Minneapolis Domestic Violence Experiment, which evaluated police responses to domestic calls, is a classic example of a categorical instrumental variable (IV) model due to non-compliance with random assignment. In this setting, where the instrument, treatment, and outcome all take finitely many values, a general methodological framework for rigorous analysis has remained elusive, despite certain established results for binary IVs in the literature. For any categorical IV model, we derive a simple, closed-form characterization of the set of joint potential outcome distributions compatible with the observed data. We show that this characterization forms a system of non-redundant inequalities that unifies several IV models under varying independence and exclusion restriction assumptions. Building on this partial identification framework, we construct confidence intervals with simultaneous finite-sample coverage for linear functionals of the joint counterfactual distribution, such as pairwise average treatment effects, utilizing a tail bound on the Kullback-Leibler divergence. Finally, we develop specialized, highly efficient optimization algorithms to compute these intervals under the derived constraints. We demonstrate our method with a reanalysis of the Minneapolis data.

Interference and Consistency Violations

A Two-Stage Experiment Design for Causal Inference under Interference Mayleen Cortez-Rodriguez* Mayleen Cortez-Rodriguez, Christina Yu, Matthew Eichhorn,

Network interference is becoming increasingly relevant in our interconnected world. While most approaches to causal inference under interference rely on knowledge of the underlying network to make headway, recent work uses low-order potential outcomes models and a staggered rollout experimental design to obtain unbiased causal effect estimators without requiring network information. However, the required polynomial extrapolation can lead to prohibitively high variance. To address this, we propose a two-stage experiment that selects a sub-population in the first stage and restricts treatment rollout to this sub-population in the second stage. We prove theoretical guarantees for the bias and variance of a polynomial interpolation-style estimator under this design, showing improved performance even without network knowledge. For settings where the researcher may have access to some network knowledge, we also explore the role of clustering in the first stage. Bias increases with the number of edges cut in the clustering of the interference network, but variance depends on qualities of the clustering that relate to homophily and covariate balance. There is a tension between clustering objectives that minimize the number of cut edges versus those that maximize covariate balance across clusters, highlighting an interesting direction for future work.

Interference and Consistency Violations

Statistical Methods for Causal Effects of Multi-component Interventions in Longitudinal Observational Studies with Interference Ke Zhang* Ke Zhang, Ashley Buchanan, Laura Forastiere, Natallia Katenka, Donna Spiegelman, Collins lwuji,

Spillover effects arise when an intervention received by one unit affects the outcomes of units within a predefined group, referred to as an interference set. Such effects commonly occur in sociometric clusters/networks, such as HIV prevention programs, due to interactions with HIV transmission risk among individuals. HIV prevention programs are often delivered as intervention packages to reduce HIV incidence in the community, which is a combination of single components to prevent or treat HIV through multiple pathways simultaneously. Disentangling component-specific effects of intervention packages at the individual and community level is essential for fully understanding the effectiveness of HIV interventions and improving package interventions. However, existing causal methods for estimating spillover effects are typically limited to settings with a single intervention (whether static or time-varying) or to multiple interventions that are analyzed as a whole, thereby unable to provide insights into which components were driving (or hindering) the effectiveness. In this study, we develop novel causal methods for cluster randomized trials with time-varying exposure to intervention package components in the presence of interference and non-compliance. We expand partial interference assumption to temporal partial interference to account for interference in time-varying settings, use marginal structure models to estimate the marginal potential outcome for the intervention compo

Interference and Consistency Violations

Multiply Robust Estimators for Controlled Direct Effects in the Presence of Interference
Jimmy Kelliher* Jimmy Kelliher, Nandita Mitra,

Controlled direct effects are important causal estimands for public health scientists and policymakers interested in understanding the mechanisms by which a treatment causes an outcome. However, in both clinical trials and in observational settings, interference can pose a threat to identification and estimation. In this work, we extend the notion of an exposure mapping to that of a generalized counterfactual mapping, in order to accommodate interference in both exposure-outcome and mediator-outcome relationships under a difference-in-differences design. After establishing identification results, we further develop multiply robust, semi-parametric efficient estimators of the controlled direct effect when counterfactual mappings are correctly specified. We then assess the small-sample performance of these estimators in various simulation settings. Finally, we apply these methods to estimate the controlled direct effect of the 2017 Philadelphia beverage tax on the volume sales of sweetened beverages, which may be mediated by neighborhood-level price heterogeneity.

Interference and Consistency Violations

Learning Exposure Mapping Functions for Inferring Heterogeneous Peer Effects Shishir Adhikari* Shishir Adhikari, Sourav Medya, Elena Zheleva,

In networked settings, individuals are influenced by the actions or behaviors of peers, such as smoking habits or vaccination preferences, yet how this influence aggregates into a composite measure of peer exposure is unknown. Existing methods for estimating peer effects rely on hand-crafted exposure mappings, such as the proportion of treated peers, implicitly assuming simplistic and often unrealistic influence mechanisms. These mappings can be misspecified and lead to biased causal effect estimates. Our work moves away from explicitly defining an exposure mapping function and instead introduces a framework that learns this function automatically. We propose EgoNetGNN, a graph neural network for heterogeneous peer effect estimation that learns the exposure mapping directly from data. The model captures complex peer influence mechanisms involving not only peer treatments but also attributes of the local neighborhood, including node, edge, and structural features. We demonstrate that misspecified mappings and naive learning strategies lead to substantial bias, while EgoNetGNN adapts to diverse and complex influence mechanisms. Across a wide range of synthetic and semi-synthetic experiments, our approach consistently achieves lower estimation error than state-of-the-art baselines. Our results suggest a shift from designing exposure mappings to learning them, to enable more robust causal inference under interference across public health, social science, and economics.

Machine Learning and Causal Inference

Finite-sample near-equivalences between Targeted Maximum Likelihood and Double Machine Learning (Augmented IPW) Alejandro Schuler* Alejandro Schuler, David Bruns-Smith, Avi Feller,

Double machine learning (DML; a generalization of augmented inverse propensity score weighting) and Targeted Maximum Likelihood Estimation (TMLE) are two influence function-based estimators popular in causal inference, with extensive debates over the relative merits of each method. In this paper, we analyze their behavior in a large, well-studied class of estimands relevant to causal inference. We first review the known fact that a natural implementation of TMLE (TMLE with a "linear update") for such estimands has a simple, closed form expression which is a minor variation of the DML estimator. We then present a new result that the DML estimator can also be written as a minor variation of the TMLE estimator. This establishes a finite-sample near-equivalence between DML and linear-update TMLE where the two are related by a single scaling factor. By analyzing the scaling factor we show that TMLE generally debiases the naive plugin estimate more aggressively than DML at the cost of inflating the debiasing weights and incurring more variance. Our results show that, for these estimands, DML can be interpreted as a "regularized" form of TMLE using the linear submodel, and confirm that the choice of submodel for TMLE can substantially impact its performance.

Machine Learning and Causal Inference

Frugal, Flexible, Faithful: Causal Data Simulation via Frengression Linying Yang* Linying Yang, Robin Evans, Xinwei Shen,

Machine learning has revitalized causal inference by combining flexible models and principled estimators, yet robust benchmarking and evaluation remain challenging with real-world data. In this work, we introduce frengression, a deep generative realization of the frugal parameterization that models the joint distribution of covariates, treatments and outcomes around the causal margin of interest. Frengression provides accurate estimation and flexible, faithful simulation of multivariate, time-varying data; it also enables direct sampling from user-specified interventional distributions. Model consistency and extrapolation guarantees are established, with validation on real-world clinical trial data demonstrating frengression's practical utility. We envision this framework sparking new research into generative approaches for causal margin modelling.

Machine Learning and Causal Inference

Stable Causal Estimation with Transport Maps Yidan Xu* Yidan Xu, Yixin Wang, Long Nguyen,

This paper develops a causal estimation framework with Optimal Transport (OT) map for observational studies. Importance weighting based adjustment is known to have high sampling variance under lack of strong overlap, which could happen under high dimensional covariate space or small sample size. On the other hand, OT map proves to be a more stable method for tackling distribution shift, which can be applied to discrete distribution and is able to accommodate various data types.
Given these advantages, we propose a framework for causal estimation with OT map under superpopulation setting. With a binary treatment regime, we provide identifiability conditions and propose estimators for average treatment effects. To accommodate lack of overlap conditions, we extend the methodology with linear interpolation maps between treatment and control covariate distributions. Across different choices of distribution discrepancy and user selected upper bound, we demonstrate that the bias owing to lack of overlap can be reduced under outcome regression model misspecification.

Machine Learning and Causal Inference

Prediction Markets as Mechanisms for Price Discovery and Market Efficiency Kiet Le* Kiet Le, Khoa Le, Nghi Le,

Prediction markets have recently experienced rapid growth, yet their role in financial information production and market efficiency remains underexplored in the finance literature. This paper studies whether prediction markets improve price discovery for publicly traded firms and how they interact with traditional information intermediaries. We develop a framework in which informed agents have incentives to trade on prediction markets to monetize private information, potentially reallocating informed trading away from equity markets. Exploiting the quasi-exogenous listing of firm-specific corporate events on major prediction market platforms such as Polymarket and Kalshi, we implement a difference-in-differences, event-study design and AIPW method to identify causal effects. Using data on prediction market events linked to S&P 500 firms, combined with CRSP/Compustat stock data and IBES analyst forecasts, we examine changes in analyst coverage, forecast dispersion, abnormal returns around earnings announcements, and post-event price dynamics. Our results shed light on whether prediction markets crowd out or crowd in analyst activity, reduce insider trading incentives in equity markets, and enhance overall price efficiency. The findings contribute to the literature on information aggregation, market design, and the evolving role of alternative trading venues in modern financial markets.

Machine Learning and Causal Inference

Estimation and Inference for Causal Explainability Weihan Zhang* Weihan Zhang, Zijun Gao,

Understanding how much each variable contributes to an outcome is a central question across disciplines.
A causal view of explainability is favorable for its ability in uncovering underlying mechanisms and generalizing to new contexts.
Based on a family of causal explainability quantities, we develop methods for their estimation and inference.
In particular, we construct a one-step correction estimator using semi-parametric efficiency theory, which explicitly leverages the independence structure of variables to reduce the asymptotic variance.
For a null hypothesis on the boundary, i.e., zero explainability, we show its equivalence to Fisher's sharp null, which motivates a randomization-based inference procedure.
Finally, we illustrate the empirical efficacy of our approach through simulations as well as an immigration experiment dataset, where we investigate how features and their interactions shape public opinion toward admitting immigrants.

Machine Learning and Causal Inference

Causal Inference Post Identification Is Just Functional Estimation: A Critique and Counterexamples Abel Mesfin* Abel Mesfin, Elijah Tamarchenko, Katherine Keith, Rohit Bhattacharya,

The computation of causal parameters, often expressed in terms of potential outcomes, is a fundamental task in the empirical and social sciences.
Here, we provide a critique of the common adage: Once the identifying functional is obtained for a target causal parameter, causal inference reduces to a pure statistical or functional estimation problem. While this remains true in idealized scenarios, we highlight how common statistical and machine learning practices for dealing with real-world challenges in functional estimation—variable selection for high-dimensional settings and class weighting for handling class imbalance—can, in fact, be counterproductive for the purposes of causal effect estimation. In particular, we show that steps in functional estimation can sometimes increase variance or even introduce bias in causal effect estimates by effectively “undoing” the prior step of causal identification. We ground these arguments in well-known results from the literature on causal graphs and causal discovery, provide empirical evidence of our claims through simulations, and ultimately, synthesize these results into concrete recommendations for handling the bias-variance tradeoff in functional estimation while being mindful of causal identification.

Machine Learning and Causal Inference

Partial Identification with Unobserved Confounding Using the Rashomon Effect Srikar Katta* Srikar Katta, Jon Donnelly, Emanuele Borgonovo, Cynthia Rudin,

Many quantities of scientific interest often take the form of non–pathwise differentiable or higher-order functionals for which standard semiparametric inference tools fail. We introduce a general strategy for statistical inference based on Rashomon sets—the set of all machine learning models compatible with the data distribution. We provide uniform coverage guarantees even when the estimand is non-smooth and lacks an influence-function representation, often found in machine learning tasks focused on model auditing. Importantly, our inferential framework easily handles settings in which identifying assumptions are violated, such as no unobserved confounding. Through semi-synthetic experiments, we demonstrate that our bounds achieve nominal coverage and remain informative in realistic settings with unmeasured confounding. We conclude with an application to credit risk assessment, illustrating how our framework enables principled inference on feature relevance despite both model uncertainty and omitted variables.

Machine Learning and Causal Inference

Efficient and Flexible Heterogeneous Treatment Effect Estimation with Random BART Features Cory McCartan* Cory McCartan, Melody Huang,

Bayesian Additive Regression Trees (BART) models have shown promise for flexibly estimating causal response functions and heterogeneous treatment effects. However, they require custom Markov chain Monte Carlo (MCMC) sampling algorithms for computation, which limits their scalability and applicability to other model classes and data structures. We show how a recent reformulation of BART as an random-features approximation to a certain Gaussian Process can be applied to the estimation of heterogeneous treatment effects on large data in a variety of causal settings, including survival outcomes, differences-in-differences, and spatial data.

Machine Learning and Causal Inference

Precise Estimates from Safe Data Integration in a Paired, Cluster-Randomized Field Trial
Adam Sales* Adam Sales, Johann Gagnon-Bartsch, Mingyu Feng, Kevin Huang,

Randomized controlled trials (RCTs) in social or biomedical science often rely on covariate and outcome data drawn from a larger administrative database. For instance, a recent paired, clustered RCT of an online homework platform used covariates—demographics and prior achievement measures—and outcomes—standardized test scores—from the state's educational database. To avoid confounding, typical analyses use only data from randomized subjects, discarding the remainder of the database.
In this study, we show how to substantially improve the precision of effect estimates by using the entire dataset, including both randomized and non-randomized subjects, without introducing any new confounding and relying largely on standard modeling techniques, and demonstrate the method by re-analyzing data from the online homework study. We used observational data to train a random forest model predicting student test scores from student- and school-level covariates, and used it to generate predicted outcomes for students in the randomized schools. Then, we modified the hierarchical linear model from the original analysis by adding the predicted outcomes as an additional regressor, reducing the standard error by 15% — the improvement expected from increasing the sample size by over a third. Since it requires few additional modeling skills and no additional assumptions, this method can be easily adapted wherever auxiliary observational data are available.

Machine Learning and Causal Inference

Automated, Efficient, and Model-Free Covariate Adjustment under Stratified Randomization Raphael Kim* Raphael Kim, Michele Santacatterina, Ramin Zabih, Ivan Diaz,

Covariate adjustment and stratified randomization have shown to improve precision and power in clinical trials. Recently, methods have been
proposed to provide valid asymptotics for covariate adjustment using M-estimation under stratified randomization. However, leveraging efficiency gains with these methods require pre-specification of a small set of the covariates that are most predictive of the outcome, which is difficult in practice since most trials measure dozens of baseline covariates that are predictive of the outcome. In this work, we build on existing literature to propose an automatic, efficient, and model-free covariate adjustment method which permits data-driven covariate selection. Through extensive simulations and analysis, we showcase the simplicity and improved precision of our method when the covariate set is not known a priori.

Machine Learning and Causal Inference

Leveraging Large Language Models to Improve Precision in Randomized Controlled Trials
Jaylin Lowe* Jaylin Lowe, Adam Sales, Johann Gagnon-Bartsch,

Large language models (LLMs) are increasingly used in statistical research and applications. However, they are also notorious for unreliable or biased information. Here, we explore whether LLMs can be used to improve the precision of randomized controlled trials (RCTs) in a safe and rigorous way. Following similar work on leveraging observational data, we incorporate LLM predictions in an RCT analysis. While this method of improving precision is not new, the value of using LLM predictions in this manner is an open question. We discuss how useful LLM predictions are and how different datasets and prompts impact their usefulness.

LLM predictions add little value when the RCT already includes highly predictive covariates. However, if few such covariates exist or the data is well-suited for LLMs—like text—LLM predictions become more beneficial. Familiar, easy-to-predict outcome variables also help. Our basic approach asks the LLM to predict outcomes for each observation, but this often produces overly similar results. Instead, we ask the LLM to compare pairs of observations and predict which will have a higher outcome. We use the selection frequency as a covariate. We can also extract additional covariates from the LLM, such as writing quality or creativity in text-based RCTs. We combine all covariates to generate a final prediction for each observation, achieving greater precision than either the single prediction or standard covariate adjustment without the LLM predictions.

Machine Learning and Causal Inference

Doubly Robust Policy Learning for Multi-dimensional Stochastic Interventions through Auto-debiased Neural Networks Sylvia Cheng* Sylvia Cheng, Alejandro Schuler,

Designing a policy for assignment mechanisms of multiple continuous treatments (e.g. varying dosages of different drugs) creates high-dimensional complexity and remains challenging. Trial and real-world data often lack enough variability due to the exponential growth of the treatment-covariate space, which leads to positivity violations due to little support in rare covariate profiles.

As an alternative, we propose a method of learning an optimal multi-dimensional shift intervention policy by integrating Targeted Maximum Likelihood Estimation with neural nets. It learns a shift under the stochastic treatment regime to modify treatment values based on their natural observed values. Our framework is statistically rigorous to learn an optimal multi-dimensional shift and evaluate its debiased effect with semiparametrical efficiency and doubly robustness. Using a two-step data-splitting process and one compact neural net, it achieves promising run speed. Furthermore, it handles multi-dimensional shift learning via an influence-curve-based loss, optimizing the expected outcome while penalizing variance when deviating from the natural policy. It offers great flexibility for treatment designs and mitigates positivity violations in static regimes, as learned treatment–covariate combinations are supported and less likely to be sparse. We evaluated the method in simulations and showed that the learned shift parameters and causal effect converge to their population truth.

Machine Learning and Causal Inference

Efficient Subgroup Analysis via Optimal Trees with Global Parameter Fusion Zhongming Xie*
Zhongming Xie, Joseph Giorgio, Jingshen Wang,

Identifying and making statistical inferences on differential treatment effects—commonly known as subgroup analysis in clinical research—is central to precision health. Subgroup analysis allows practitioners to pinpoint populations for whom a treatment is especially beneficial or protective, thereby advancing targeted interventions. Tree-based recursive partitioning methods are widely used for subgroup analysis due to their interpretability. Nevertheless, these approaches encounter significant limitations, including suboptimal partitions induced by greedy heuristics and overfitting from locally estimated splits, especially under limited sample sizes. To address these limitations, we propose a fused optimal causal tree method that leverages mixed-integer optimization to facilitate precise subgroup identification. Our approach ensures globally optimal partitions and introduces a parameter-fusion constraint to facilitate information sharing across related subgroups. This design substantially improves subgroup discovery accuracy and enhances statistical efficiency. We provide theoretical guarantees by rigorously establishing out-of-sample risk bounds and comparing them with those of classical tree-based methods. Empirically, our method consistently outperforms popular baselines in simulations. Finally, we demonstrate its practical utility through a case study on the Health and Aging Brain Study–Health Disparities dataset, where our approach yields clinically meaningful insights.

Machine Learning and Causal Inference

Conditional Distributional Treatment Effects: Doubly Robust Estimation and Testing
Saksham Jain* Saksham Jain, Alex Luedtke,

Beyond conditional average treatment effects, treatments may impact the entire outcome distribution in covariate-dependent ways, for example, by altering the variance or tail risks for specific subpopulations. We propose a novel estimand to capture such conditional distributional treatment effects, and develop a doubly robust estimator that is minimax optimal in the local asymptotic sense. Using this, we develop a test for the global homogeneity of conditional potential outcome distributions that accommodates discrepancies beyond the maximum mean discrepancy (MMD), has provably valid type 1 error, and is consistent against fixed alternatives—the first test, to our knowledge, with such guarantees in this setting. Furthermore, we derive exact closed-form expressions for two natural discrepancies (including the MMD), and provide a computationally efficient, permutation-free algorithm for our test.

Machine Learning and Causal Inference

Identifying Recurring Payments In Financial Transaction Data Nate Bradshaw* Nate Bradshaw, Dylan Zwick, Joshua Jensen, Robert Ball,

Recurring payments for goods and services are a significant and growing source of expenses for individuals and institutions. While some payments arise from habitual behavior (e.g. daily coffee purchases), others are true recurring transactions such as subscriptions, utilities, and leases. Distinguishing between these patterns is essential for financial transparency but is challenging due to irregularities in real-world transaction data.

This work investigates methods for identifying recurring payments at the transaction level. We evaluate periodicity detection techniques, including Fourier transforms and convolution-based approaches, and establish a Z-score baseline based on the variance of time gaps between transactions. While this baseline achieves high recall (94.14%) on real-world data from Weber State University, its precision (28.39%) is limited by its inability to separate recurring signals from habitual noise.

To address this limitation, we engineer 25 features capturing temporal and transactional structure and train multiple machine learning models, including XGBoost and multi-layer perceptrons, using synthetic, real, and hybrid datasets. We show that augmenting real data with synthetic examples that model irregular habitual behavior improves precision while maintaining strong recall. Our results indicate that an XGBoost model trained on combined data provides the most effective balance, enabling more accurate identification of recurring expenses.

