Society for Causal Inference
EST. 2020

ABSTRACT BOOK

2025 American Causal Inference Conference

May 13-16, 2025

Machine Learning and Causal Inference

ML-assisted Randomization Tests for Detecting Treatment Effects in A/B Experiment
Wenxuan Guo* Wenxuan Guo, JungHo Lee, Panos Toulis,

Experimentation is widely utilized for causal inference and data-driven decision-making across disciplines. In an A/B experiment, for example, an online business randomizes two different treatments (e.g., website designs) to their customers and then aims to infer which treatment is better. In this paper, we construct randomization tests for complex treatment effects, including heterogeneity and interference. A key feature of our approach is the use of flexible machine learning (ML) models, where the test statistic is defined as the difference between the cross-validation errors from two ML models, one including the treatment variable and the other without it. This approach combines the predictive power of modern ML tools with the finite