Machine Learning and Causal Inference

Kernel von Mises Formula of the Influence Function Yaroslav Mukhin* Yaroslav Mukhin,

The influence function (IF) of a statistical functional is the Riesz representer of its derivative, also known as its first variation and Fisher-Rao gradient. It is a key object for numerical optimization over probability measures, semiparametric efficiency theory, standard constructions of efficient estimators, and an arsenal of inference methods for these estimators. Yet, deriving the IF analytically is often an obstruction for practitioners. To automate this task, we develop a novel spectral representation of the IF that lends itself to a low-rank functional estimator in a reproducing kernel Hilbert space (rkHs). Our estimator (i) does not require analytic derivations by the user, (ii) relies on kernel Principal Component Analysis and numerical pathwise derivatives along these components. We present the derivation of the representation and prove consistency of the low-rank rkHs estimator.

Machine Learning and Causal Inference

Markov-Blanket-Guided training for Tabular Foundation Models Shu Wan* Shu Wan, Abhinav Gorantla, Kasim Selcuk Candan, Huan Liu,

The Markov blanket of a target variable constitutes the minimal and information-theoretically optimal feature set for prediction in a directed acyclic graph (DAG). Despite its central role in causal discovery and probabilistic graphical models, modern predictive systems rarely leverage Markov blanket structure explicitly during training. Meanwhile, tabular foundation models such as TabPFN introduce a new paradigm for supervised learning by training on large corpora of synthetically generated datasets derived from random DAGs. This generative perspective creates an opportunity to incorporate structural signals directly into the learning process.

In this work, we propose MB-Guide, a Markov-Blanket-Guided training strategy for tabular foundation models. Instead of treating all features symmetrically, MB-Guide uses the Markov blanket of the target variable, available by construction in synthetic DAG-based data generation, as a structural supervision signal during training. The model is encouraged to prioritize blanket variables while suppressing irrelevant ones, aligning representation learning with the theoretically optimal predictive set.

Empirical results demonstrate several desirable properties. First, MB-guided training improves computational efficiency by reducing effective feature redundancy during learning. Second, models trained with MB-Guide exhibit a strong ability to recover the target’s Markov blanket at inference time, enhancing structural interpretability.

Matching, Weighting

Balancing Act: Comparing Coarsened Exact Matching and Entropy Balancing in Cigna's 2025 Value of Integration Study Aran Canes* Aran Canes, Kamala Swayampakala, Lukas Halim, Robert Wojewoda,

Background: This study represents the 2025 edition of Cigna's Value of Integration (VOI) study, which evaluates the impact of integrated benefits on healthcare costs.
Methods: Using a large observational dataset with approximately 1.8 million treated customers and 95,642 controls, we compared Propensity Score (PS) stratification, Coarsened Exact Matching (CEM) and Entropy Balancing (EB) for ATT estimation. PS stratification was explored using 5, 10 and 20 strata but failed to achieve balance on key confounders at any stratification level.
Results: Pre-adjustment estimates indicated substantial confounding. CEM retained 83% of treated customers but excluded a subset with markedly higher outcomes leading to an ATT of $241. EB retained the full treated population at the cost of reduced ESS (83%) for the controls and produced an ATT of $501, closer to the naive difference. No evidence of dominance by individual controls was observed.
Conclusions: Divergent ATT estimates from CEM and EB reflected differences in overlap handling and estimand definition rather than estimator instability. EB preserved the full treated distribution and efficiently absorbed overlap constraints, while CEM produced a more conservative estimate applicable to an overlap-restricted treated subpopulation. Cigna elected to publish the CEM estimate since it reflects outcomes with strong covariate overlap. These results underscore the importance of aligning estimator choice with the target estimand.

Matching, Weighting

Doubly Robust Estimation of the Average Probabilistic Index Sarah Boese* Sarah Boese, Rui Wang, Tom Chen,

There are many trial settings when researchers want to report a marginal summary of treatment effect but targeting the Average Treatment Effect (ATE) is inappropriate. When outcomes are semi-continuous ordinal or composed of multiple endpoints with ordered levels of severity, like hierarchical composite endpoints, then rank-based comparisons are often tractable when treatment effect differences are not. Targeting the average probabilistic index (API), or the probability that a treatment unit does better than a control unit, is the most appropriate measure of treatment effectiveness in these scenarios. A locally semi-parametric efficient estimator for the probabilistic index has been established under a restricted moment model framework when outcomes for all treatment units are observed. In practice, outcomes are rarely fully observed. We develop a novel class of doubly robust estimators for the marginal probabilistic index under the missing at random framework and identify the semi-parametric efficient estimator from that class. We compare three estimators from our class: an inverse probability weighted estimator, a doubly robust estimator which assumes working independence and a locally semi-parametric efficient estimator under the semi-parametric transform model. We use machine learning tools like SuperLearner to estimate propensity score nuisance functional needed to fit our models.

Matching, Weighting

Achieving Covariate Balance in Infant RSV Prevention through Cardinality Matching with Multiple Treatment Options Lauren Liao* Lauren Liao, Karen Jacobson, Andrew Watson, Sally Stephens, Nicola Klein, Samuel Pimentel,

Treatment variation, such as different treatment options, often occurs in practice. A study that primarily aims to compare treatment versus control may also examine the effectiveness of different treatment options. For example, when comparing individuals receiving any treatment versus none, the study may initially focus on the overall treatment effect, ignoring variation among treatment options, while later analyses examine the effects of each option separately. Traditional matched designs targeting the overall comparison (any treatment vs. none) may fail to guarantee balanced comparisons for additional analyses involving different treatment options. We propose a single matched study design that ensures balanced comparisons for the overall comparison and separate treatment options between the treated and control subjects. We leverage and extend the cardinality matching approach to create covariate balance constraints for overall and separate treatment option comparisons and impose additional constraints to ensure each individual can only receive one treatment option. We demonstrate this method in a study of newborn infants at risk for respiratory syncytial virus (RSV), evaluating both the overall effect of receiving any RSV protection and the effects of distinct treatment options for protection (maternal vaccination or infant monoclonal antibody treatment).

Matching, Weighting

A Weighting Framework for Clusters as Confounders In Observational Studies Luke Keele* Luke Keele,

Units in observational studies are often clustered in groups, such as students in schools and patients in hospitals. Researchers then seek to adjust for (possibly unmeasured) cluster-level covariates and contexts that may influence both treatment assignment and outcomes. In this paper, we introduce a unified framework to evaluate the constraints assumed by estimation methods that adjust for cluster membership. We develop a weighting framework to show that different approaches differentially control global balance (differences between treated and control units across clusters) and local balance (differences within clusters). We first review traditional model-based inverse propensity score weighting (IPW) focusing on IPW with a hierarchical propensity score model, which is the current standard in the literature. We show that this approach does not impose any constraints on local balance. We then outline a more general balancing weight estimator that include constraints on global and local balance but uses regularization within these constraints. We next show that a form of the newly proposed Generalized Mundlak approach also fits into this framework, with model-based IP weights that adjust for cluster-level attributes rather than cluster indicators. We also propose a novel Mundlak balancing weights estimator, which is well-suited to this context and can be applied even if there are smaller clusters where all the units are treated or untreated. We then compare these methods in

Matching, Weighting

Which Covariates to Adjust for? Specification-robust Causal Inference in Observational Studies Aditya Ghosh* Aditya Ghosh, Dominik Rothenhäusler,

In observational causal inference, domain knowledge often leaves multiple covariate adjustments plausible, yet which sets satisfy ignorability is untestable. Different adjustment sets can yield conflicting estimates of the average treatment effect, and standard remedies (adjusting for their union or intersection, or reporting the union or convex hull of confidence intervals) can fail or produce intervals whose width does not vanish with sample size. We propose a specification-robust procedure that returns a single point estimate and a confidence interval that is valid as long as at least one candidate adjustment set is valid and has width shrinking at the fast, parametric rate. Our approach mirrors how trimming and overlap weighting handle overlap violations: We shift the target to a reweighted population, closest in KL-divergence to the original population, for which credible, specification-robust inference is feasible. We also provide diagnostic plots to assess the population shift and an extension to protect any function of the covariates used for reweighting, similar to calipers in matching. Synthetic and real-data examples demonstrate that our procedure provides substantially tighter confidence intervals than the convex hull while maintaining nominal coverage.

Mediation Analysis, Mechanisms

Semiparametric Inference for Causal Path-Specific Effects in Longitudinal Studies Xiaxian Ou* Xiaxian Ou, Razieh Nabi, Xinwei He,

We develop a semiparametric framework for estimating and conducting inference on causal path-specific effects in longitudinal studies involving multiple or repeatedly measured treatments, mediators, and confounders. The framework accommodates general longitudinal structures, including differing measurement intervals and varying numbers of mediators across time. Our framework focuses on estimation and inference for identifiable effects under the edge g-formula, allowing for binary, continuous, or multivariate mediators and certain patterns of unmeasured confounding among treatments, mediators, and the outcome. We propose multiply robust estimators derived from influence function theory that integrate data-adaptive machine learning techniques, and we establish rate conditions for their asymptotic linearity, efficiency, and robustness to nuisance misspecification. We further extend the framework to include a sensitivity analysis procedure that evaluates the impact of violations of cross-world ignorability assumptions. Simulation studies and an empirical application using the Framingham Heart Study demonstrate the method’s performance. An accompanying R package, flexPaths, was developed to facilitate implementation of the proposed methods.

Omitted Variable Bias

Multiple Regression Analysis of Unmeasured Confounding: Bounding Causal Effects by Reasoning about Randomness R. Mitchell Hughes* R. Mitchell Hughes, Brian Knaeble,

By reasoning about randomness, we can bound the uncertainty of a causal effect of interest due to omitted variable bias in a multiple regression setting. In previous work, we introduced a methodology for computing confounding intervals, enabling assessment of uncertainty due to unmeasured attributes. We have since generalized that methodology to multiple regression and developed an algorithm to partially identify a causal effect of interest in a multiple regression model when subject matter knowledge is available. Alternatively, the algorithm supports sensitivity analysis when such knowledge is absent. The strength of our approach lies in its use of coefficients of determination which allow us to calculate bounds by intuitively reasoning about randomness. We demonstrate our methodology in two example applications which highlight how the algorithm quantifies the robustness of the causal effect against omitted variable bias.

Policy Learning

Causal Discovery for Efficient Offline RL with Factored Action Spaces Cecilia Ehrlichman* Cecilia Ehrlichman, Shengpu Tang, Michael Dykstra, Maggie Makar,

Offline policy optimization is often sample-inefficient, especially when the action space is large, a problem that commonly arises in healthcare applications and multi-agent tasks. Many domains, however, admit a combinatorial action space, where sub-actions affect future states and rewards independently of one another. Past work either makes a priori assumptions about sub-action independence leading to efficient but potentially biased policy optimization, or fails to leverage potential independence, sacrificing sample efficiency. In contrast, we propose a two-step framework that leverages causal discovery for efficient policy optimization without introducing bias. Our approach (i) discovers the causal structure underlying the environment's dynamics from observational data, and (ii) exploits this structure to restrict the admissible policy class to a simpler, unbiased class. We provide theoretical guarantees characterizing settings under which our approach leads to efficient unbiased policy learning. Empirically, we demonstrate that our approach leads to more efficient policy optimization in settings with limited observational data, across both single-agent healthcare tasks and multi-agent settings.

Policy Learning

Optimal Policy Learning for Recurrent Outcomes via Instrumented Difference-in-Differences: An Application to T2DM Treatment Ritoban Kundu* Ritoban Kundu, Ashkan Ertefaie, Sean Hennessy, James Flory,

Learning reproducible, generalizable optimal treatment policies for chronic diseases requires large, representative populations observed over extended periods. While administrative health data offer an attractive foundation, their utility is often compromised by unmeasured confounding. We address this by proposing a novel framework based on Instrumented Difference-in-Differences (iDID) to estimate optimal policies for recurrent event outcomes subject to a terminating event. The iDID design is particularly advantageous as it relies on fewer and weaker assumptions than conventional instrumental variable or difference-in-differences methods. A key feature of our approach is that it explicitly addresses the fundamental challenge of avoiding policies that trivially reduce recurrent adverse events by increasing mortality, a common pitfall in policy learning for chronic disease settings. We develop a multiply robust estimator that remains consistent if any one of several subsets of nuisance models is correctly specified. Theoretical results establish the estimator's consistency and large-sample behavior. Simulations demonstrate that our estimator outperforms existing approaches in finite samples. We apply this method to a national Medicare dataset to optimize first-line Type 2 Diabetes strategies for minimizing disease-related hospitalizations.

Policy Learning

Nonparametric Estimation of Optimal Just-In-Time Adaptive Interventions for Distal Outcomes Jack Wolf* Jack Wolf, Nandita Mitrta, Ashkan Ertefaie,

Mobile and wearable technologies enable the delivery of just-in-time adaptive interventions (JITAIs)—interventions that adapt treatment delivery to an individual's rapidly changing internal state and context in real-time, real-world settings. However, existing methods for estimating optimal policies do not scale to the complexity of these designs and estimating optimal JITAIs remains challenging. In particular, JITAI settings typically involve dozens of decision points per individual, which cannot be handled using standard longitudinal causal inference methods. Advanced reinforcement learning approaches often optimize discounted sums of proximal outcomes and cannot support common questions in behavioral and clinical studies regarding end-of-study distal outcomes, which reflect long-term success rather than immediate effects. To address these challenges and align with scientific objectives, we make two methodological contributions. First, we develop a nonparametrically efficient inverse probability weighting approach for estimating optimal JITAIs for distal outcomes. Second, we introduce a data-driven policy tilting procedure that mitigates numerical positivity violations common in settings with a large number of decision points to improve finite-sample performance. We apply the proposed framework to Project MARS, a micro-randomized trial for smoking cessation that evaluated mobile health prompts recommending self-regulatory strategies to support quit attempts.

Proximal Causal Learning

A generalized front-door method when the mediator is confounded Helen Guo* Helen Guo, Beatrix Wen, Ilya Shpitser,

Unobserved confounding is a fundamental obstacle in causal inference problems. In the graphical modeling literature, a general theory has been developed that allows identification in the presence of hidden variables, with some limitations. In particular, Pearl’s celebrated front-door criterion only allows identification in the presence of treatment outcome unobserved confounding when a mediator variable exists that captures all causal influence from the treatment and outcome, and does not itself suffer from unobserved confounding.

In this paper, we propose a proximal generalization of the front-door criterion, allowing both arbitrary treatment/outcome confounding, and unobserved confounders of the mediator, provided informative proxies for the latter type of confounders are observed. In addition to deriving new identification strategies in this setting, we provide a Neyman orthogonal estimator for the resulting functional under one of these strategies with desirable efficiency properties, and evaluate its performance through simulations.

Randomized Designs and Analyses

Regression Adjustments for Double Randomization in Two-Sided Marketplaces Timothy Sudijono* Timothy Sudijono, Lihua Lei, Lorenzo Masoero, Suhas Vijaykumar, Guido Imbens, James McQueen,

Multiple randomization designs (MRDs) are a class of experimental designs used to handle interference in two-sided marketplaces. We investigate regression adjustment strategies for estimating total, spillover, and direct effects in MRDs. We derive minimum asymptotic variance estimators among a broad class of linearly adjusted estimators. Surprisingly, the optimal regression adjustments are estimable from data and are generally different from regression adjustments in classical randomized experiments. For example, one such optimal estimator for the direct effect corresponds to a weighted regression with interacted two-way fixed effects. We establish model-robustness properties, central limit theorems, and inferential methods for our estimators, relying on improved theoretical results for MRD experiments. Our results provide the analog of the CUPED regression adjustment for marketplace experiments. Numerical simulations demonstrate a considerable increase in efficiency over simpler approaches, enabling better inference when running MRDs.

Randomized Designs and Analyses

Optimal randomization-based FWER control Andy Chen* Andy Chen,

Randomization tests are finite-sample valid, model-lean, and well suited to moderate-sample analysis. In randomized trials, researchers are often presented with several potentially related hypotheses arising from multiple subgroups or multiple treatment levels. In this paper, we investigate family-wise error rate (FWER) control using individual randomization-test p-values.

Noting that sharp nulls are closed under intersection, we apply the closure principle for FWER control, constructing randomization tests directly for the induced intersection nulls, thereby avoiding the inefficiency of generic p-value aggregation. Combined with recent optimality results for a single randomization test, we show that the proposed procedure, when paired with most powerful single-test statistics, is optimal within the class of randomization-based FWER-control methods. Computationally, we develop branch-and-bound–type shortcuts and reduce the number of randomizations by sharing treatment reassignments across hypotheses.

Randomized Designs and Analyses

Efficient Statistical Estimation for Sequential Adaptive Experiments with Implications for Adaptive Designs Wenxin Zhang* Wenxin Zhang, Mark van der Laan,

Adaptive experimental designs are increasingly used in clinical trials and digital experiments, allowing treatment randomization probabilities to be updated based on sequentially accrued data, with objectives ranging from enhancing estimation efficiency to improving participant outcomes. However, the resulting dependence among observations induced by adaptive designs poses substantial challenges for efficient estimation and inference of causal estimands. Building on the Targeted Maximum Likelihood Estimation (TMLE) framework tailored for adaptive experiments, we introduce a new adaptive-design-likelihood-based TMLE (ADL-TMLE) for estimating a broad class of causal estimands from adaptively collected data, including the average treatment effect. We establish asymptotic normality and semiparametric efficiency of ADL-TMLE under relaxed positivity and design stabilization assumptions for adaptive experiments, while achieving improved finite-sample efficiency relative to prior TMLE approach that relies on inverse probability weighting of adaptive randomization probabilities. Simulation studies show that ADL-TMLE achieves substantial variance reduction across a range of adaptive experiments. Motivated by these results, we further propose a novel adaptive design that targets efficient estimation of causal estimands and outperforms standard efficiency-oriented adaptive designs. We further extend the proposed framework to broader settings including longitudinal structures.

Randomized Designs and Analyses

Causal Effects in Blinded Trials Jipcy Amador* Jipcy Amador, Michael Hudgens, Todd Schwartz,

In randomized clinical trials (RCTs), blinding (or masking) is commonly employed to reduce bias that may be introduced due to knowledge or presumptions that participants or investigators may have about the treatment being evaluated. However, in many settings it may also be of interest to evaluate the treatment effect in the absence of blinding to better predict the effect of treatment when used in the real world. In some settings understanding the effect of blinding may also be of intrinsic interest. This paper considers identification of treatment effects (in the presence or absence of blinding) and blinding effects (in the presence or absence of treatment) in blinded RCTs, under different assumptions about the treatment and blinding effects. In instances where the effect of interest is partially identified, sharp bounds are provided. Both single-blind and double-blind RCTs are considered. The methods are applied to: (i) a single-blind RCT evaluating the effect of silver diamine fluoride on caries arrest and prevention in children, and (ii) a double-blind RCT evaluating the effect of rituximab in individuals with myasthenia gravis.

Regression Discontinuity Designs

Nonparametric Regression Discontinuity Designs with Survival Outcomes Maximilian Schuessler* Maximilian Schuessler, Erik Sverdrup, Robert Tibsirani, Stefan Wager,

Quasi-experimental evaluations are critical for generating real-world causal evidence and complementing insights from randomized trials. The regression discontinuity design (RDD) is a quasi-experimental framework for estimating causal effects when treatment assignment depends on a running variable crossing a threshold. Such threshold-based rules are ubiquitous in healthcare, education, policy and beyond. However, standard RDD estimators rely on complete outcome data, an assumption often violated in time-to-event analyses such as in healthcare where censoring arises from loss to follow-up. To address this issue, we propose a nonparametric approach that leverages doubly robust censoring corrections and can be paired with existing RDD estimators. Our approach can handle multiple survival endpoints, long follow-up times, and covariate-dependent variation in survival and censoring. We discuss the relevance of our approach across biomedical applications and demonstrate its usefulness through simulations and the PLCO Cancer Screening Trial for prostate cancer, a US-based phase III clinical trial with right-censored survival outcomes. Our results show that our approach is more robust to misspecification and yields higher precision than censoring corrections based on inverse probability of censoring weighting. We have also developed an open-source software package rdsurvival that enables estimation with existing RDD approaches in the R language.

Semiparametric Inference

A Characterization of the Orthocomplement of the Tangent Space of Semiparametric Markov Models Trung Phung* Trung Phung, Ilya Shpitser,

Graphical models are ubiquitous in social and empirical science as they are intuitive and easy to use. These models belong to the broader class of Markov models, defined using solely conditional independence (CI) restrictions.

In order to estimate a finite-dimensional target parameter in such models efficiently, semi-parametric theory provides a principled framework for constructing regular and asymptotically linear estimators via influence functions (IFs). These estimators are asymptotically normal and root-n consistent. Characterizing the class of all influence functions for a target parameter is crucial for statistically efficient inference in these models.

For models that are Markov relative to directed acyclic graphs (DAGs), the orthogonal complement of the tangent space is known, implying that for any target the class of all influence functions can be derived once an influence function is obtained. On the other hand, for Markov models not equivalent to a DAG model—such as ordinary Markov models associated with undirected graphs, chain graphs, or acyclic directed mixed graphs—the orthogonal complement has not been characterized, impeding semi-parametric inference in these models.

We derive closed form expressions for the orthogonal complement of the tangent space for general Markov models and illustrate our results by characterizing the class of influence functions for the conditional mean parameter in several graphical models.

Sensitivity Analysis

Distributionally Equivalent Urns for the Truncation by Death Problem Jaffer Zaidi* Jaffer Zaidi,

The analysis of causal effects when the outcome of interest is possibly truncated by death has a long history in statistics and causal inference. The survivor average causal effect is commonly identified with more assumptions than those guaranteed by the design of a randomized clinical trial. This paper demonstrates that stochastic individual level causal effects in the ‘always survivor’ principal stratum can be identified and quantified with no stronger identification assumptions than randomization. Distributionally equivalent sufficient cause urns are defined and developed to quantify individual level ‘always survivor’ causal effects under truncation by death and censoring. Such urn models also enable sensitivity and multiverse analysis at the individual and population level. They also enable comparison of different identification strategies. We illustrate the practical utility of our methods using data from a randomized clinical trial on patients with prostate cancer. Our comprehensive methodology is the first and, as of yet, only proposed procedure that enables quantifying individual level causal effects in the presence of truncation by death and censoring using only the assumptions that are guaranteed by design of the clinical trial.

Sensitivity Analysis

Sensitivity to Attrition for Inferences from an RCT Benjamin Cher* Kenneth Frank, Kenneth Frank, Qinyun Lin,

Most randomized field experiments experience some attrition. Moreover, the extent of attrition may differ by treatment condition in systematic, non-random ways, biasing estimates of treatment effects and contributing to invalid inferences. We address concerns about non-random attrition by quantifying the conditions necessary in the attritted data to nullify an inference based on observed data. We do so non-parametrically by quantifying what must be the overall mean in the attritted data and the estimated treatment effect in the attritted data to nullify an inference. We also derive results based on a correlational framework, deriving what the correlation between any predictor and outcome must be in the attritted data to nullify an inference. In the empirical example, we show that the attritted students in the Tennessee Class Size study would have to have experienced a negative effect of small classes to nullify the inference that small classes have positive effects on achievement. While this provides no certainty about the inference, it does quantify the conditions necessary to nullify the inference, informing scientific interpretations and corresponding policy debate.

Sensitivity Analysis

Inferring Comprehensive Cohort Causal Effects in the Presence of Unmeasured Confounders and Missing Outcomes Shiyao Xu* Shiyao Xu, Razieh Nabi, Daniel Scharfstein,

Randomized control trials (RCTs) are considered the gold standard approach for estimating causal effect. However, RCTs often enroll patients who are not representative of a broader population. To address these limitations, we consider the comprehensive cohort study (CCS) design, where clinically eligible patients are first asked to enroll in an RCT, and if they decline, are asked to participant in a parallel observational study. Data on baseline covariates, treatments and outcomes are collected on all patients. In this paper, we present a methodological framework for estimating the comprehensive cohort causal effect (CCCE) – the difference in mean potential outcomes had all patients in the CCS received treatment A vs. treatment B, in the presence of unmeasured confounding in the observational arm (handled via sensitivity analysis) and outcome missingness (assumed to be missing at random). We apply our methods to the TOIB study, a CCS to determine the effect of topical versus oral non-steroidal anti-inflammatory drugs (NSAIDs) in managing knee pain among older adults with chronic knee pain. We also conduct a simulation to evaluate the performance of our approach.

Sensitivity Analysis

Identification Limits of Proximal Inference: Sharp Closed-Form Bounds Guilherme Duarte* Guilherme Duarte,

Proximal causal inference exploits proxy variables to address unobserved confounding, but existing results largely focus on point identification under strong completeness or functional assumptions. This paper studies the identification limits of proximal inference when only a unique proxy is available and shows that, in this setting, the average treatment effect (ATE) is generally partially identified. We derive the first sharp bounds for the ATE in this class of proximal models and show that the bounds admit closed-form expressions. Sharpness is established by explicitly characterizing all observationally equivalent data-generating processes consistent with the proximal assumptions. The analysis reveals that the identified set arises from slackness in an associated cubic program, providing a transparent geometric interpretation of the failure of point identification. We further characterize the magnitude of the bias induced by adjusting for proxies rather than the latent confounder itself, thereby quantifying the limits of proxy-based adjustment. These results clarify the informational content of proximal assumptions and position partial identification as an inherent feature of proximal causal inference with limited proxy information.

Sensitivity Analysis

Using large language models for sensitivity analysis in causal inference: cases studies on Cornfield inequalities and E-values Qingyan Xiang* Qingyan Xiang, Jiahao Zhang, Bojian Feng,

Unmeasured confounding is a central challenge in causal inference from observational studies. Sensitivity analysis methods such as Cornfield inequalities and E-values assess robustness to unmeasured confounding, but are often difficult for interdisciplinary researchers to compute and interpret. Recent advances in large language models (LLMs) offer accessible tools to support sensitivity analyses, yet their reliability has not been evaluated. We assess four LLMs (ChatGPT, Claude, DeepSeek, and Gemini) using case studies from smoking, back pain, Alzheimer's disease, and environmental health research. Performance is evaluated by (1) E-value calculation accuracy, (2) qualitative interpretation of robustness to unmeasured confounding, and (3) identification of plausible unmeasured confounders. ChatGPT, Claude, and Gemini accurately reproduce reported E-values, whereas DeepSeek shows small biases. All models generate conclusions consistent with effect sizes and identify biologically plausible unmeasured confounders. To our knowledge, this is the first work using cases studies to evaluate the performance of LLMs on sensitivity analysis. The results suggest that structured prompting enables LLMs to support sensitivity analysis, which can fruther inform researchers to improve their study design and decision-making in observational studies.

Sensitivity Analysis

Inferential impacts of spatial confounding in national-scale air pollution health analyses
James Celi Kitch* James Celi Kitch, Sophie Woodward, Danielle Braun, Michelle Bell, Francesca Dominici, Daniel Mork,

Objective: The health burden attributable to Alzheimer's disease and related dementias (ADRD) is expected to increase substantially in the coming decades. Mitigating this increase requires understanding factors that influence ADRD progression, and recent national-scale studies have begun to address this. However, existing approaches do not sufficiently incorporate spatial information to address unobserved spatial confounding. Difficulties in developing realistic simulation studies further complicate evaluation of inferential impacts from unobserved spatial confounding.

Methods: We leverage a novel causal inference benchmarking tool to generate semi-synthetic datasets from real, nationwide health data. On this data, we evaluate seven statistical models in their ability to estimate the Exposure-Response Curve (ERC) between fine particulate matter (PM2.5) exposure and ADRD-related hospitalizations, measuring model bias and coverage of the true ERC. We assess the impact of spatial features by masking spatially autocorrelated variables, emulating settings of unobserved spatial confounding.

Results: Models that consider spatial proximity are more robust to unobserved spatial confounding, supporting increased use in national-scale epidemiology studies. We also find that commonly used inferential procedures, such as an m-out-of-n block bootstrap, are insufficient for spatial confounding. Finally, we apply these methods to estimate the PM2.5–ADRD relationship in Medicare data.

Sensitivity Analysis

Bounding Disparities under Selective Reporting Elsa Palumbo* Elsa Palumbo, Edward Kennedy, Leah Jacobs,

Understanding disparities in outcomes across subpopulations is a central problem in the social sciences. However, accurately quantifying the magnitude of a disparity is challenging when the observed data are unrepresentative. The case of selective reporting is especially difficult, since standard biased sampling tools do not apply. In this paper, we develop a nonparametric sensitivity model to address this challenge. We derive tight bounds on the conditional mean outcome for each treatment group and, from those, bounds on the covariate-adjusted mean outcome as well. Using this result, we provide doubly robust estimators for the covariate-adjusted bounds, assuming the margin condition in order to handle the non-smoothness of our target parameters. Finally, we implement our sensitivity analysis framework on traffic stop data from the Racial and Identity Profiling Act (RIPA), bounding the relative risk of police use of force against civilians with and without perceived mental illness. At 5% missing records, we predict a three to four times higher risk for those with perceived mental illness. The existence of a disparity is robust to substantial missingness, allowing up to 35% unrecorded traffic stops.

Sensitivity Analysis

Covariate Measurement Errors as An Omitted Variable Bias Problem Minh Duy Pham* Minh Duy Pham, Chad Hazlett,

A necessary but rarely stated assumption in causal inference is that all variables, especially the covariates we condition on to identify the causal effects, are measured without error. However, in reality, measurement errors are endemic, and we often adjust for error-prone proxies or measurements of the true confounders instead. In this work, we approach the problems of covariate measurement errors as one of residual unmeasured confounding: conditioning on an error-prone proxy of a confounder will most likely not fully adjust for its confounding effects on the treatment and the outcome, leaving (a portion of) it unobserved. If so, how unreliable must our measurements be for the residual confounding to ruin our treatment effect estimates? We extend and reparameterize the omitted variable bias approach in Cinelli and Hazlett (2020) to describe how measurement (un)reliability influences the bias in an intuitive and interpretable manner. In doing so, we highlight the merits of and recommend a novel approach of applying sensitivity analysis to covariate measurement error problems.

Weighting

Generalized Entropy Calibration for Inference with Partially Observed Data: A Unified Framework Mst Moushumi Pervin* Mst Moushumi Pervin, Hengfang Wang, Jae Kwang Kim,

Missing data is an universal problem in statistics. We develop a unified framework for estimating parameters defined by general estimating equations under a missing-at-random (MAR) mechanism, based on generalized entropy calibration weighting. We construct weights by minimizing a convex entropy subject to (i) balancing constraints on a data-adaptive calibration function, estimated using flexible machine-learning predictors with cross-fitting, and (ii) a debiasing constraint involving the fitted propensity score (PS) model. The resulting estimator is doubly robust, remaining consistent if either the outcome regression (OR) or the PS model is correctly specified, and attains the semiparametric efficiency bound when both models are correctly specified. Our formulation encompasses classical inverse probability weighting (IPW) and augmented IPW (AIPW) as special cases and accommodates a broad class of entropy functions. We illustrate the versatility of the approach in three important settings: semi-supervised learning with unlabeled outcomes, regression analysis with missing covariates, and causal effect estimation in observational studies. Extensive simulation studies and real-data applications demonstrate that the proposed estimators achieve greater efficiency and numerical stability than existing methods. In particular, the proposed estimator outperforms the classical AIPW estimator under the OR model misspecification.

Applicants in Social Sciences

The Causal Effect of Volatility: Estimation by Marginal Structural Models Ian Lundberg* Ian Lundberg, Nanum Jeon, Hao Liang,

We apply marginal structural models to a social science question about the causal effect of economic volatility. Our question is motivated by social science literature showing a major shift in the U.S. labor market. While it was historically common for a U.S. worker to experience a long-term, stable careers with a single employer (e.g. in the 1950s), the labor market experiences of present-day workers are more often characterized by frequent job changes. Thus the life course trajectory of economic well-being often involves more rises and falls today than in the past. We would like to understand how economic volatility at ages 25–34 affects the probability of being married at age 35. Because volatility is a treatment that unfolds across time periods, we use marginal structural models. But standard MSMs that focus on the cumulative treatment cannot answer our applied question: a trajectory that rises and falls repeatedly is more volatile than one that is consistently average, even though the cumulative treatment is the same. We therefore generalize the usual functional form of MSMs to directly incorporate the volatility of the treatment, measured by a summary of year-to-year changes. We produce estimates using panel data on a probability sample of U.S. adults. Broadly, our study illustrates one example where generalizations of cumulative treatment MSMs can be useful in a new class of social science applications.

Applicants in Social Sciences

A Bayesian State-Space Approach with Dynamic Covariates for Disentangling Anticipatory and Intervention Effects Damiano Baldaccini* Damiano Baldaccini, Alessandra Mattei, Fiammetta Menchetti,

Evaluating public policies is particularly challenging when their nationwide scope precludes the existence of a suitable control group and when advance announcements generate anticipatory effects that must be disentangled from the effects of the policy itself. We address these challenges within the potential outcomes framework. We formally define the causal estimands of interest, explicitly distinguishing between the effects of the policy announcement and those of the policy implementation. We then introduce a set of identifying assumptions under which these causal effects can be estimated using a Bayesian state-space model that exploits the time-series structure of the data. The proposed model is estimated over the full time horizon and incorporates a dynamic treatment variable that captures the exogenous shocks induced by both the policy announcement and the subsequent intervention. Through a series of simulation studies, we assess the performance of our approach and compare it with standard counterfactual forecasting methods for causal inference in the absence of a control group. Finally, we apply the proposed methodology to evaluate an Italian transportation policy that incentives green vehicles and penalizes polluting ones.

Applicants in Social Sciences

A Platform for Personalized Financial Interventions Simon Smith* Simon Smith, Brian Knaeble,

Consumers struggle to make and follow realistic budgets. Yet traditional budgeting platforms (such as the top 10 Forbes' budgeting apps of 2026) categorize expenses by vendor, hiding the specific products which make up the total cost. This poster introduces a platform to tackle this over-generalization by recording expenses at the product level: scanning receipts and storing itemized product and price information. The platform's product-level schema helps consumers make informed financial decisions and provides detail-rich input for the development of causal AI. Visitors to this poster can expect to interact with the technology and discuss how it fits into larger models of consumer financial behavior.

Applicants in Social Sciences

Surrogate Representation Inference for Text and Image Annotations Kentaro Nakamura* Kentaro Nakamura,

As researchers increasingly rely on machine learning models and LLMs to annotate unstructured data, such as texts or images, various approaches have been proposed to correct bias in downstream statistical analysis. However, existing methods tend to yield large standard errors and require some error-free human annotation. In this paper, I introduce Surrogate Representation Inference (SRI), which assumes that unstructured data fully mediate the relationship between human annotations and structured variables. The assumption is guaranteed by design provided that human coders rely only on unstructured data for annotation. Under this setting, I propose a neural network architecture that learns a low-dimensional representation of unstructured data such that the surrogate assumption remains to be satisfied. When multiple human annotations are available, SRI can be extended to further correct non-differential measurement errors that may exist in human annotations. Focusing on text-as-outcome settings, I formally establish the identification conditions and semiparametric efficient estimation strategies that enable learning and leveraging such a low-dimensional representation. Simulation studies and a real-world application demonstrate that SRI reduces standard errors by over 50% when machine learning classification accuracy is moderate and provides valid inference even when human annotations contain non-differential measurement errors.

Applications in Health and Biology

The Self-Masking Model for Imputing Missing Electronic Health Record Data Yidan Zhang*
Yidan Zhang, Eric Slud, Razieh Nabi, Daniel Scharfstein,

In the early hours of an emergency department encounter, laboratory tests are selectively ordered for patients who are suspected of having abnormal findings. As a result, laboratory measurements recorded in the electronic health record (EHR) are often subject to informative missingness, since the absence of a test may itself convey clinical information. This missing not at random (MNAR) process poses substantial challenges for downstream analyses that require complete laboratory profiles. To address this problem, we focus on binary laboratory variablescoded as normal/abnormal and develop a missing data imputation scheme under the followingassumptions: (1) the probability a laboratory value is missing depends only on the underlying(possibly unobserved) value of that variable, and (2) the joint distribution of laboratory resultsarises from a latent multivariate probit model, which captures dependence across laboratoriesthrough correlated latent Gaussian variables. We estimate the model parameters using an EM algorithm applied to a pairwise composite likelihood and then impute the missing laboratory values via MCMC sampling with adaptive tuning. We illustrate the proposed method using EHR data from a cohort of patients with suspected sepsis presenting to emergency departments within the Intermountain Healthcare system.

Applications in Health and Biology

A Protocol for Comparing the Causal Impact of Pre- and Perinatal Factors on Autism Rachel Hanger* Rachel Hanger, Amy Cochran, Olivia Pokoski, Sarah Furnier, Maureen Durkin, Camara Gregory, Dadit Hidayat,

Over the past several decades, autism has been pushed to the forefront of the medical zeitgeist. Recent government policy emphasizes identification of contributing factors to autism through efforts such as the Autism CARES Act of 2024. However, it is both hard to identify such factors and measure their impact due to limitations of observational studies. An autistic child cannot have a perinatal factor retroactively changed and then be reevaluated for autism. In this work, we present a protocol for analyzing case-control data from a multi-site study of autism. We use Bayesian Additive Regression Trees (BARTs) to calculate two important metrics for candidate pre- and perinatal factors: the individual causal relative risk (the relative change in an individual's chance of having autism if one factor's value is changed) and the population causal relative risk (the relative change in the prevalence of autism in general if the entire population had the same factor value changed). We carefully justify the assumptions required to identify these causal effects from case-control data. In addition, we describe how to build on these analyses so that we can attribute specific percentages of the likelihood of having autism to each factor in both individualized and generalized public health contexts, and present a guideline for future applications and analyses, medical or not.

Applications in Health and Biology

Identifying Misreporting Rates in the Absence of Ground Truth Data Dylan Zapzalka* Dylan Zapzalka, Muskaan Mittal, Jenna Wiens, Maggie Makar,

Strategic agents are often incentivized to misreport their features to obtain favorable outcomes from machine learning models. While prior research utilizes causal inference to estimate misreporting rates for binary features, these existing methods rely on the restrictive assumption of having access to a ground truth dataset. In this work, we relax this requirement by leveraging two datasets with directional misreporting: one where agents misreport features in only one direction, and another where they only misreport in the opposite direction. We give the conditions under which the misreporting rate is identifiable using causal effect estimation by integrating these two complementary data sources. For scenarios where exact identification is not possible, we provide sensitivity analysis bounds for the misreporting rate. Finally, we empirically validate our theoretical findings using both semi-synthetic data and a real-world Medicare dataset, demonstrating the practical utility of our method.

Applications in Health and Biology

Handling Informative Missing Data in Electronic Health Records: Imputation with the Semiparametric Gaussian Copula Model Yongjun Lee* Yongjun Lee, Anna Guo, Daniel Scharfstein, Razieh Nabi,

Emulating a clinical trial based on electronic health records requires careful adjustment for pre-treatment confounders. However, these variables are often subject to informative missingness. Assumptions about missing data are untestable.
Missing data directed acyclic graphs (mDAGs) provide a framework to represent these assumptions and can be used to evaluate whether the conditional distribution of the missing data given the observed data is identified. If identified, this conditional distribution can be used to impute the missing data.
In this paper, we propose a novel imputation framework based on the Semiparametric Gaussian Copula model (SGCM), built on a latent multivariate Gaussian structure that is linked to the variables of interest through monotonic increasing transformations for continuous variables and a set of cut points for ordered categorical variables.
We illustrate our approach to impute missing confounders in a target trial emulation designed to evaluate treatments for patients with non-small cell lung cancer based on data from the Flatiron Health electronic health records database. We also conduct a simulation study to investigate the performance of our approach.

Applications in Health and Biology

Causal Inference for Noisy Sequence Count Data Tinghua Chen* Tinghua Chen, Justin Silverman,

Modern sequencing technologies, including microbiome profiling, single-cell RNA sequencing, and bulk gene expression assays, generate high-dimensional count data that serve as noisy, indirect measurements of latent biological quantities such as microbial abundance or gene expression. Analyses of these data are challenged by compositional constraints, sparsity, sampling variability, and pervasive confounding. Despite this, existing causal inference methods typically define potential outcomes on the observed count scale, even though treatments act on latent biological states rather than on stochastic, technology-dependent counts, resulting in causal effects that lack clear biological interpretation. We propose a causal inference framework that explicitly targets latent potential outcomes underlying sequencing data by jointly modeling latent biological quantities and the measurement processes that generate observed counts. This approach enables causal estimands to be defined on the latent biological scale, characterizes identifiability and uncertainty under realistic sequencing models, and supports scalable estimation of latent average treatment effects using a combination of parametric and flexible nonparametric components. The proposed methods are evaluated through simulations and applications to microbiome, single-cell, and gene expression datasets, providing a principled foundation for biologically interpretable causal inference from sequencing data.

Applications in Health and Biology

When Do Observational Causal Conclusions Survive? Evidence from Patient Access to Medical Records Minh Nguyen* Minh Nguyen,

Causal claims in healthcare often rest on observational associations whose credibility depends more on identifying assumptions than on statistical precision. We study whether access to personal medical records increases the probability that individuals use information in healthcare decisions, treating the application as a laboratory for systematic stress-testing rather than single-model confirmation. Across outcome regression, weighting, and doubly robust estimators, access is associated with increases of roughly 15–20 percentage points. Alternative adjustment sets, learner choices, and subsampling perturb magnitudes but rarely direction. Strengthening overlap attenuates effects without eliminating them. Placebo and negative-control exercises yield estimates near zero, and perceived usefulness remains associated with access even among respondents who have not acted. No diagnostic is decisive alone, but together they substantially narrow simple null explanations. At the same time, an asymmetry emerges: while the average effect is stable, mechanisms and individual prioritization are not. Specifications that leave the mean largely intact often reorganize who appears to benefit. The results suggest that observational evidence can be informative about whether access matters, yet remain underdetermined about why or for whom. We offer both new evidence on behavioral responses to information access and a framework for evaluating credibility across multiple dimensions of inferential f

Applications in Health and Biology

Sparse Group LASSO for Causal Network Discovery in High-Dimensional Multivariate Time Series: An Application to Swine Disease Surveillance Zhengyuan Zhu* Zhengyuan Zhu, Alan Moore, Lynna Chu,

Learning causal relationships in high-dimensional multivariate time series is essential for understanding complex dynamical systems. In many applications, variables naturally form overlapping groups reflecting shared structure, such as spatial regions or related processes. Standard causal discovery methods often ignore this structure, which can reduce stability and statistical power when the number of possible interactions is large.

We propose a causal discovery framework based on sparse group LASSO that incorporates structured sparsity in a vector autoregressive model with lagged causal effects. The method encourages similar sparsity patterns across related groups while maintaining overall sparsity of causal links. A two-stage procedure uses sparse group LASSO to screen candidate links and ordinary least squares to estimate causal strengths.

We illustrate the method using swine disease surveillance data consisting of weekly case counts for multiple viruses across multiple U.S. states. Some viruses may precede or trigger outbreaks of others, and outbreaks in one state may lead to outbreaks in other states. Identifying such cross-virus and cross-location causal relationships can help improve outbreak monitoring. The learned causal network provides information that can support earlier detection of emerging outbreaks and reduce detection delay.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

Stochastic interventions for studying the health effects of environmental mixtures

Zhuochao Huang* Zhuochao Huang, Antonelli Joseph,

Evaluating the causal health effects of multivariate continuous exposure mixtures, such as air pollutants, is a critical public health challenge. A primary obstacle is the frequent violation of the positivity assumption, which renders the effects of standard deterministic interventions unidentified without unreliable extrapolation. In this paper, we develop a novel causal inference framework to address this challenge. We extend exponential tilting to multivariate exposures and address the critical question of how to compare different intervention directions fairly. This establishes a systematic framework for defining and evaluating various policy-relevant causal estimands, allowing researchers to address diverse scientific questions. We develop numerous methodological advancements, including efficient one-step estimation strategies, a Riemannian BFGS algorithm to solve a constrained manifold optimization problem, semiparametric efficiency bounds for causal estimands, minimax rates for estimators, and establishing asymptotic normality. We demonstrate our framework’s utility by applying it to a large-scale, real-world environmental health dataset to identify the optimal strategy for reducing adverse health outcomes associated with a PM2.5 chemical mixture.

Bayesian Causal Inference

When IRT Point Scores Stand in for Latent Confounders: How Test Information Shapes Bias and Interval Validity in Causal Adjustment Weiran Li* Weiran Li, Bruno Zumbo, Xiangyi Liao,

Applied evaluations often adjust for baseline differences using test based scores even when treatment uptake is driven by an unobserved latent variable that also predicts outcomes. Let theta denote this baseline latent confounder. Because theta is unobserved, analysts often plug in an IRT point score, typically the EAP, and treat it as error free. Fully joint IRT measurement, treatment, and outcome models can address latent confounding, but they are often hard to specify, fit, and communicate in routine evaluations, creating a gap between recommended joint modeling and common workflows. We link instrument information and targeting to both bias and interval validity of causal adjustment with IRT point scores. We simulate data from a coherent model with 2PL measurement, logistic treatment selection driven by theta, and a linear outcome with a nonzero treatment effect, varying test length, discrimination, and difficulty targeting. Unadjusted analyses show large bias, about 0.26 to 0.28, and near zero nominal 95 percent coverage. Plug in EAP adjustment reduces bias under high information, but interval validity depends sharply on where the test is informative: under low information, coverage collapses to about 0.04 to 0.08, and even under higher information coverage remains below nominal at about 0.69 to 0.81. These results motivate uncertainty propagation via posterior draws and principled pooling while retaining existing analytic workflows.

Bayesian Causal Inference

Transporting Principal Causal Effects Across Strata Veronica Ballerini* Veronica Ballerini, Francesca Dominici,

In mediation analysis, decomposing treatment effects into natural direct and indirect effects relies on cross-world independence assumptions involving a priori counterfactuals. Principal stratification instead defines causal effects within principal strata (PS) of joint potential mediator values, avoiding such assumptions, but how to decompose effects within PS into direct and mediated components remains unclear. Direct effects are identifiable only in PS where, by definition, there is no mediated effect, and no general framework exists for separating the two components elsewhere without strong assumptions. Building on recent works on transportability, we introduce a formal approach for transporting direct principal effects across PS. We give identifying assumptions enabling full or partial transportability and tests for mediated effects. The peculiarity with respect to the literature on transportability is that PS are “latent;” the effects are only weakly identifiable. We address this with a Bayesian approach that does not require full identification and propagates the PS membership’ uncertainty. We illustrate our method using Medicare data on over 30 million beneficiaries, integrating claims and high-resolution PM2.5 exposure.

Bayesian Causal Inference

Why Even Bayesians Need to Worry about Multiple Comparisons George Perrett* George Perrett, Jennifer Hill, Marc Scott, Christopher Buglino,

Researchers asking causal questions are often interested not only in the average treatment effect but also subgroup specific treatment effects that allow for more nuanced understanding of who benefits from an intervention. However, this pursuit can lead to issues with multiple comparisons. While previous research has demonstrated that Bayesian methods with regularizing prior distributions are more conservative than their frequentist counterparts and can lead to better outcomes than either ignoring the issue or using corrections (Bonferroni, FDR), the extent to which Bayesian methods eliminate the problem of multiple comparisons has been overstated. Critically, we demonstrate a setting common in social sciences where Bayesian regularizing priors are not sufficient to control false positive claims. Moreover, we show that this is not only limited to false positive claims but extends to sign errors. We characterize this setting as dominated by “shrinkage to the wrong place” and present a diagnostic to guide researchers to more appropriate models.

Bayesian Causal Inference

Early Stopping for Time-Varying Treatment Effects Sam van Meer* Sam van Meer, Alberto Abadie,

When treatment effects evolve over time, early estimates can be misleading about long run program effectiveness. We model the trajectory of treatment effects with a semi-local linear trend model, where treatment estimates are noisy measurements of this latent process. Using filtering methods, we construct sequential confidence intervals for long run summary statistics such as cumulative or discounted treatment effects at each point in time as data accumulates. These intervals remain valid at arbitrary stopping times, allowing researchers to terminate data collection whenever evidence becomes sufficiently strong. We apply the framework to data from ASOS, an online retail platform, and demonstrate that our approach enables rejection of the null hypothesis well before the intended four week experimental horizon while maintaining the type I error guarantee.

Causal Discovery

LLM-Augmented Human-in-the-Loop Causal Discovery Chi Zhang* Chi Zhang, Scott Mueller, Rumen Iliev, Laura Libby, Laith Ulaby, Candice Hogan,

Causal structure discovery is a challenging problem. Some core challenges include the limitation of data and scalability to large models. Data alone are almost never sufficient for learning one single causal structure, and real-world data can be noisy. Expert knowledge is helpful in filling in the blanks and correcting the errors, but it becomes cognitively difficult for human experts when the graphs are large. In this work, we address those challenges by developing a novel causal discovery approach that combines constraints from the data, knowledge from large-language models (LLMs), and human expertise. This approach aims to improve the accuracy of the discovered causal structure by iteratively applying information from the three sources. It lowers the burden of human experts by using LLMs to infer causal relationships at the same time. We derive theorems on the accuracy of the proposed algorithm, and empirically evaluate the performance through comparisons with baseline algorithms on synthetic datasets.

Causal Discovery

Expert-Augmented Causal SHAP: Recovering DAG-Consistent Feature Importance via Iterative Causal Discovery and Domain Knowledge Andrew Wilson* Andrew Wilson, Alexandra Pasi, Aimee Harrison, Justin Ross, Jenny Alderden,

Feature importance methods such as SHAP are widely used to help explain machine learning models in health research and other areas. Yet these methods can misattribute feature importance when features have causal structure. Specifically, mediators can absorb credit from upstream causes; a topological artifact that inflates their apparent importance at the expense of upstream causal drivers.

We propose an expert-augmented workflow for DAG-consistent feature attribution that combines constrained causal discovery with an interactive DAG interface to resolve ambiguous edges before computing causal SHAP. The workflow supports iterative refinement through required and forbidden edges and compares standard, DAG-constrained, and adjustment-set SHAP.

We evaluated the approach in synthetic data generated under known DAGs using the R package simcausal and in real-world MIMIC-IV data. In simulations, standard SHAP over-attributed importance to mediators, whereas causal SHAP re-ranked features (Kendall's $\tau=0.42$ between methods) and reduced mediator inflation. In MIMIC-IV, the expert-informed DAG helped distinguish upstream severity drivers from downstream interventions. These results support expert-guided DAG refinement as a practical route to more causally coherent explanations in structured clinical machine learning.

Causal Fairness, and Bias/Discrimination

Debiasing Alternative Data for Credit Underwriting Using Causal Inference Chris Lam* Chris Lam,

Alternative data provides valuable insights for lenders to evaluate a borrower’s creditworthiness, which could help expand credit access to underserved groups and lower costs for borrowers. But some forms of alternative data have historically been excluded from credit underwriting because it could act as an illegal proxy for a protected class like race or gender, causing redlining. We propose a method for applying causal inference to a supervised machine learning model to debias alternative data so that it might be used for credit underwriting. We demonstrate how our algorithm can be used against a public credit dataset to improve model accuracy across different racial groups, while providing theoretically robust nondiscrimination guarantees.

Causal Fairness, and Bias/Discrimination

Individualized Inference for Causal Fairness through Conformal Mediation Analysis Cheng Yu* Cheng Yu, Zhimei Ren,

Ensuring causal fairness is a critical concern across a wide range of applications. However, assessing fairness at the individual level and identifying units that have experienced unfair outcomes remain challenging tasks. In this paper, we focus on the direct effect of a sensitive attribute on an outcome, potentially mediated by other variables, and propose a novel framework for individualized statistical inference. Our approach integrates causal mediation analysis with conformal prediction to enable inference on causal fairness at the individual level. To control the false discovery rate (FDR) in selecting individuals subjected to unfair treatment, we further develop a multiple testing procedure based on the conformalized e-values with conditional calibration. We formalize the notion of individual causal fairness and demonstrate the utility and novelty of our methodology through extensive simulations and two real-world applications.

Causal Inference and SUTVA/Consistencies Violations

Coarsened but Confused: Why Composite Exposures Often Fail in Causal Inference Nicholas Bakewell* Nicholas Bakewell,

In health research, high-dimensional binary treatments are often reduced to composite exposures (CEs) via coarsening functions, typically weighted linear combinations (e.g., medication indicators summarized as a unit-weighted linear combination to form a polypharmacy CE). In causal inference, CEs may be treated as deterministic nodes, often assuming no direct effects from underlying indicators to distal outcomes and informational equivalence. This implies CEs are causally efficacious and sufficient summaries containing necessary information from underlying indicators. However, the latter may not hold, as coarsening functions are often non-invertible and outcome-agnostic. While discussed under multiple versions of treatment theory, this literature implicitly assumes CEs are sufficient, resulting in tautological identification arguments. Further, estimands of CEs represent weighted-averaged effects based on underlying version distributions, but do not allow meaningful per-indicator interpretation as done in practice, and inference is invalid as it ignores uncertainty in the CE. Assumptions under which CE effects may be interpreted as such are formalized: equal conditional effects, no interactions, monotonicity, exchangeability under permutation, sufficiency, and homogenous effect modification across indicators. Simulations demonstrate that even when assumptions hold, further categorization induces bias. These results highlight fundamental problems with CEs for causal inference.

Causal Inference and SUTVA/Consistencies Violations

Learning and Testing Exposure Mappings of Interference using Graph Convolutional Autoencoders Mara Mattes* Mara Mattes, Martin Huber, Jannis Kueck,

Interference or spillover effects arise when an individual's outcome (e.g., health) is influenced not only by their own treatment (e.g., vaccination) but also by the treatment of others, creating challenges for evaluating treatment effects. Exposure mappings provide a framework to study such interference by explicitly modeling how the treatment statuses of contacts within an individual's network affect their outcome. Most existing research relies on a priori exposure mappings of limited complexity, which may fail to capture the full range of interference effects. In contrast, this study applies a graph convolutional autoencoder to learn exposure mappings in a data-driven way, which exploit dependencies and relations within a network to more accurately capture interference effects. As our main contribution, we introduce a machine learning-based test for the validity of exposure mappings and thus test the identification of the direct effect. In this testing approach, the learned exposure mapping is used as an instrument to test the validity of a simple, user-defined exposure mapping. The test leverages the fact that, if the user-defined exposure mapping is valid (so that all interference operates through it), then the learned exposure mapping is statistically independent of any individual's outcome, conditional on the user-defined exposure mapping. We assess the finite-sample performance of this proposed validity test through a simulation study.

Causal Inference Education

From Identification to Implementation: Practical Strategies for Initiating Field Experiments in Government & Non-Profit Research Partnerships Jessie Harney* Jessie Harney,

Initiating RCTs with government agencies and non-profit organizations often hinges less on methodological sophistication and more on effective communication and relationship-building. Many potential partners running innovative programs face some combination of limited statistical training, weak data infrastructure, binding resource constraints, and prior negative or extractive experiences with academic research, sometimes shaped by gatekeeping or other forms of marginalization. As a result, fruitful collaborations can be easily stymied. This talk provides practical strategies for developing effective, ethical, collaboratively-designed RCTs with governments and non-profits. Drawing on lessons from (transparently, an early-career applied researcher), core components of initiating RCTs are proposed: accessible communication of research design & methodology; early & sustained incorporation of partner expertise; explicit attention to providing value to the partner; flexibility and creativity in randomization strategies; and acknowledgment of the ways experiments have been used to cause harm. Rather than treating these collaboration strategies as external to methodological work, I argue that they shape prospects and design of experiments themselves. The session discusses short-term strategies (e.g., communication tips, proposed researcher training, & engaging data-savvy boundary-spanners) and proposed long-term solutions to reducing structural barriers to rigorous, ethical RCTs.

Causal Inference in Networks

Regression Adjustments for Disentangling Spillover Effects David Ritzwoller* David Ritzwoller,

Empirical analyses that characterize the mechanisms that mediate spillover effects often do so by relating responses to a treatment, shock, or policy change with a measure of economic proximity, such as geographic distance, technological similarity, trade costs, or migration flows. Typically, such efforts are based on regressions that associate outcomes with proximity-weighted averages of the treatments received by other units. We show that regressions with this structure measure how the association between one unit's outcome and another unit's treatment correlates with the proximity between the two units. We then argue that, if the proximity measure of interest is associated with other channels that mediate spillover effects, causal interpretations of such relationships are susceptible to confounding. For instance, if technologically similar firms tend to be geographically proximate, then a positive association between technological similarity and the intensity of the productivity spillovers between firms might arise spuriously. We give conditions under which the effect of a proximity measure on spillover intensity can instead be recovered by regressing outcomes on averages of other units' treatments, reweighted by residualized versions of the proximity measure under consideration. We show that estimates obtained in this manner achieve the optimal rate of convergence and give a resampling procedure for constructing estimates of the associated uncertainty.

Causal Inference in Networks

Minimax Rates for Estimating Causal Effects in Network Experiments Vardis Kandiros*
Christopher Harshaw, Vardis Kandiros, Fredrik Sävje, James Robins,

While a growing literature has investigated statistical methods for network experiments, optimal rates of estimation remains a largely open question. Unlike the mature efficiency theory developed within the iid super-population framework, the design based framework (where treatment is the only source of randomness) is generally lacking in statistical lower bounds. For example, we are not aware of work which directly establishes minimax rates of ATE estimation even in the no-interference setting. A central challenge is that a statistical lower bound for randomized experiments should incorporate not only the optimal choice of estimator, but also the optimal choice of the experimental design.

In this paper, we study minimax rates of estimation in network experiments, where optimality is with respect to both the choice of estimator and design. We develop information theoretic lower bounds which depend on both the underlying network and the causal effect, thru the conflict graph introduced by Kandiros et al (2024). For many networks, these lower bounds match the upper bounds obtained by the Conflict Graph Design of Kandiros et. al (2024), establishing the minimax rate of estimation for these networks. However, the upper and lower bounds do not match for all networks. We conjecture that this gap reflects an issue of computational complexity in the following sense: the exact minimax rate of estimation may be NP-Hard to approximate in general.

Causal Inference in Networks

Adaptive Experimental Design for Efficient Causal Estimators under Neighborhood and Temporal Interference Fei Fang* Fei Fang, Laura Forastiere,

Network interference poses fundamental challenges for experimental design, as incompatibilities among nearby units in simultaneously attaining the exposure conditions defining the target estimand can lead to low estimation efficiency. We develop an adaptive design for estimating causal effects under neighborhood interference to improve efficiency. We first consider experiments conducted on multiple networks, where network structure, potential outcomes, and covariates are sampled from a common distribution over time. In this regime, we build on the conflict graph design of Kandiros et al. (2025), which reduces conflicts in realizing target exposure conditions by assigning treatments according to an importance ordering of neighboring units. In the proposed adaptive conflict graph design, exposure sampling probabilities and importance orderings are updated based on observed history and chosen to minimize estimator variance. We then study adaptive design on a single fixed network observed repeatedly over time, where temporal carryover and dependence arise. To address this setting, we propose a block-adaptive design that applies the adaptive conflict graph design at the beginning of each block, with estimation leveraging observations from the carryover period. We jointly minimize estimator variance over block length, exposure sampling probabilities, and importance orderings in an adaptive manner. Synthetic data applications show efficiency gains of our designs.

Design of Experiments

Gaussianized Design Optimization for Covariate Balance in Randomized Experiments
Wenxuan Guo* Wenxuan Guo, Tengyuan Liang, Panos Toulis,

Achieving covariate balance in randomized experiments enhances the precision of treatment effect estimation. However, existing methods often require heuristic adjustments based on domain knowledge and are primarily developed for binary treatments. This paper presents Gaussianized Design Optimization, a novel framework for optimally balancing covariates in experimental design. The core idea is to Gaussianize the treatment assignments: we model treatments as transformations of random variables drawn from a multivariate Gaussian distribution, converting the design problem into a nonlinear continuous optimization over Gaussian covariance matrices. Compared to existing methods, our approach offers significant flexibility in optimizing covariate balance across a diverse range of designs and covariate types. Adapting the Burer-Monteiro approach for solving semidefinite programs, we introduce first-order local algorithms for optimizing covariate balance, improving upon several widely used designs. Furthermore, we develop inferential procedures for constructing design-based confidence intervals under Gaussianization and extend the framework to accommodate continuous treatments. Simulations demonstrate the effectiveness of Gaussianization in multiple practical scenarios.

Design of Experiments

AllocOT: Constrained Treatment Assignment via Semi-Discrete Optimal Transport Mingxun Wang* Qiuran Lyu, Mingxun Wang,

Treatment assignment is central to causal inference but often must respect real-world constraints, including adaptive updates, fairness across demographic groups, and budget or capacity limits. Moreover, as the number of treatment arms grows and constraints become complex, generic solvers such as Gurobi can become computationally expensive.
We propose Constrained Allocation via Optimal Transport (AllocOT), a semi-discrete optimal transport framework that casts constrained assignment as transporting the covariate distribution to a finite set of treatment arms. AllocOT gives assignment rules for each unit by minimizing an overall cost; for outcome-driven allocation, the cost is the negative predicted outcome under each arm, so minimizing cost maximizes expected benefit. Practical requirements are imposed as linear quota constraints on aggregate assignment frequencies, for example, fixed arm ratios, subgroup-specific exposure bounds for fairness and balance, and capacity or budget caps for scarce or costly interventions.
We validate AllocOT performance across simulation studies covering varying numbers of arms and constraint regimes. The result shows that AllocOT delivers high-quality allocations with substantially improved scalability and runtime compared with generic optimization solvers, making it a practical approach for large-scale constrained treatment assignment in causal inference.

Design of Experiments

Dynamic Adaptive Rerandomization for Efficient Sequential Trials Under Budget Constraints Kateryna Husar* Kateryna Husar,

We introduce the Dynamic Adaptive Rerandomization (DAR) framework to improve the statistical efficiency of the Average Treatment Effect (ATE) estimator in resource-constrained sequential randomized controlled trials (RCTs). DAR combines two linked Bayesian procedures for concurrent adaptive learning and robust estimation. In each batch, a Thompson Sampling-based Multi-Armed Bandit (MAB) policy selects covariates with the highest posterior predictive importance while respecting the measurement budget. Measured covariates are then used in rerandomization to minimize the variance of the batch-specific ATE estimator ($hattau_k$). The overall Average Treatment Effect (tau) is estimated sequentially using a Bayesian weighted averaging approach. The $hattau_k$’s are combined using inverse-variance weighting, where uncertainty in the variance estimates ($sigma_k^2$) is explicitly modeled. This structure naturally assigns greater weight to batches rerandomized on more predictive covariates, as these estimators exhibit smaller sampling variances. Covariate predictive utility is subsequently reassessed to update the importance prior, closing the adaptive loop. This dual-path mechanism allows for identification of influential covariates, optimizes resource use, and yields a precise, fully quantified posterior for the ATE.

Design-Based Causal Inference

Using a Two-Parameter Sensitivity Analysis Framework to Efficiently Combine Randomized and Non-randomized Studies Ruoqi Yu* Ruoqi Yu, Bikram Karmakar, Jessica Vandeleest, Eleanor Schwarz,

Causal inference is vital for informed decision-making across fields such as biomedical research and social sciences. Randomized controlled trials (RCTs) are considered the gold standard for internal validity of inferences, whereas observational studies (OSs) often provide the opportunity for greater external validity. However, both data sources have inherent limitations preventing their use for broadly valid statistical inferences: RCTs may lack generalizability due to their selective eligibility criterion, and OSs are vulnerable to unobserved confounding. This paper proposes an innovative approach to integrate RCT and OS that borrows the other study's strengths to remedy each study's limitations. The method uses a novel triplet matching algorithm to align RCT and OS samples and a new two-parameter sensitivity analysis framework to quantify internal and external validity biases. This combined approach yields causal estimates that are more robust to hidden biases than OSs alone and provides reliable inferences about the treatment effect in the general population. We apply this method to investigate the effects of lactation on maternal health using a small RCT and a long-term observational health records dataset from the California National Primate Research Center. This application demonstrates the practical utility of our approach in generating scientifically sound and actionable causal estimates.

Design-Based Causal Inference

Testing individual-level null without imputation Zijun Gao* Zijun Gao,

An individual-level null specifies restrictions on each unit’s potential outcomes. Fisherian randomization inference provides model-lean, finite-sample–valid tests for a class of such nulls via imputation of potential outcomes. However, for general individual-level nulls, such as a single contrast across multiple treatment levels, the potential outcomes may no longer be imputable. Fisherian randomization inference therefore does not directly apply, and existing approaches typically introduce additional individual-level restrictions and effectively tests a stronger null.

We develop an e-value–based test for general individual-level nulls that circumvents imputation. The key idea is to construct e-values whose validity is directly implied by the restrictions in the null. The proposed test is finite-sample valid, with validity determined solely by the treatment assignment. It also permits data-dependent e-values without sample splitting, yielding adaptive tests with high power. We illustrate the efficacy of our method by testing zero interaction effects in factorial designs.

Design-Based Causal Inference

Exact Fisherian P-Values for Multi-Armed Bandits Adam Sales* Adam Sales, Ethan Prihar,

Randomized multi-armed-bandits experimental designs in which successive subjects are adaptively randomized subjects between conditions: randomization probabilities are based on previous users' outcomes. One heuristic, Thompson sampling, essentially randomizes users to conditions according to the posterior probabilities that each condition is optimal, conditional on previous users' responses. However, from a scientific perspective, RMABs present a challenge: the adaptivity of RMABs induces a complex dependence structure between the observations which invalidates usual approaches to statistical inference from randomized experiments, which assume some degree of independence between observations.
This paper illustrates a simple, but overlooked, solution: simulation-based exact p-values for Fisher's strict null hypothesis of no effect. This approach requires a complete record of randomization probabilities, treatment assignments, and outcomes—and sufficient computational power—but little else. We illustrate this approach using a new dataset of over 200 RMABs conducted on an online homework platform, where the outcome of interest was students' correctness on the next problem, and compare randomization-based p-values using a variety of test statistics to naïve chi-squared tests.

Design-Based Causal Inference

Crime, Salience, and the Housing Market Nency Dhameja* Nency Dhameja,

How do housing markets respond to nearby crime? Do buyers price crime based on objective risk, or do they react more strongly to highly visible events? Using 347,455 repeat home sales from 148,349 Chicago properties between 2008 and 2022, matched to 4.7 million geocoded crime incidents, this paper examines how crime is capitalized into housing prices.
Preliminary analysis across 120 distance–time specifications shows that 109 (90.8%) produce negative point estimates, with the largest effects concentrated at the shortest distances and most recent time windows. A near–far identification strategy sharpens this descriptive pattern into a causal test by comparing crime occurring immediately around a property (0–0.05 miles) with crime slightly farther away in the surrounding neighborhood (0.05–0.30 miles), under property and Community Area × Month fixed effects. Crime within the immediate block significantly reduces sale prices by −0.195%, while crime slightly farther away has no detectable effect, indicating highly localized capitalization. Within violent crimes, daytime incidents affect prices, whereas nighttime incidents of the same categories do not. Taken together, the results suggest that housing markets respond disproportionately to visible, unexpected crime events, producing block-level price distortions driven by salience rather than long-run risk.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Pharmacometrics modeling in combination with G-formula to adjust for time-varying confounder for time-to-event analysis Siyan Xu* Siyan Xu, Harriet Longley, Thomas Dumortier, Yu-Yun Ho,

Treatment crossover (XO) from control to experimental in oncology complicates estimation of treatment effect of overall survival (OS). A hypothetical estimand is: the probability of survival in control arm under a counterfactual scenario where XO is disallowed. Naïve censoring at XO time yields biased estimates when XO is driven by post randomization variables that also affect survival. In prostate cancer trials, prostate specific antigen (PSA) influences both XO and OS, making PSA a time varying confounder of the relationship between time to XO and time to death. Assuming PSA is the sole confounder, conditioning on PSA can render time to XO independent of time to death, enabling unbiased estimation of the survival even when censoring at XO time.
We estimate hypothetical estimand using g-formula, with a pharmacometrics modelling framework and parametric time-to-event analysis. Trial data is simulated from an “assumed ground truth” kinetic-pharmacodynamic model where PSA is the only time-varying confounder that impacts both XO and death hazards. G-formula based survival estimates were compared to the known “ground truth” survival curve.
Simulations show that, under correct model assumptions and no unmeasured confounders, the g-formula approach accurately estimates the hypothetical estimand. This finding highlights the value of integrating pharmacometric modeling with causal inference to improve the estimation of treatment effect when time-varying confounders are present.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

A Latent Factor Panel Approach to Spatiotemporal Causal Inference Jiaxi Wu* Jiaxi Wu, Alexander Franks,

Unmeasured confounding can severely bias causal effect estimates from spatiotemporal observational data, especially when the confounders do not vary smoothly in time and space. In this work, we develop a method for addressing unmeasured confounding in spatiotemporal contexts by building on models from the panel data literature and methods in multivariate causal inference. Our method is based on a factor confounding assumption, which posits that effects of unmeasured confounders on exposures and outcomes can be captured by a shared latent factor model. Factor confounding is sufficient to partially identify causal effects, even when there is interference between units. Additional assumptions that limit the degree of spatiotemporal interference, reasonable in most applications, are sufficient to point identify the effects. Simulation studies demonstrate that the proposed approach can substantially reduce omitted variable bias relative to other spatial smoothing and panel data baselines. We illustrate our method in a case study of the effect of prenatal PM2.5 exposure on birth weight in California.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Synthetic Control with Disaggregated Data Lea Bottmer* Lea Bottmer,

The synthetic control estimator is widely used to evaluate aggregate-level policies, but researchers increasingly face settings with rich, disaggregated data (e.g., county-level outcomes within states) that raise new questions about aggregation choice. Existing approaches incorporate such data by estimating separate synthetic controls for each disaggregated treated unit, enlarging the donor pool with disaggregated control units, or both. These strategies can improve fit but also amplify noise, with little guidance on how to balance these trade-offs. This paper develops a general framework for synthetic control with disaggregated data that nests the classical synthetic control estimator and other existing approaches. Within this framework, I propose a multi-level SC (mlSC) estimator that formalizes the aggregation choice as a data-driven regularization problem. The estimator flexibly regularizes toward the classical synthetic control estimator while exploiting additional variation from the disaggregated data. In simulations calibrated to four empirical settings, mlSC matches or outperforms existing approaches. Two applications—Minnesota's cigarette tax and minimum wage effects on teen employment—illustrate its practical value.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Distributionally Robust Synthetic Control: Ensuring Robustness Against Highly Correlated Controls and Weight Shifts Taehyeon Koo* Taehyeon Koo, Zijian Guo,

he synthetic control method estimates the causal effect by comparing the treated unit's outcomes to a weighted average of control units that closely match its pre-treatment outcomes, assuming the relationship between treated and control potential outcomes remains stable before and after treatment. However, the estimator may become unreliable when these relationships shift or when control units are highly correlated. To address these challenges, we introduce the Distributionally Robust Synthetic Control (DRoSC) method, which accommodates potential shifts in relationships and addresses high correlations among control units. The DRoSC method targets a novel causal estimand defined as the optimizer of a worst-case optimization problem considering all possible weights compatible with the pre-treatment period. When the identification conditions for the classical synthetic control method hold, the DRoSC method targets the same causal effect as the synthetic control; when these conditions are violated, we demonstrate that this new causal estimand is a conservative proxy for the non-identifiable causal effect. We further show that the DRoSC estimator's limiting distribution is non-normal and propose a novel inferential approach. We demonstrate its performance through numerical studies and an analysis of the economic impact of terrorism in the Basque Country.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

In Defense of the Pre-Test: Valid Inference when Testing Violations of Parallel Trends for Difference-in-Differences Jonas Mikhaeil* Jonas Mikhaeil, Christopher Harshaw,

The difference-in-differences (DID) design is a key identification strategy which allows to estimate causal effects under the parallel trends assumption. While the parallel trends assumption is counterfactual and cannot be tested directly, researchers often examine pre-treatment periods to check whether the time trends are parallel before treatment is administered. Recently, researchers have been cautioned against using preliminary tests which aim to detect violations of parallel trends in the pre-treatment period. We argue that preliminary testing should play an important role within the DID research design. We propose a new and more substantively appropriate conditional extrapolation assumption, which requires to conduct a preliminary test to determine whether the severity of pre-treatment parallel trend violations falls below an acceptable level before extrapolation to the post-treatment period is justified. This stands in contrast to prior work which can be interpreted as either setting the acceptable level to be exactly zero (in which case preliminary tests lack power) or assuming that extrapolation is always justified (in which case preliminary tests are not required). Under mild assumptions, we provide a consistent preliminary test as well confidence intervals which are valid when conditioned on the result of the test. The conditional coverage of these intervals overcomes a common critique made against the use of preliminary testing within the DID design.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Causal Supply–Demand Decomposition with Movers Zhu Shen* Zhu Shen, Jose Zubizarreta,

Mover designs are widely used to analyze regional differences by leveraging individual relocations to separate place-specific effects from population composition. Although commonly used to address causal questions, these designs are rarely formulated within an explicit causal framework, making the resulting estimates difficult to interpret for policy-relevant decomposition. We develop a causal framework for movers that defines origin–destination–time–specific causal effects and makes the associated identification assumptions explicit. These cohort-specific mover effects serve as building blocks for a general supply–demand decomposition of regional differences. We show that, in multi-period and multi-region settings, observed cross-sectional outcome differences can be decomposed into causal supply effects, defined by holding cohort composition fixed, and demand effects arising from population heterogeneity. To estimate these effects, we propose a weighting-based estimator that recovers cohort-specific mover effects by approximately balancing pre-move covariates between movers and appropriate stayer comparison groups. Applying the framework to Medicare data on patient moves across Hospital Referral Regions, we uncover substantial heterogeneity in mover effects by origin, destination, and timing, and show how place-specific factors and population composition jointly contribute to geographic variation in health care utilization.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Difference-in-Differences in the Presence of Unknown Interference Javier Viviens* Javier Viviens, Fabrizia Mealli,

The stable unit treatment value (SUTVA) is a crucial assumption in the Difference-in-Differences (DiD) research design. It rules out hidden versions of treatment and any sort of interference and spillover effects across units. Even if this is a strong assumption, it has not received much attention from DiD practitioners and, in many cases, it is not even explicitly stated as an assumption, especially the no-interference assumption. In this technical note, we investigate what the DiD estimand identifies in the presence of unknown interference. We show that the DiD estimand identifies a contrast of causal effects, but it is not informative on any of these causal effects separately, without invoking further assumptions. Then, we explore different sets of assumptions under which the DiD estimand becomes informative about specific causal effects. We illustrate these results by revisiting the seminal paper on minimum wages and employment by Card and Krueger (1994).

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Resource Onshore, Restriction Offshore: A Causal Decomposition of the AH Premium via Heckman-HCW Yujue Wang* Yujue Wang,

Despite the deepening of financial integration via the Stock Connect program in China's stock market, the persistent AH Premium continues to challenge the Law of One Price. This paper proposes a political wedge mechanism, arguing that conflicting institutional beliefs price state ownership as a "resource" (implicit guarantee) onshore but a "restriction" (agency cost) offshore. Integrating the Heckman correction with the HCW panel method (Hsiao, Ching, & Wan 2012), I construct a counterfactual benchmark to isolate causal effects from firm-quality selection bias. Empirical results reveal a dual-layer mechanism. First, state ownership drives a policy-led selection, significantly increasing the propensity to dual-list ($beta=0.149$). Second, while negative selection on unobservables ($text{IMR}, beta=-0.84$) explains part of the discount, the average treatment effect on valuation remains robustly positive. This persistence underscores the stubbornness of political beliefs: even as market barriers diminish, offshore investors structurally price a "sovereign discount" that defies simple arbitrage. This ingrained belief is only partially mitigated by high-transparency proxies, specifically HK-Macau-Taiwan backgrounds ($beta=-0.56$) and firm size ($beta=-0.64$). Ultimately, these findings demonstrate that valuation divergence is not merely a transient friction, but a resilient reflection of institutional incompatibility that capital channel liberalization alone cannot eliminate.

Difference in Differences, Synthetic Control, Methods for Panel and Longitudinal Data

Semi-parametric Estimation Under a Stationarity Assumption With Applications to Quasi-Experimental Designs Gary Hettinger* Gary Hettinger,

Semi-parametric causal inference typically relies on assumptions that connect causal estimands to directly observable data, enabling non-parametric estimation of nuisance functions and the construction of doubly robust estimators. However, many important problems such as positivity violations, transportability across populations, and controlled interrupted time series, require extrapolation beyond the observed data support, making fully non-parametric identification impossible. In these settings, part of the data-generating process must be specified parametrically, often through assumptions such as stationarity or stable trends, while other components remain non-parametrically estimable.

In this work, we formulate such problems as a semi-parametric estimation problem under required extrapolation and provide formal theory to what is done often implicitly in practice. Further, we propose a unified semi-parametric framework that separates parametric extrapolation components from nonparametric confounding mechanisms and characterize the resulting efficiency-robustness tradeoffs. The framework is illustrated via a controlled interrupted time series analysis evaluating a nutritional excise tax policy. Our results clarify attainable bounds on robustness and efficiency under extrapolation and provide practical guidance for causal inference in modern observational studies.

Dynamic Treatment Regimes

Evaluating Dynamic System-Level Congestion Pricing Regimes with Endogenous Interference Aleksander Holleran* Aleksander Holleran,

Vehicle tolling operates as a dynamic treatment regime applied at the system level, where interventions affect shared congestion states that feed back into future behavior. Such settings violate standard causal assumptions through interference and state dependence.
We study dynamic treatment regimes in a transportation setting using an agent-based simulation with endogenous state evolution. The simulation is embedded in the physical road geometry and resolves vehicle movement at one-second intervals, allowing congestion to emerge from spatially constrained driving behavior and mode choice. Time-varying tolls are determined by rules that map observed conditions to prices. Under dynamic regimes, tolls are updated continuously. Congestion functions as a shared, time-varying state through which policy effects propagate, generating interference across agents and over time.
We compare policy regimes across synthetic seasons of varying demand, crash, and closure conditions. Seasons allow agents to learn expected travel times from experience, generating behavioral responses to policy-driven congestion relief. Regime performance is assessed using aggregate welfare measures that combine travel time and toll payments, evaluated with both individual-specific values of time (VOT) and median VOT to support distributionally neutral comparisons. This framework supports regime-level policy evaluation in settings where interference, learning, and feedback are intrinsic features.

Dynamic Treatment Regimes

Robust Estimation of Stochastic Intervention Effects under Multistage Missingness
Shengfang Song* Shengfang Song, Hongxiang Qiu, Honglei Chen, Christine Parks, Zhehui Luo,

We develop estimation and inference methods for stochastic intervention effects in longitudinal studies with outcome-dependent sampling and multistage missing data. We consider estimation of the mean counterfactual outcome under a proportional modified treatment policy for a semi-continuous exposure, with outcomes measured decades after exposure and data collected through three stages.

Under a nonparametric model, we derive the efficient influence function and propose a one-step estimator that integrates stochastic treatment reweighting with stage-specific missingness mechanisms. The estimator accommodates semi-continuous exposures and monotone nonresponse, extending existing stochastic intervention methods that typically assume fully observed data or simpler missingness structures. Assuming correct specification of the treatment mechanism, the estimator is sequentially multiply robust, remaining consistent if at each stage either the outcome model or the missingness mechanism is correctly specified. When all nuisance functions are consistently estimated, the estimator is asymptotically linear and achieves the semiparametric efficiency bound.

This work fills a key gap by providing a single estimator that jointly handles stochastic interventions, outcome-dependent sampling, and multistage missingness. Simulation studies are underway, and an application to data from the Agricultural Health Study and its Pesticide and the Sense of Smell sub-study is ongoing.

Generalizability/Transportability

Data Fusion with Distributional Equivalence Test-then-pool Linying Yang* Linying Yang, Xing Liu, Robin Evans,

Randomized controlled trials (RCTs) are the gold standard for causal inference, but their high costs and recruitment challenges often limit the feasibility of fully powered studies. A common remedy is to supplement the current control arm with data from historical trials. However, naively pooling external controls can introduce bias when the populations differ. Existing test-then-pool (TTP) approaches attempt to guard against this risk, but standard implementations are prone to both loss of power and uncontrolled bias when distributions differ.
We propose a new TTP framework that fuses control arms while rigorously controlling the Type-I error rate of the final treatment effect test. Our method leverages kernel two-sample testing via maximum mean discrepancy (MMD) to capture distributional differences, and equivalence testing to avoid introducing bias due to under-powered fusion test, providing a more flexible and informative criterion for pooling. To ensure valid inference, we introduce partial bootstrap and partial permutation procedures for approximating null distributions in the presence of heterogeneous controls. We further establish the overall validity of the fused treatment effect test and provide guidance on selecting equivalence margins to balance power and robustness.

Generalizability/Transportability

Generalized projection tests for function-valued parameters with applications to testing structural causal assumptions Rui Wang* Rui Wang, Albert Osom, Bo Zhang,

Structural identification assumptions are central to the causal inference literature. In practice, it is often crucial to assess their validity or to test implications that follow from them. In many settings, such tests can be framed as evaluating whether a function-valued parameter equals zero. In this paper, we propose a class of generalized projection tests based on series estimators for testing such function-valued parameters. We establish conditions under which the proposed tests are valid and illustrate their applicability through examples from the data fusion and instrumental variables literatures. Our approach accommodates flexible machine learning methods for estimating nuisance parameters. In contrast to existing approaches, the limiting distribution of the proposed test statistics is straightforward to compute under the null hypothesis. We apply our method to test the equality of conditional COVID-19 incidence rates across vaccine arms in the COVID-19 Variant Immunologic Landscape (COVAIL) trial.

Generalizability/Transportability

Beyond the Experiment Window: Prospective Impacts Under Long-Term Ranking Dynamics
Lo-hua Yuan* Lei Shi, Lo-hua Yuan, Peng Ding, Navin Sivanandam,

Short A/B tests for ranking systems can be myopic when seasonality, user evolution, and feedback loops drive outcomes beyond the experiment window. We target the prospective long-term average treatment effect (PLATE): the cumulative effect of sustaining a new ranker versus the incumbent over a future horizon for the experimental population. Estimating PLATE from short experiments requires adjusting for time-varying post-treatment covariates, imputing long-run outcomes when the new ranker is not fully represented in historical logs, and transporting information across experimental and observational data under covariate shift. We propose BSTAR (Blip Surrogate TrAnsfeR), combining structural nested mean model blips, surrogate-index identification that treats the displayed result set as a mediator, and causal transfer learning for generalization across data sources. Under sequential randomization, surrogacy, and transferability assumptions, BSTAR identifies PLATE and yields a practical estimation pipeline with bootstrap inference. Simulations calibrated to marketplace ranking experiments show reduced bias and MSE versus inverse-propensity weighting a nd surrogate-index baselines, enabling earlier and more reliable long-term impact estimates.

Generalizability/Transportability

Robust trial augmentation using external data Guanbo Wang* Guanbo Wang, Rickard Karlsson, Pier De Bartolomeis, James Robins, Eric Tchetgen Tchetgen, Issa Dahabreh,

Randomized trials often have sample sizes that are too small to produce precise estimates of treatment effects. One approach for improving trial efficiency is to incorporate external data, from previously completed trials or observational studies, into the estimation process. When the external data are aligned with the trial data and statistical models for nuisance functions are correctly specified, using the external data can yield consistent estimates and enhance efficiency. However, some degree of misalignment or misspecification is usually expected and can threaten the validity of trial analyses incorporating external data. Here, we develop a class of estimators that exploit randomization to ensure consistency and asymptotic normality, even when the external data are misaligned with the trial. We also propose a procedure that uses members of this class of estimators to construct a combined estimator that is consistent and asymptotically normal, and can leverage external data even when that data are misaligned with the trial, or when models for nuisance functions are misspecified or have slow convergence rates. We show that the efficiency of the combined estimator is no lower than that of each of its component estimators (including the efficient trial-only estimator, if it is used as a component for the combined estimator). Our methods allow investigators to use external data to improve the trial’s efficiency without concern for misalignment between the external data and t

Graphical Models

Recursive proximal identification when common causes or mediators are unobserved
Beatrix Wen* Helen Guo, Ilya Shpitser,

Nonparametric identifiability of the causal effect in the presence of hidden variables is characterized by the ID algorithm. In cases where the causal effect is not identified nonparametrically, prior work has augmented the ID algorithm via proximal causal learning methods, allowing some unobserved confounders to be handled via proxies, and others via the standard machinery of the ID algorithm based on the fixing operator.

We present a novel generalization of the ID algorithm that allows identification if either confounders or mediators are unobserved, provided informative proxies for these unobserved variables are available. Our generalization is based on a novel reformulation of the ID algorithm via a fixing operator that resembles sequential applications of either the backdoor or the front-door criterion.

Heterogeneous Treatment Effects

Efficiency Gain of Covariate-Adjusted Differential Variance Estimators in Randomized Controlled Trials Hani Zaki* Hani Zaki, Tania Janaudis-Fereirra, Philippe Boileau, Mireille Schnitzer,

Contrasts of potential outcomes' variances have recently been proposed to detect treatment effect heterogeneity, even when treatment effect modifiers are missing or mismeasured. Adjusted causal machine learning estimators of these causal estimands have been derived and shown to be doubly robust and asymptotically linear under mild conditions. They were successfully applied to detect treatment effect heterogeneity in the re-analysis of randomized controlled trials (RCTs). However, it was not previously demonstrated that the covariate-adjusted estimators are more efficient compared to unadjusted estimators in the context of RCTs. Best practices for inference of these new causal estimands in RCTs are therefore unclear, particularly in situations where the asymptotic guarantees of the adjusted estimators may not be achieved due to small sample sizes. To address this gap, we derive unadjusted estimators and compare their asymptotic and finite-sample behavior to that of the adjusted estimators. We show theoretically and empirically that the adjusted estimator is asymptotically more efficient, but that empirically the unadjusted estimators can have smaller variance in small-sample settings. We then apply the adjusted and unadjusted estimators to data from an RCT aimed at evaluating the effectiveness of a virtual home-based physical rehabilitation program for patients living with long COVID.

Heterogeneous Treatment Effects

Generalized Causal Rule Ensembles: Interpretable Heterogeneous Exposure–Response Functions for Continuous Exposures Suhwan Bong* Suhwan Bong, Heejun Shin, Francesca Dominici,

In many areas of scientific research, it is often of interest to estimate an exposure response function that has a causal interpretation, and most importantly to also assess whether the shape of that ERF is heterogeneous across subgroups of the population. However, most of the recent work on heterogeneous causal effects is limited to the case where the exposure is binary. In this paper, we introduce the Generalized Causal Rule Ensemble (GCRE), a data-driven, rule-based framework that uncovers and summarizes heterogeneity in continuous exposures. The GCRE can be summarized in two steps. First, in a subgroup discovery step, we rely on targeted-smoothing Bayesian Additive Regression Trees (tsBART) to identify covariate-based decision rules whose effects are smooth in the exposure, and then applies group LASSO on spline expansions of exposure to select a sparse, stable subset of subgroups. Second, in a inference step, we estimate subgroup-specific deviation functions and ERFs for the selected rules with favorable theoretical guaranties. In simulations, GCRE recovers heterogeneous ERF that are invisible in dichotomized analyses and improves estimation accuracy relative to standard ERF estimators. We illustrate the approach in a ZCTA-level study of the health effects of PM 2.5 among U.S. Medicare beneficiaries in 2021, and provide summaries of how the ERF varies between sociodemographic subgroups.

Heterogeneous Treatment Effects

Constructing influence sets for heterogeneous treatment effect models Melody Huang*
Melody Huang, Ana Kenney, Tiffany Tang,

Evaluating the performance of heterogeneous treatment effect (HTE) models (i.e., metalearners) is inherently challenging. Unlike standard supervised learning problems, the prediction target–the conditional average treatment effect–is unobservable. As a result, existing model evaluation tools (i.e., omnibus goodness-of-fit tests, or relative error measures) must rely on indirect diagnostics; however, these approaches can highly sensitive to certain observations in a study. In the following paper, we propose a novel framework for identifying influence sets: subsets of observations whose removal induces the largest change in an HTE model's predictive behavior. Our approach extends classical influence function methodology to account for dependence across observations, enabling the exact characterization of the impact of removing multiple points simultaneously. This extension generalizes to complex metalearners requiring multi-stage estimation. We show that identifying the most influential subset can be formulated as a mixed integer quadratic optimization problem, yielding global optimality guarantees for the resulting influence set. Moreover, our approach is general and can be adapted to other metrics of interest beyond shifts in predictive behavior. We illustrate the utility and flexibility of our approach in a case study evaluating the impact of cash transfer program.

Heterogeneous Treatment Effects

The Parachuted Hybrid CATE Estimator with Bootstrap Methods for Inference Xianlin Sun*
Xianlin Sun, Stephen Man Sing Lee,

This paper introduces a novel hybrid estimator for the Conditional Average Treatment Effect (CATE) that achieves optimal convergence rates and robustness against model mis-specification. Our proposed “parachute” estimator exhibits the oracle property: it achieves the fast parametric convergence rate when either the propensity score or outcome model is correctly specified, and gracefully degrades to the slower non-parametric rate when both are mis-specified. We make two primary contributions.
First, we derive the complete asymptotic distribution of the joint vector of parametric and non-parametric CATE estimators. This theoretical novelty culminates in Theorem 1, establishing their joint asymptotic normality. Corollary 1 characterizes the asymptotic distribution of our hybrid estimator, revealing its adaptive “parachute” mechanism. The complexity of this distribution motivates our second contribution.
Second, we establish the theoretical validity of bootstrap methods for constructing confidence intervals. Theorem 2 proves the consistency of the bootstrap for estimating the hybrid estimator’s distribution, providing a practical tool for statistical inference. This builds upon Chatterjee and Bose’s (2005) framework for M-estimation. Detailed proofs are in the Appendix. We conclude with a simulation study demonstrating our estimator’s empirical performance and bootstrap coverage properties.

Instrumental Variables

Design-based nested instrumental variable analysis Zhe Chen* Zhe Chen, Bo Zhang, Xinran Li,

Two binary instrumental variables (IVs) are nested if individuals who comply under one binary IV also comply under the other. This situation often arises when the two IVs represent different intensities of encouragement or discouragement to take the treatment-one stronger than the other. In a nested IV structure, treatment effects can be identified for two latent subgroups: always-compliers and switchers. Always-compliers are individuals who comply even under the weaker IV, while switchers are those who do not comply under the weaker IV but do under the stronger one. In this article, we introduce a novel pair-of-pairs nested IV design, where each matched stratum consists of four units organized in two pairs. Under this pair-of-pairs design, we develop design-based inferential methods for estimating the always-complier sample average treatment effect (SATE) and switcher SATE. In a nested IV analysis, IV assignment is randomized within each IV pair; however, whether a study unit receives the weaker or stronger IV may not be randomized. To address this complication, we then propose a novel partly biased randomization scheme and study design-based inference under this new scheme. Using extensive simulation studies, we demonstrate the validity of the proposed method and assess its power under different scenarios. Applied to PLCO trial data, we identified 52.2% always-compliers and 26.7% switchers, with sigmoidoscopy showing potential benefit for always-compliers but not switchers.

Instrumental Variables

A Modified Instrumental Variable Approach for Modeling Engagement in Mobile Health
Alexis Fleming* Alexis Fleming, Andrew Spieker,

Engagement with mobile health interventions has gained increasing attention as mobile devices become more widely accessible, with identifying thresholds for effective engagement being a primary focus. However, in these settings, the exclusion restriction assumption is unreasonable. Previous methodological advances have addressed this challenge by modifying instrumental variable (IV) approaches through sensitivity analyses to understand how engagement influences improvements in outcomes when the exclusion restriction cannot be reasonably assumed. These methods, however, have largely been limited to settings with two treatment arms and a continuous measure of engagement.

We propose a modified instrumental variable (MIV) approach that builds on this prior work and broadens its application to real-world interventions. Our method allows for multiple treatment groups and supports alternative distributions of the engagement variable, thereby expanding the range of settings in which engagement effects can be assessed. Under correct specification of a sensitivity parameter, the class of local average treatment effects (LATEs) of interest becomes identifiable.

Estimation proceeds through two-stage least squares to recover a weighted average of the LATEs over a specified range of the sensitivity parameter. We evaluate the proposed approach using simulation studies and demonstrate its ability to recover LATEs across varying combinations of treatment and engagement variable types.

Instrumental Variables

Two-stage least squares with clustered data Anqi Zhao* Anqi Zhao, Peng Ding, Fan Li,

Clustered data are common in empirical research. To estimate the causal effect of a possibly endogenous treatment, a common approach—which we call the canonical two-stage least squares (2sls)—is to fit a 2sls regression of the outcome on treatment status with instrumental variables (IVs) for point estimation, and apply cluster-robust standard errors in inference. When both the treatment and IVs have variation within clusters, a natural alternative—which we call the two-stage fixed effects (2sfe)—is to include cluster indicators in the 2sls specification, thereby incorporating cluster information in point estimation as well. This paper clarifies the trade-off between the canonical 2sls and 2sfe within the local average treatment effect (LATE) framework, and makes the following contributions. First, we establish the validity and relative efficiency of the canonical 2sls and 2sfe for large-sample Wald-type inference of the LATE when clusters are homogeneous. We show that, when the true outcome model includes a cluster fixed effect, 2sfe is more efficient than the canonical 2sls when the variation in cluster fixed effects dominates that in unit-level errors. Second, we show that with heterogeneous clusters, 2sfe recovers a weighted average of cluster-specific LATEs, whereas the canonical 2sls does not. Third, we develop a joint asymptotic analysis of the canonical 2sls and 2sfe under homogeneous clusters and propose a Wald-type test for detecting cluster heterogeneity.

Instrumental Variables

The Multiplicative Quasi-Instrumental Variable Model Jiewen Liu* Jiewen Liu, Eric Tchetgen Tchetgen, Chan Park, David Richardson,

We introduce the Quasi-Instrumental Variable (QIV) model, a framework for causal inference with unmeasured confounding that leverages an instrument that may be imperfectly exogenous. We allow the candidate instrument to have a direct effect on the outcome not mediated by the treatment, thus violating the standard IV exclusion restriction. Despite this, we establish nonparametric identification of the population average treatment effect on the treated (ATT) under a multiplicative treatment model, in which the QIV and the hidden confounder combine multiplicatively to govern treatment uptake. This multiplicative structure arises naturally, when treatment occurs only if both two independent instrument-driven and confounder-driven causal mechanisms are present, or when uptake follows a class of latent index model. Importantly, the QIV model is agnostic to treatment-effect heterogeneity with respect to hidden confounders. Identification is achieved via a modified Wald ratio estimand, which corrects the bias due to the exclusion restriction violation, and we propose a new class of estimators that are multiply robust and semiparametric efficient. Finally we propose a straightforward falsification test for the proposed QIV model, and we evaluate the approach in extensive simulations and an application to evaluate the causal effect of having three or more children on mothers' labor-market engagement.

Instrumental Variables

Identifying the Cumulative Average Effect of a Multi-Phase Treatment with Noncompliance by Leveraging Multisite Instrumental Variables Guanglei Hong* Guanglei Hong, Fan Yang, Zhengyan Xu, Xu Qin,

When evaluating a multi-phase intervention, the cumulative average treatment effect (ATE) is often a causal estimand of primary interest. However, among individuals who do not respond well to the treatment in an earlier phase, some may subsequently display noncompliant behaviors. At the same time, exposure to the earlier-phase treatment is expected to both directly and indirectly influence an individual's potential outcomes. Building on an instrumental variable (IV) strategy for multisite trials, we clarify the conditions under which the cumulative ATE of a multi-phase treatment can be identified by employing the randomization of the initial treatment assignment as the instrument. Our strategy relaxes both the conventional exclusion restriction and sequential ignorability assumptions. We assess the performance of the new strategy through simulation studies and clarify data requirements for estimation. Additionally, we reanalyze data from the Tennessee class size study, in which students and teachers were randomly assigned to either small or regular class types in kindergarten (Phase I) with noncompliance emerging in Grade 1 (Phase II) and again in Grade 2 (Phase III). Applying our new strategy, we estimate the cumulative ATE of receiving three consecutive years of instruction in a small versus regular class.

Instrumental Variables

Few Clusters, Many Problems: A Clustered Wild Bootstrap for Instrumental Variables Estimation with Evidence from School-Board Gender Representation on Achievement Gaps
Sam Lee* Sam Lee,

Empirical researchers often estimate instrumental variables (IV) models to address endogeneity in key regressors. When observations are correlated within clusters, however, conventional inference can fail, particularly when the number of clusters is small. Standard cluster-robust variance estimators may understate uncertainty in these settings. I introduce a finite-sample wild cluster bootstrap procedure for just-identified IV models under arbitrary cluster dependence. The method recovers the consistency properties of the Liang–Zeger cluster-robust variance estimator and is shown to coincide with a restricted wild cluster bootstrap under the null hypothesis. Monte Carlo simulations demonstrate that the proposed procedure achieves rejection rates close to nominal levels and uniformly improves upon existing cluster-robust approaches. An empirical application estimates the causal effect of female school-board representation on gender gaps in scholastic achievement in California from 2014–2024. While the estimates show no significant reduction in within-district gender achievement disparities, the results provide applied researchers with a transparent and reliable framework for inference in clustered IV settings.

Intercurrent Event

Implementing the principal stratum strategy for intercurrent events with survival outcomes: a tutorial Xiaoxiao Zhou* Xiaoxiao Zhou, Fan Li,

The International Council for Harmonization (ICH) E9 (R1) addendum introduces the estimand framework to standardize how treatment effects are evaluated. The framework comprises five attributes, including intercurrent events (ICEs) and five strategies to address them. Among these strategies, the principal stratum strategy is the most conceptually and technically challenging because it defines treatment effects on unobserved strata. Its application to survival outcomes is particularly inaccessible to practitioners. This tutorial reviews the methodology and implementation of the estimand framework with the principal stratum strategy to address ICEs with survival outcomes. We illustrate using a clinical trial in oncology and focus on a simple case with binary treatment and a single binary ICE of discontinuation of the assigned treatment. We define the causal effects and review two main methods for estimating
the effects: the mixture model method and the weighting method. For each method, we elaborate the associated assumptions, models, sensitivity analysis, software and provide example R code. We conduct simulation studies that mimic the real study to study the operation characteristics of these methods.

Interference and Consistency Violations

Generalizing Causal Effects under Partial Interference in Two-Stage Sampling Designs
Yihan Bao* Yihan Bao, Laura Forastiere,

Causal inference under interference is increasingly important in public health and social science applications, yet most existing methods are either targeted to finite sample estimands or to super-population estimands implicitly assuming simple random sampling. In practice, many studies rely on complex multistage survey designs, where clusters are sampled with unequal probabilities and only a subset of individuals within each cluster are observed. We develop a general framework under clustered interference for estimating causal effects and generalizing them from the sample to the target population when data arise from such two-stage sampling designs. We define causal estimands as contrasts of average potential outcomes under hypothetical interventions and derive inverse probability weighted estimators that jointly account for treatment assignment, cluster-level sampling, and individual-level sampling. We establish identification conditions under varying assumptions on the interference set, including settings where potential outcomes depend only on sampled units as well as more general interference structures involving non-sampled units. When full treatment information is unavailable, we derive bias expressions and characterize conditions under which consistency is obtained. Simulation studies illustrate the finite-sample behavior of the proposed estimators. The methods are applied to estimate the effects of bed net use on malaria prevalence among children in Uganda.

Machine Learning and Causal Inference

Doubly Robust Estimation of Treatment Effects with Missing Outcomes in Longitudinal Studies Asteria Chilambo* Asteria Chilambo, Zach Branson,

Longitudinal studies are central to understanding dynamic treatment effects, but their analysis is complicated by within-unit temporal dependence and sequentially missing outcomes. Standard methods, such as outcome regression or inverse-probability weighting can address missingness, but may be biased under model misspecification or when nonparametric models are used. Although doubly robust estimators offer protection against such misspecification, existing theory largely focuses on cross-sectional data or single-time-point missingness. We develop a doubly robust framework for estimating mean potential outcomes under a sequential missing-at-random (SMAR) assumption. We derive the efficient influence function for the mean potential outcome under a fixed treatment regime and construct a doubly robust estimator that is root-n consistent and asymptotically normal, provided that cross-fitting is used and the nuisance functions are estimated at n^(-1/4) rates. The estimator allows flexible, data-adaptive estimation of nuisance components. Our method is motivated by and applied to a longitudinal randomized clinical trial of mindfulness-based interventions for irritable bowel syndrome, in which outcomes are collected via smartphone surveys and occasional nonresponse induces complex missingness structure. The method relies on a strong SMAR assumption using observed history (prior non-missing outcomes); future work will relax the assumption to account for unobserved past outcomes.

Machine Learning and Causal Inference

Causal Inference with High-Dimensional Unstructured Treatments Kevin Christian Wibisono* Kevin Christian Wibisono, Yixin Wang,

Causal inference with high-dimensional treatments, such as texts, images, or medical treatment sequences, poses unique challenges: standard causal estimands like the average treatment effect (ATE) are often ill-defined due to overlap violations. Existing approaches typically assume that the treatment of interest is known a priori through pre-defined attributes such as topic or sentiment. In contrast, we propose a data-driven framework that learns the treatment itself. Specifically, we introduce the maximally influential feature (MIF), a latent binary treatment that maximizes the causal effect on the outcome while satisfying overlap. To ensure interventions are meaningful, we decompose each treatment into immutable content and mutable style components, intervening only on the latter. We establish theoretical identifiability of the learned causal estimand, propose a flexible estimator, and introduce a treatment budget that enables the discovery of multiple causal dimensions. Our approach further allows us to nudge or modify treatments in the direction of increased MIF, providing a principled way to causally improve the outcomes. Finally, we demonstrate the effectiveness of our framework across text, image, and treatment sequence applications.

Machine Learning and Causal Inference

Incremental Causal Effects for Time to Treatment Zhichen Zhao* Zhichen Zhao, Zhichen Zhao, Andrew Ying, Ronghui (Lily) Xu,

We consider time to treatment initiation, which commonly arises in preventive medicine, such as disease screening and vaccination, and in non-fatal health conditions, such as HIV infection prior to AIDS onset. While traditional causal inference has focused on deterministic interventions that assign treatment according to fixed rules, including whether or when treatment is assigned and allowing dependence on subject characteristics, we study the incremental causal effect of intervening on the intensity of treatment initiation.

We establish identification and derive the efficient influence function for this effect without requiring the commonly assumed positivity condition. Building on this characterization, we propose efficient nonparametric estimators based on augmented inverse probability weighting that can attain fast convergence rates while accommodating flexible machine learning estimation of nuisance functions. We also develop general efficiency theory for the proposed estimators.

We illustrate the finite-sample performance of our methods through simulation studies and apply them to a rheumatoid arthritis study to evaluate the incremental effect of time to methotrexate initiation on joint pain, as well as to a Norwegian women’s study to evaluate the incremental effect of time to subsequent HPV testing on detection of cervical intraepithelial neoplasia grade 2 or 3 (CIN2+).

Machine Learning and Causal Inference

Causal Inference with Multiple Latent Textual Treatments Arisa Sadeghpour* Arisa Sadeghpour,

Researchers are increasingly interested in understanding the causal effect of texts on human behavior, e.g. the effect of social media posts on persuasion. Recent work introduces a framework for estimating the isolated causal effect of focal language, adjusting for other, non-focal attributes of the text (Lin et al., 2025). While this framework considers settings with only one focal attribute, there are often several focal treatments of interest within texts. As with factorial studies, interaction effects of these focal treatments are often of interest. We leverage recent advances in observational factorial studies (Yu & Ding, 2026) to identify and estimate the isolated causal effects of multiple focal treatments and their interactions. We demonstrate the proposed approach through simulation and applications and offer practical guidance for estimation.

Machine Learning and Causal Inference

Landscape Analysis of the Causal Inference Literature: A Topic Modeling and Bibliometric Study Gabrielle Gauthier-Gagné* Gabrielle Gauthier-Gagné, Tibor Schuster,

Background. The fast-growing causal inference literature makes reviews of methodological approaches and applications challenging and quickly outdated. To overcome this limitation, we used topic modeling and bibliometric analysis to synthesize the causal inference literature.
Methods. We retrieved 349,466 deduplicated records from OpenAlex using causal inference related terms. We applied BERTopic, which uses deep learning embeddings to cluster documents by semantic similarity, to uncover latent topics in article titles and abstracts. We used citation network centrality to estimate topic influence and calculated yearly topic trends.
Results. Topic modeling uncovered 335 topics. The most central topics were methodological including econometric causality, structural equation modeling, Mendelian randomization, and regression discontinuity. Trends emerged in application areas; blockchain technology, gut microbiome, and education technology are growing while agriculture, brain connectivity and tuberculosis show declining prevalence.
Conclusion. Topic modeling enabled quick, transparent, and updateable synthesis of hundreds of thousands of causal inference articles, revealing core methodological topics and fluctuating application domains.

Machine Learning and Causal Inference

Dynamic Conformal Prediction of Survival with Time-varying Covariates Yuyao Wang* Yuyao Wang, Larry Han,

Time-to-event prediction plays a central role in many causal inference problems, where stakeholders care not only whether an event will occur but also how long remains until it does. While conformal prediction provides distribution-free uncertainty quantification for time-to-event outcomes, existing methods often yield only one-sided lower prediction bounds and do not leverage time-varying covariates, limiting their usefulness for dynamic planning and decision-making. In this work, we develop two-sided dynamic conformal prediction intervals for individual event times among survivors that adapt to evolving covariate histories as new information becomes available. The proposed method achieves asymptotically valid marginal coverage at each decision time under the conditional independent censoring assumption given covariate history. Through simulation studies, we show that the resulting dynamic prediction intervals are substantially narrower on average than existing conformal survival approaches while maintaining nominal coverage. We further illustrate the method using publicly available benchmark time-to-event data sets.

Machine Learning and Causal Inference

Debiased Machine Learning for Conformal Prediction of Counterfactual Outcomes Under Runtime Confounding Keith Barnatchez* Keith Barnatchez, Kevin Josey, Rachel Nethery, Giovanni Parmigiani,

Data-driven decision making frequently relies on predicting counterfactual outcomes. In practice, researchers commonly train counterfactual prediction models on a source dataset to inform decisions on a possibly separate target population. Conformal prediction has arisen as a popular method for producing assumption-lean prediction intervals for counterfactual outcomes that would arise under different treatment decisions in the target population of interest. However, existing methods require that every confounding factor of the treatment-outcome relationship used for training on the source data is additionally measured in the target population, risking miscoverage if important confounders are unmeasured in the target population. In this paper, we introduce a computationally efficient debiased machine learning framework that allows for valid prediction intervals when only a subset of confounders is measured in the target population, a common challenge referred to as runtime confounding. Grounded in semiparametric efficiency theory, we show the resulting prediction intervals achieve desired coverage rates with faster convergence compared to standard methods. Through numerous synthetic and semi-synthetic experiments, we demonstrate the utility of our proposed method.

Machine Learning and Causal Inference

Text-as-Treatment Causal Estimation with Sparse Autoencoders Amar Venugopal* Amar Venugopal, Amir Feder, Omri Feldman, Jann Spiess,

Large language models (LLMs) have rich internal representations of language, the study of which can enable the design of controlled experiments with latent language treatments. Recent work in sparse autoencoders (SAEs) allows for intervention on specific concepts embedded in text, generating new texts that vary in the intensities of those concepts. However, these methods are highly sensitive to the choice of concepts and hyperparameters. In this paper we present a novel hypothesis generation methodology that discovers concepts of interest in labeled textual data and identifies the optimal SAE features and layers for such interventions. Using semi-synthetic datasets, we show that the downstream experiments used to validate these hypotheses present a unique challenge for causal inference with latent treatments. Specifically, we demonstrate that the estimation of the conditional average treatment effect (CATE) suffers from significant bias due to inherent positivity violations and treatment leakage. We characterize the estimation bias induced in this setting and propose a solution based on covariate residualization. Our results show that this approach effectively mitigates estimation error, providing a robust foundation for causal effect estimation in text-as-treatment settings.

Machine Learning and Causal Inference

Few-shot causal learning for new treatments and outcomes using task embeddings Sophie Woodward* Sophie Woodward, James Kitch, Claudio Battiloro, Mauricio Tec, Francesca Dominici,

Estimating heterogeneous treatment effects is a central problem in causal inference, with applications in personalized medicine, public policy, and online marketing. Existing methods focus on predicting the effects of fixed treatments on fixed outcomes and do not address settings in which new treatments or new outcomes are introduced. In many such settings, a small amount of data from the new treatment-outcome pair may be available, for example from an early-phase clinical trial. We study the problem of estimating the conditional average treatment effect (CATE) for a new treatment-outcome pair, given limited data and borrowing information from previously observed treatments and outcomes. Specifically, we view CATE estimation for each treatment-outcome pair as a task, and propose a framework that uses task embeddings—vector representations that encode structural or semantic relationships across treatments and outcomes—to predict the CATE function across tasks. We subsequently estimate the CATE for the new task by combining the embedding-based CATE predictor learned across tasks with a CATE estimator fit using data from the new task alone. This yields a data-fusion estimator that can reduce variance relative to task-only estimation under some regularity conditions. Experiments on semi-synthetic benchmarks and large-scale medical claims data evaluate performance and illustrate the roles of covariate shift, number of tasks, task sample size, and embedding distance.

Machine Learning and Causal Inference

From Iterative Targeting to One-Step Updates: Convex-Dual Affine Universal Least Favorable Models for Heterogeneity, Distributional, and Policy-Risk Estimands Kaiwen HOU* Kaiwen HOU, Mark van der Laan,

Modern causal inference increasingly targets nonlinear estimands, for which valid inference with flexible nuisance learning relies on semiparametrically efficient procedures built around the efficient influence function (EIF). In practice, efficiency is often pursued via iterative local targeting in TMLE updates, repeatedly computing the EIF and taking small steps, which can be computationally costly and numerically unstable.

We identify a broad class of settings where the universal least favorable model (ULFM) admits a semi-explicit solution. The key structure is that the ULFM score equation reduces to an equation pointwise in the nuisance functions and coupled only through finitely many scalar moments (e.g., normalizing constants or low-order moments). We term these mean-field ULFMs. This structure yields one-step TMLE updates that enforce the EIF score condition without iterative recomputation, often requiring only the solution of a low-dimensional auxiliary ODE.

We derive mean-field ULFMs for: i) density functionals, including density power integrals and distributional treatment effect; ii) heterogeneity functionals, including centered moments of CATE; iii) policy risk functionals, including the variance of policy value; and iv) propensity and overlap functionals. Simulations show that ULFM-based one-step targeting can improve numerical stability and reduce sensitivity to step-size tuning relative to iterative local targeting while maintaining finite-sample performance.

Machine Learning and Causal Inference

Screens to Smarts: Regularized Apprenticeship Learning with Attention for Inferring Behavioral Pathways Linked to Health Knowledge Gains in a Digital Intervention Rahul Ladhania* Rahul Ladhania, Rema Padman, Shreyas Vajjhala,

Gamified digital interventions are increasingly used to promote health knowledge and behavior, yet limited evidence exists on which in-intervention behavioral pathways are linked to downstream learning outcomes. We analyze gameplay telemetry from an 11-week school-based digital health intervention in India to discover interpretable gameplay strategies associated with improvements in children's health knowledge. We model gameplay trajectories as behavioral sequences, and develop a novel inverse reinforcement learning framework that incorporates L1 regularization and an attention mechanism to extract interpretable latent behavioral patterns from high-dimensional, small-sample telemetry data. Survey-based knowledge gains serve as a downstream reward signal to recover latent gameplay strategies associated with improved learning outcomes. We find that students who more consistently engage with health-aligned mechanics (e.g., exercising, using power-ups effectively) exhibit larger knowledge gains. Incorporating regularization and attention substantially improves alignment between inferred behavioral patterns and observed outcomes, yielding transparent representations that highlight the most behaviorally relevant gameplay features. Our findings demonstrate how telemetry-driven behavioral modeling can support mechanism discovery and hypothesis generation, informing design of future experiments, adaptive interventions, and personalization strategies in digital health and education.

Machine Learning and Causal Inference

Random Forest Counterfactual Similarity for Causal Inference Bernardo Modenesi* Bernardo Modenesi, Sima Najafzadehkhoei,

Causal inference fundamentally relies on the ability to construct credible counterfactuals, i.e., to identify which units are “similar enough” in pre-treatment covariates to support causal comparisons. In practice, similarity is often imposed using Euclidean distances on standardized covariates, linear regression adjustment, or one-dimensional balancing summaries such as the propensity score. These choices can be brittle in tabular data with nonlinear interactions, heterogeneous feature relevance, mixed variable types, and missingness, where small changes in preprocessing or feature scaling can substantially alter which counterfactuals are deemed comparable. We propose a proximity-based notion of counterfactual similarity learned from random forests, yielding a data-adaptive metric that emphasizes covariate dimensions that matter for partitioning the population while down-weighting irrelevant variation. We show how this learned similarity can be used as a unifying primitive for several causal workflows: (i) proximity-based matching and stratification, (ii) proximity-weighted estimators that localize adjustment to data-supported neighborhoods, and (iii) estimation of heterogeneous treatment effects by comparing proximity-defined counterfactual outcome models across treatment groups. We further introduce diagnostics for counterfactual quality based on local overlap and neighborhood stability, enabling transparent assessment of where causal conclusions are well supported.

Machine Learning and Causal Inference

Replay and Ground: Causal Offline Evaluation of Language Models Jikai Jin* Jikai Jin, Vasilis Syrgkanis,

Evaluating language models offline is challenging when deployment logs are confounded: routing decisions depend on latent user or task factors that simultaneously influence quality outcomes, making naive observational comparisons unreliable. We address this problem in a setting where three data sources are available: a large confounded observational log (OBS), a small randomized experiment (EXP), and an offline replay simulator, and the evaluation target is each language models's causal value: the expected reward under a policy that routes all traffic to that agent.
We make two main contributions. First, we show that causal value is nonparametrically identified by combining replay-generated mediators with a reward surface fit on randomized EXP data, without requiring the observational log to be unconfounded. Second, building on this identification result, we develop hybrid reward-model estimators that exploit OBS at scale — either by learning a deconfounded representation from OBS auxiliary labels, or by grounding an OBS-trained reward model with a small EXP-estimated bias correction — and pair these with both a direct plug-in and a doubly robust value estimator. Empirically, no single estimator uniformly dominates: hybrid methods exhibit systematic, predictable crossovers across reward nonlinearity, confounding strength, and EXP budget.

Machine Learning and Causal Inference

Surrogate Augmentation for Causal Inference on Censored Survival Outcomes Yaroslav Mukhin* Yaroslav Mukhin, Tereza Oprea, Arielle Anderer, Christina Yu, Jelena Bradic,

Missing data on a long-horizon outcome is a common challenge for evaluating the impact of an intervention: Experiments are constrained by budgets and timelines; observational studies face drop-out. We show how to leverage surrogate outcomes to increase power for causal effect estimation with a right-censored survival outcome. Censoring creates a loss of information, but there is also an opportunity to regain statistical efficiency by employing auxiliary variables. E.g., in a clinical trial evaluation of the effect of a cancer treatment on survival, disease progression, or absence thereof, is informative of a censored survival time. With missing outcomes, covariates become informative for causal parameters that, absent missing data, do not depend on the latter. This result holds without structural or semiparametric restrictions, e.g., full-mediation or proportional hazards assumptions, but the size of the gain depends on the predictability of the missing outcome by the covariate. Surrogate outcomes are informed by the effect of the treatment, and allow conditioning on the information set at the time of censoring, strictly improving on the gains from baseline covariates. We derive an efficiency bound that reveals the interplay between (i) survival hazard, (ii) censoring hazard, and (iii) time-adapted forecasts of the primary event-time. We demonstrate the gains with semi-synthetic data from 93 metastatic breast cancer clinical trials.

Machine Learning and Causal Inference

Comparing Causal Forest and BART for Estimating Treatment Effect Heterogeneity with Cluster Data: An Application to LLM Evaluation JIA QUAN* JIA QUAN, Walter Leite,

Causal Forests (Wager & Athey, 2018) and Bayesian Additive Regression Trees (BART; Hahn, Murray, & Carvalho, 2020) based causal inference are widely used to estimate conditional average treatment effects (CATEs), yet they operationalize heterogeneity differently. As large language models (LLMs) are increasingly deployed in high-stakes domains, rigorous causal evaluation of model choices becomes essential. We apply both methods to data from an educational application generating decodable reading passages to estimate heterogeneous effects of a fine-tuned versus off-the-shelf LLM on story quality across prompt types, grade levels, and linguistic benchmarks. This application involves data conditions frequently encountered in AI evaluation research: clustered observations, high-dimensional moderators, skewed outcomes, and complex interactions. We identify three implementation issues that critically affect BART's performance: outcome-scale sensitivity, cluster parameterization, and estimand specification in which marginalizing over random effects attenuates individual CATEs. After resolving these, both methods recovered nearly identical ATEs (Causal Forest: g = 0.213; BART: g = 0.210), with strong agreement in ITE ranking and moderator importance.
Our findings demonstrate how modern causal ML methods can provide principled, heterogeneous treatment effect estimates for generative AI systems, offering a framework for cross-disciplinary evaluation of model deployment decisions.

Marginal Structural Models

Efficient Counterfactual Mean Estimation Implies Efficient Marginal Structural Model Estimation Jacob M Chen* Jacob M Chen, Ilya Shpitser,

Marginal structural models (MSMs) are a class of causal models that allow for the estimation of causal effects from observational data that are widely used due to their interpretability as well as generalizability to continuous treatments and multiple treatments in longitudinal settings. An MSM posits that the expectation of the counterfactual outcome had treatments been intervened on – sometimes also conditional on a subset of observed pretreatment covariates – is a function indexed by a finite set of parameters. We aim to estimate the parameters of this function, which allows us to infer causal quantities of interest, such as the average causal effect (ACE) or conditional ACE. Estimation strategies for parameters of an MSM are well studied under the coarsening at random (CAR) assumption, which states that all variables affecting treatment assignment are observed. Here, we show how to estimate the parameters of any MSM as long as the counterfactual mean is identified even when CAR does not hold, such as in the frontdoor and proximal causal learning settings, using a loss minimization technique. Our estimator for the MSM parameters inherits desirable properties from the mean estimator, if the mean estimator possesses such qualities, such as robustness to misspecification of a subset of nuisance models and asymptotic normality. Our results will allow practitioners to employ MSMs in a wider range of settings, especially when unmeasured confounding is unavoidable.

Matching, Weighting

Weighting-based Identification and Estimation Techniques in Graphical Models of Missing Data Anna Guo* Anna Guo, Razieh Nabi,

In this paper, we propose a novel algorithm to identify complete data distributions in graphical models of missing data, without imposing any restrictions on the complete data distribution and only requiring the missingness mechanisms to factorize according to a conditional directed acyclic graph. Our view aligns with prior work on missing data that frames identification using causal graphical models with hidden variables, where missingness indicators are viewed as “treatments” that could potentially be intervened on.
Selection bias is the primary obstacle to identification in missing data models under the interventionist perspective. To address this, our identification algorithm generates a tree data structure that facilitates tracking selection bias and provides insight into how it can be avoided. Building on this framework, we develop recursive weighting strategies for estimating missingness mechanisms and for conducting statistical analyses of the complete data law, extending inverse probability weighting methods to missing-not-at-random settings. We demonstrate the effectiveness of our approach through simulation studies, comparing it with classical methods such as multiple imputation and the EM algorithm across a range of analysis tasks. An accompanying R package, flexMissing, implements all proposed procedures.

Matching, Weighting

Low-rank Covariate Balancing Estimators under Interference Souhardya Sengupta*
Souhardya Sengupta, Kosuke Imai, Georgia Papadogeorgou,

A key methodological challenge in observational studies with interference between units is twofold: (1) each unit's outcome may depend on many others' treatments, and (2) treatment assignments may exhibit complex dependencies across units. We develop a general framework for constructing robust causal effect estimators to address these challenges. We first show that, without restricting the patterns of interference, the standard inverse probability weighting (IPW) estimator is the only uniformly unbiased estimator when the propensity score is known. In contrast, no estimator has such a property if the propensity score is unknown. We then introduce a "low-rank structure" of potential outcomes as a broad class of structural assumptions about interference. This framework encompasses common assumptions such as anonymous, nearest-neighbor, and additive interference, while flexibly allowing for more complex study-specific interference assumptions. Under this low-rank assumption, we show how to construct an unbiased weighting estimator for a large class of causal estimands, even when the true propensity score is unknown. If the true propensity score is known, we can obtain an unbiased estimator that is more efficient than the IPW estimator by leveraging a low-rank structure. We establish finite sample and asymptotic properties of the proposed estimators, develop a data-driven procedure to select among candidate low-rank structures, and validate our approach through empirical studies.

Matching, Weighting

Quantifying Practical Overlap in Causal Inference via KL Projections Geondo Park* Geondo Park, Juyeon Kim, Kwonsang Lee,

Assessing overlap is an essential task in causal inference, as limited overlap undermines identifiability and leads to unstable estimators. In practice, overlap is most often evaluated through visual inspection of propensity score distributions. Although overlap is frequently discussed in connection with methods that improve estimation, such as trimming or overlap weighting, there is limited work on directly measuring how much overlap is present.
We propose a likelihood-based framework for quantifying practical overlap using Kullback-Leibler projections. The approach defines a common component as the distribution that best approximates the treated covariate distribution while remaining representable as a mixture component of the control population. This construction yields a smooth, distribution-level characterization of overlap that avoids explicit density estimation. An overlap parameter is defined as the largest fraction of the control population that can be retained while maintaining sufficient distributional proximity to the treated group. Simulations and empirical examples with known overlap challenges illustrate how the proposed measure provides a principled early-stage diagnostic prior to causal effect estimation.

Matching, Weighting

Weighted average treatment effect with unknown weights Georgy Kalashnov* Georgy Kalashnov,

I derive semi-parametrically efficient estimate of a weighted average treatment effect, where weights depend on the data generating process and therefore are unknown. Examples of such estimates include average treatment on the treated, overlap weights. The need to (whether explicitly, or implicitly) estimate the weights restricts the outcome adjustment function we can use to keep the efficiency. E.g. for ATE we should use the prediction of the countrerfactual that is not observed as an adjustment function, for ATT — the prediction of control outcome, for overlap weights, the prediction of observed Y. This restriction creates three interpretable terms in the asymptotic variance of the estimate: a weighted variance of the idiosyncratic error, a (weighted) variance of the conditional average treatment effect, and most importantly, the systematic error created by the forced choice of adjustment function. This results is useful in several ways: 1) it nests and extends a large number of results, which either concentrate on some specific weights, e.g. Hahn (1998), Freedman (2009), Lin (2013), or on a known weighting function, e.g. Li, Morgan, Zaslavsky (2018), Chernozhukov, Newey, Singh (2022) 2) it allows to inform bias variance tradeoff in the task of choosing an estimand to pursue (e.g. should we use ATE or ATT, or overlap weights), under different additional assumptions on the data generating process.

Matching, Weighting

Development and Evaluation of Ensemble Propensity Score Matching: A Comparison with the Covariate Balancing Propensity Score Yasutaka Hasegawa* Yasutaka Hasegawa, Takanobu Osaki, Hideyuki Ban, Takayuki Arai, Tsutomu Kikuchi,

Covariate balance is essential for estimating the average treatment effect on the treated (ATT) from observational data. Propensity score matching (PSM) is widely used, but reliance on a single propensity score (PS) model can be brittle under model misspecification. We propose Ensemble Propensity Score Matching (Ensemble PSM), which fits multiple PS estimators and selects the matched sample that minimizes overall imbalance. We estimate PS using logistic regression, LASSO, elastic net, gradient-boosted trees, and a neural network; for each PS we create 1:1 caliper matches without replacement, compute standardized mean differences (SMDs) across all covariates, and select the candidate match with the smallest mean SMD. We evaluated the method using a real-world health guidance dataset (treated n=3,574; controls n=9,668; 31 covariates) and benchmarked it against CBPS-PSM (PS estimated via the covariate balancing propensity score, then matched under the same 1:1 caliper-without-replacement design) targeting the ATT. Ensemble PSM achieved a mean SMD of 0.00266 and a maximum SMD of 0.00754 versus 0.00796 and 0.02381 for CBPS-PSM, corresponding to 66.6% and 68.3% reductions, respectively. Ensemble PSM also outperformed single-model PSM baselines (e.g., logistic-regression PSM: 0.01123/0.02775). These findings suggest that balance-driven ensemble selection can improve the robustness of covariate adjustment for causal effect estimation in observational health service evaluations.

Matching, Weighting

Resampling with Control Reuse: A Valid Bootstrap for Fixed-M Nearest-Neighbor Matching
Xiang Meng* Xiang Meng, Aaron Smith,

Inference after fixed-M nearest-neighbor matching remains challenging because matching induces nonstandard dependence through control reuse. Abadie and Imbens (2008) showed that the naive bootstrap fails when the number of matches is fixed, prompting a substantial literature proposing alternative resampling methods. More recently, Lin and Han (2024) establish that the naive bootstrap is consistent when the number of matches diverges, demonstrating that the classical inconsistency is fundamentally a fixed-M phenomenon. While valid bootstrap procedures are now available when M goes to infinity, a general and practically implementable solution for valid inference under fixed M—the regime most commonly used in applications—remains lacking.

We therefore study inference directly in the fixed-M setting and develop a weighted unit bootstrap for the Average Treatment Effect on the Treated (ATT) that remains valid under fixed-M matching. The central insight is that valid resampling must replicate the unit-level covariance structure induced by shared controls. Rather than re-matching the data or perturbing outcomes independently, the proposed procedure resamples entire units together with their induced matching weights, thereby preserving the dependence created by control reuse.

We show theoretically that the weighted unit bootstrap consistently approximates the asymptotic variance in the general fixed-M framework. This perspective also yields a unified explanation for the failure o

Mediation Analysis, Mechanisms

Validating Causal Mechanisms through Replicability: A Unified Bayesian Framework for Mediation Analysis Ester Alongi* Ester Alongi, Gianmarco Altoè, Giovanni Parmigiani,

How can we ensure that an estimated mediation effect has a causal interpretation? Causal mediation analysis is a powerful tool for understanding the mechanisms transmitting an effect from an exposure to an outcome. However, estimating the natural indirect effect relies on the nonrefutable assumption of sequential ignorability, specifically the absence of unmeasured mediator-outcome confounding. While various sensitivity analyses have been proposed to assess the robustness of mediation estimates against hypothetical assumption violations, they do not directly address whether such confounding is present.
We propose a systematic replicability approach as an empirical tool to detect whether an estimated mediation effect reflects sample-specific unmeasured mediator-outcome confounding. We introduce a unified Bayesian hierarchical framework embedding causal mediation within a multifaceted replicability structure evaluating four dimensions: natural indirect effect consistency across independent studies; meta-analytic inference implied by a common generative model; consistency between studies and the shared meta-analytic structure; consistency between an existing body of evidence and a new study from the same generative model.
We apply this framework to a Mendelian randomization case study using GTEx data, evaluating whether a genetic variant’s effect on downstream metabolic genes in subcutaneous adipose tissue is causally mediated by the expression of the KLF14 transcription factor.

Mediation Analysis, Mechanisms

Efficient estimation of pathway effects mediated by intermediate events in multi-state models Yuhao Deng* Yuhao Deng, Haoyu Wei, Donglin Zeng, Rui Song, Xiao-Hua Zhou,

Cardiovascular and microvascular events are leading causes of death in patients with type 2 diabetes. While a randomized controlled trial indicated that X reduces cardiovascular and microvascular risks as well as mortality, the mechanism by which X prevents vascular events and death is not fully understood. In this work, we consider hypothetical interventions in each transition of disease progression. Through such interventions, we distinguish the effects along specific pathways from the total effect on each event. Our proposed framework enables three key applications: estimating path-specific treatment effects, identifying which events are influenced by treatment, and inferring dynamic treatment strategies. Based on multi-state models, we derive multiply robust, nonparametrically efficient estimators for the counterfactual cumulative incidences and treatment effects, accompanied by inference procedures. By analyzing data from a randomized controlled trial, we find that X significantly reduces the risk of non-fatal expanded major adverse cardiovascular events as well as microvascular events. Beyond existing results, our new methods recover the following potentially useful clinical findings. The reduction in all-cause mortality associated with X is primarily mediated by its effects on expanded major adverse cardiovascular events, and importantly, sustained adherence to X is crucial for achieving an effective reduction in cardiovascular risk

Multilevel Causal Inference

Estimation, Inference, and Sensitivity for Spillover Effects in Two-Stage Observational Studies via Matching Zhiwei Xiao* Zhiwei Xiao, Samuel Pimentel,

Spillover effects play a crucial role in driving large-scale substantive impacts of treatments in the health and social science, yet their estimation remains challenging in observational studies with complex interference. Two-stage observational designs, in which both clusters and units within clusters select into treatment, provide valuable opportunities to measure such treatment effects. However, standard tools for confounding control and permutation inference do not apply directly to these studies because of the two-stage structure. We introduce two new study designs for two-stage observational studies that enable unbiased point estimation of spillover effects, valid finite-sample permutation inference, and sensitivity analysis for unobserved confounding. The first design leverages cluster-level matching followed by random selection of a subset of control subjects under a propensity model. This estimator is compatible with permutation tests, and we show that it is marginally unbiased across all possible realizations of the random subset. The second design employs multi-level matching, eliminating the need for random selection of controls at the cost of a reduced sample size. The two-stage sensitivity analyses we develop for both designs allow separate quantification of unmeasured confounding at the cluster and individual levels. We demonstrate our methods using a two-stage study of deworming medication in Kenyan primary schools.

Multiple treatments/positivity violations

Vector Incremental Treatment Effects for Causal Inference with Multiple Binary Treatments Denis Agniel* Denis Agniel, Max Rubinstein, Sharon-Lise Normand, Marcela Horvitz-Lennon,

Estimating causal effects with m > 1 binary treatments faces challenges: positivity violations when combinations are rare, exponential growth in potential outcomes (2^m), and difficulty choosing meaningful treatment contrasts. We extend incremental propensity score interventions to vector treatments, defining causal effects through intervention distributions scaling odds of each treatment. These effects capture the joint effect of the many treatments. The method selects an intervention treatment distribution Q_δ minimizing a user-specified divergence (e.g., KL, Hellinger, f-divergence) from the original propensity score distribution P_0 subject to marginal odds constraints: Q_δ(A_k = 1|X) / {1 – Q_δ(A_k = 1|X)} = δ_k × P_0(A_k = 1|X) / {1 – P_0(A_k = 1|X)}. This preserves the joint treatment distribution in the intervention distribution and allows the intervention to affect treatments differently. We derive the efficient influence function, which is not a straightforward extension of the univariate incremental propensity score one, and derive doubly robust estimators from it. The efficient influence function accounts for uncertainty in nuisance estimation as well as Q_δ computation. We establish asymptotic normality and provide inference. We apply these methods to study the effect of quality of care (captured by many binary indicators of quality) on clinical outcomes (treatment discontinuation, mortality) for New York Medicaid patients diagnosed with schizophrenia.

Policy Learning

Treatment Policy Design in the Presence of Measurement Error Chang Liu* Chang Liu, Mats Stensrud, AmirEmad Ghassami,

In many applications, treatments are assigned based on unit features, leading to personalized treatment policies. This often involves optimizing objective functions with counterfactual quantities such as the conditional ATE (CATE). In practice, some key features may be measured with error and ignoring such errors can introduce systematic bias. We study how to design personalized treatment policies when some features are latent and only noisy measurements are available. We ask how policies can be constructed when the CATE is not identifiable, and how uncertainty from measurement error can be incorporated into policy learning. We propose two frameworks for treatment policy design based on partial identification, focusing on the measurement mechanism—the conditional distribution of observed measurements given unobserved features. The first framework builds on ideas from proximal causal inference. The second adopts a Bayesian approach, using a reparametrization of the likelihood to obtain the posterior distribution of the parameters of interest. We use data augmentation for latent variables and a Gibbs sampler with Hamiltonian Monte Carlo updates for model parameters. Quantiles of the posterior distribution are used to construct bounds for the CATE. Treatment policies are then derived by comparing the null value of the CATE to these bounds. The proposed methods provide a coherent way to account for measurement error and model uncertainty in personalized treatment policy learning.

Policy Learning

Optimal Plug-in Treatment Rules under Heterogeneous Network Interference Elena Dal Torrione* Elena Dal Torrione, Laura Forastiere,

This paper studies optimal treatment policies under heterogeneous network interference, where treating different individuals affects welfare differently through their direct response and influence on others. In settings without interference, the welfare-maximizing policy treats individuals with positive conditional average treatment effects. Under interference, however, individual outcomes depend on others' treatments, and simple plug-in treatment rules are generally unavailable without further assumptions on the underlying potential outcome structure. We first derive general first-order optimality conditions in terms of conditional direct effects for units with given covariates and conditional indirect effects on other units. Unlike no-interference settings, we show that under interference optimal treatment rules may be stochastic, and we provide sufficient conditions for when the optimal policy is stochastic or deterministic. Next, we propose a quadratic outcome model that incorporates pairwise treatment interactions and nests standard linear-exposure models. We derive conditions under which the optimal policy admits a closed-form characterization and discuss implications for plug-in treatment rule estimation. Finally, we extend our analysis to welfare maximization under budget constraints and to the complementary problem of minimizing budget subject to outcome constraints.

Policy Learning

Off-policy evaluation using debiased calibration Jae-kwang Kim* Jae-kwang Kim, Yuyang Li, Yumou Qiu,

Off-policy evaluation (OPE) estimates the expected performance of an evaluation policy using data collected under a distinct behavior policy, enabling policy assessment when direct experimentation is infeasible. Standard OPE methods typically assume stable covariate distributions across datasets—a condition often violated in practice due to temporal or environmental data shifts.
This paper introduces a generalized entropy calibration (GEC) weighting framework to improve OPE under covariate shift. A self-normalized estimator with GEC weighting is first developed for stationary settings, and its efficiency and double-robustness properties are established. Building on this foundation, a doubly-weighted estimator is proposed to further correct for selection bias induced by covariate shift, using two sets of calibration weights to sequentially adjust for policy and covariate discrepancies across datasets.
The resulting multi-calibrated framework accommodates multiple data-transfer mechanisms and attains multiple robustness. Extending beyond conventional doubly robust estimators, the proposed method achieves table-wise quadratic robustness, ensuring consistency when any one of four model combinations is correctly specified. We also provide an alternative proof of the semiparametric efficiency bound for OPE under covariate shift and show that the proposed estimator attains this bound when all models are correctly specified.
Theoretical validation and numerical studies are provided

Proximal Causal Inference

Proximal Inference for Hidden Outcomes Helen Guo* Helen Guo, Ilya Shpitser, Elizabeth Ogburn,

Nonparametric proxy variable methods have emerged as a powerful approach for addressing hidden variables in causal inference. One influential line of this work recovers latent factors which compose the target functional via eigendecomposition tasks. Within this framework, we identify the full law under discrete hidden outcomes and develop influence-function–based estimators for causal effects, resulting in Neyman-orthogonal estimation with desirable efficiency properties. Although methods for handling missing and mismeasured outcomes exist, this work provides what we believe to be the first proximal inference results for hidden outcomes. We illustrate our estimation approach through simulation studies.

Randomized Designs and Analyses

Toward a Causal Framework for Crossover Trials: From Estimands to Estimation Richard Liu* Richard Liu, Mats Julius Stensrud, Michele Santacatterina,

Crossover trials are commonly used in clinical research to improve efficiency in small sample settings, as each participant serves as their own control across treatment periods. Despite their widespread use, however, a principled framework for defining, identifying, and estimating causal effects in this design remains lacking. In particular, causal estimands are often poorly defined, and the identification assumptions required are rarely made explicit. As a result, it is often unclear which estimand is being targeted and under what conditions commonly used estimation procedures achieve efficiency gains or introduce bias. A key example of this ambiguity lies in the definition and identification of carryover effects and washout periods, which are inconsistently handled across both study designs and analytical approaches. This methodological gap has been highlighted in recent FDA guidance advocating for the use of a formal estimand framework. In this work, we leverage tools from causal inference and modern statistical methodology to address these challenges. We propose a set of causal models tailored to the crossover design, clarify the corresponding causal estimands and their identification assumptions, and examine when canonical statistical methods yield unbiased and efficient estimates. In addition, we also introduce doubly robust estimators to further enhance robustness while preserving efficiency. Further, we formalize the notion and the role of carryover effects.

Randomized Designs and Analyses

Compound Causal Selection Decisions: An Almost SURE Approach Timothy Sudijono* Timothy Sudijono, Jiafeng Chen, Lihua Lei, Liyang Sun, Tian Xie,

This paper proposes methods for producing selection decisions in a Gaussian sequence model. These decisions directly target the following problem: given unknown, fixed parameters $mu_{1:n}$ and known $sigma_{1:n}$ with observations $Y_i sim textsf{N}(mu_i, sigma_i^2)$, the decision maker would like to
select a subset of indices S so as to maximize utility $sum_{iin S} (mu_i – K_i)$,
for some known costs K. This problem appears in the management of experimentation programs: large collections of related A/B tests in the online technology industry where selections are made based on ATE estimates. It also appears as a special case of empirical welfare maximization, where subpopulations are selected on the basis of CATE estimates.

Inspired by Stein’s unbiased risk estimate (SURE), we introduce an almost unbiased estimator, called ASSURE, for the expected utility of a proposed
decision rule. ASSURE allows a user to choose a welfare-maximizing rule from a pre-specified class by optimizing the estimated welfare, thereby producing selection decisions that borrow strength across noisy estimates. We show that ASSURE produces decision rules that are asymptotically no worse than the optimal but infeasible decision rule in the pre-specified class. We apply ASSURE to the
selection of Census tracts for economic opportunity, the identification of discriminating firms, and the analysis of p-value decision procedures in A/B testing.

Randomized Designs and Analyses

Improving Variance Estimation for Covariate Adjustment with Binary Outcomes Kaitlyn Lee*
Kaitlyn Lee, Courtney Schiffman, Alex Ocampo, Michel Friesenhan, Christina Rabe, Michael Rosenblum,

Covariate adjustment is a general method for improving precision when estimating treatment effects in randomized trials and is recommended by the FDA in its recent guidance when baseline variables are prognostic for the primary outcome. We focus on a method highlighted in that guidance called “standardization” (or “g-computation”) for estimating the marginal treatment effect. We address the question of how to reliably estimate variance for binary outcomes when marginal outcome probabilities are close to 0 or 1. We propose an influence function-based leave-one-out cross-validated (LOO-CV) variance estimator for the standardized difference-in-means average treatment effect. Through simulation studies, we show that this estimator provides appropriate type I error control and performs reliably in challenging settings where existing methods can yield inflated type I error or fail entirely, such as when outcome events are rare or sample sizes are small. In addition to having desirable statistical properties, we derive a closed-form expression for the proposed estimator, enabling straightforward and reliable implementation by study statisticians. The robust finite-sample performance and ease of implementation suggest the LOO-CV variance estimator is a prudent default choice for standardization in clinical trials.

Randomized Designs and Analyses

Design-Aware Variance Reduction for Switchback Experiments: A Comparative Study Sergei Pankratev* Sergei Pankratev,

Switchback experiments and other clustered randomized designs are widely used on online platforms, but the clustered, time-dependent nature of these designs can make standard variance reduction methods behave differently than in standard A/B tests. We evaluate design-aware variance reduction methods for switchbacks—CUPED, CUPAC (ML-based covariate adjustment), and doubly robust (DR) estimators—relative to a baseline switchback analysis with cluster-robust standard errors. Through a hierarchical simulation framework that varies key regime parameters—number of clusters, cluster-size imbalance, within-cluster autocorrelation, carryover, and predictive signal strength—we evaluate validity (false positive rate and confidence interval coverage) and efficiency (standard error reduction, power, and minimum detectable effect as a function of run length). We also include a sensitivity analysis for cross-cluster spillovers to quantify bias and inference degradation under mild interference. The primary outcome is a practitioner-oriented regime map: when CUPED, CUPAC, or DR are most beneficial, and when time and cluster dependence and finite-cluster effects limit improvements.

Randomized Designs and Analyses

Learning Treatment Effects while Treating under Priority Queues Johnna Sundberg* JungHo Lee, Johnna Sundberg, Bryan Wilder,

A recurring challenge in social programs is allocating scarce resources, such as housing assistance, under uncertainty about the program's benefits. In practice, resources are often prioritized toward individuals judged to have higher need, and applicants are commonly first categorized into priority tiers rather than randomized outright. Motivated by this, we introduce an experimental design that randomizes incoming applicants into priority queues using assignment probabilities based on their risk scores; within each queue, treatment offers are made in priority order and then first-in-first-out as budget becomes available. We formulate the choice of queue-assignment probabilities as a convex optimization problem for efficient estimation of the average treatment effect and priority-based local effects under noncompliance, formalizing the trade-off between more conservative identification assumptions and achievable statistical power. The framework will be deployed with the Allegheny County Department of Human Services in Pittsburgh, Pennsylvania, to provide housing assistance to people experiencing homelessness.

Regression Discontinuity Designs

Distributional Discontinuity Design Kyle Schindl* Kyle Schindl, Larry Wasserman,

We introduce distributional discontinuity design, a framework for studying distributional causal effects for a scalar outcome at the boundary of a discontinuity in treatment assignment (a generalization of the regression discontinuity design). Our causal estimand is the Wasserstein distance between limiting conditional outcome distributions above and below the treatment discontinuity; a single scale-interpretable measure of distribution shift. We show that this weakly bounds the average treatment effect, where equality holds if and only if the treatment effect is purely additive. Moreover, we show that the Wasserstein distance can be decomposed into squared differences in L-moments, thereby quantifying the contribution from location, scale, skewness, etc. to the overall distributional distance. This decomposition provides a novel way of encoding the heterogeneity in the treatment effect. Next, we extend this framework to distributional kink designs by evaluating the Wasserstein derivative at a deterministic policy kink; this describes the flow of probability mass through the kink. In both settings, we allow the treatment assignment to be either sharp or fuzzy. Finally, we apply our method on real data by re-analyzing several natural experiments to compare our distributional effects to traditional causal estimands.

Regression Discontinuity Designs

Bayesian Inference in Longitudinal Regression Discontinuity Designs Laura Forastiere*
Alessandra Mattei, Laura Forastiere, Fabrizia Mealli,

We study longitudinal regression discontinuity (RD) designs in which treatment is dynamically assigned over time through a sequence of cutoff rules. We focus on a two-period setting,allowing the forcing variable in the first period to affect the forcing variable in the second period. Within the potential outcomes framework, we formally characterize longitudinal RD designs as local sequentially latent regular designs. We define causal estimands that capture the effects of treatment sequences for well-defined, though generally unobserved, subpopulations where local overlap conditions and local versions of SUTVA hold. Inference relies on local longitudinal unconfoundedness, imposing conditional independence between potential outcomes for the main outcome and (a) the first-period forcing variable, and (b) the potential outcomes of the second-period forcing variable. To identify the subpopulations for which valid causal inference is possible, we extend the Bayesian model-based finite mixture approach proposed by Forastiere et al. (AOAS, 2025) to the longitudinal setting, probabilistically clustering observations into subpopulations in which the required assumptions are more or less likely to hold, based on their observable implications. We then derive posterior distributions of the target causal effects by marginalizing over uncertainty in subpopulation membership. We apply the proposed methodology to the evaluation of Italian university student-aid policies on academic outcome.

Sensitivity Analysis

On the role of the Potential Outcome Association Structure for Principal Causal Effects
Arianna Nuti* Arianna Nuti, Alessandra Mattei, Bjoern Bornkamp, Tianmeng Lyu,

In clinical trials, subgroup analyses based on biomarkers that may lie on the causal pathway between treatment and outcome can provide valuable clinical insight. Assessing causal effects among patients who respond to treatment according to their biomarker levels, namely those with a sufficient biomarker reduction, may help physicians tailor therapy early after treatment initiation. We formalize this research question using the principal stratification framework, recognized in the ICH E9(R1) addendum as an estimand strategy for addressing post-treatment variables. Within this framework, the target causal estimand is the principal average causal effect for the principal stratum of responders, defined as patients whose biomarker value under treatment would fall below a pre-specified threshold. We investigate how this principal causal effect depends on the association structure between the potential outcomes for the biomarker and the primary endpoint Assuming a multivariate normal joint distribution of all potential outcomes, we explicitly express the causal effect as a function of the association parameters. We derive a closed-form formula, showing the sensitivity to non-identifiable association parameters. Our results provide insights into the role of the association parameters and highlight how causal conclusions may depend on assumptions about them. These findings may be informative even beyond the normal setting to assess the plausibility of this type of assumptions.

Sensitivity Analysis

Causal Effects of Modified Treatment Policies under Limited Overlap: A Partial Identification Approach Taehyeon Koo* Taehyeon Koo, Kara Rudolph, Caleb Miles,

Modified treatment policies (MTPs) define causal effects of interventions on continuous or multivariate treatments by mapping observed exposures to counterfactual values. Identification of MTP effects typically relies on positivity assumptions that are frequently violated in practice, particularly in settings with high-dimensional or highly correlated treatments. We propose a partial identification framework for causal effects of modified treatment policies under limited overlap. Our approach decomposes the target estimand into a point-identified component over regions where overlap holds and a non-identifiable component arising from extrapolation beyond these regions. For the latter, we derive partial identification bounds under the assumption of Lipschitz continuity on the conditional mean of potential outcomes, extending smoothness-based identification methods to multivariate continuous treatments. We show that the length of these bounds is minimized by projecting counterfactual treatment–covariate values onto the overlap region and introduce a regularized projection operator to address the non-regularity induced by hard projections. We develop efficient influence function–based estimators for the resulting bounds, establish their asymptotic normality, and construct valid confidence intervals for partially identified causal effects. Simulation studies demonstrate favorable finite-sample performance, and an application to pesticide data from the CHAMACOS cohort.

Sensitivity Analysis

Bounding causal effects with an unknown mixture of informative and non-informative missingness Max Rubinstein* Max Rubinstein, Denis Agniel, Larry Han, Marcela Horvitz-Lennon,

In experimental and observational data settings, researchers often have limited knowledge of the reasons for missing outcomes. To address this uncertainty, we propose bounds on causal effects for missing outcomes, accommodating the scenario where missingness is an unobserved mixture of informative and non-informative components. Within this mixed missingness framework, we explore several assumptions to derive bounds on causal effects, including bounds expressed as a function of user-specified sensitivity parameters. We develop influence-function based estimators of these bounds to enable flexible, non-parametric, and machine learning based estimation, achieving root-n convergence rates and asymptotic normality under relatively mild conditions. We further consider the identification and estimation of bounds for other causal quantities that remain meaningful when informative missingness reflects a competing outcome, such as death. We conduct simulation studies and illustrate our methodology with a study on the causal effect of antipsychotic drugs on diabetes risk using a health insurance dataset.

Sensitivity Analysis

Sensitivity Analysis for the Attributable Fraction in Stratified Observational Studies Zhong Zheng* Zhong Zheng, Iris Horng, Dylan Small,

The attributable fraction measures the proportion of observed outcomes that can be causally attributed to an exposure. In stratified observational studies, such quantities are often estimated from collections of 2×2 tables that adjust for measured covariates through stratification. Because treatment assignment is not randomized, inference depends on the assumption of no unmeasured confounding. We develop a sensitivity analysis for attributable fractions in stratified observational studies. Existing sensitivity methods for stratified designs are typically based on Mantel–Haenszel or chi-squared statistics and are not well suited for attributable effects, which are nonlinear functionals of stratum-specific risks and exposure prevalences. Our approach evaluates the worst-case departure from the null within a sensitivity model that bounds treatment odds across individuals. The procedure can be formulated as an optimization problem involving the conditional mean and variance of a stratified test statistic. We derive computational methods to evaluate the objective efficiently under the conditional randomization distribution and provide algorithms that scale to many strata.

Sensitivity Analysis

Pseudo-RIR for Interpreting Low-Power Pre-Trend Tests in Difference-in-Difference Estimator Xuesen Cheng* Xuesen Cheng,

Difference-in-difference (DiD) estimators often check identifying assumptions by testing for “pre-trends.” A common example is an event-study specification, where researchers test whether lead coefficients are jointly zero. However, standard pre-trend tests can have low power, and conditioning on “passing” a pretest can lead to misleading reassurance and distorted inference. This paper proposes Pseudo-RIR to make a “passed” pre-trend test more interpretable. The idea builds on Robustness of an Inference to Replacement (RIR). When a pre-trend test does not reject, Pseudo-RIR quantifies how large a fraction of the pre-treatment treated observations would need to be increased by the threshold amount for the parallel trends test to become significant (fail to pass). This converts nonrejection into a concrete robustness statement: “How fragile is the absence of detected pre-trends?”
The study then shows—using broad simulation evidence across realistic DiD settings—how Pseudo-RIR maps into the power of common pre-trend tests under serial correlation, limited pre-periods, clustering, and different forms of trend violations. The paper offers a systematic sensitivity analysis plan: (i) report Pseudo-RIR whenever pre-trend tests “passed”; (ii) when Pseudo-RIR indicates meaningful risk, complement with partial-identification style sensitivity analysis under bounded trend deviations (like Honest DiD); and (iii) summarize robustness for the main estimation effect with an RIR-style index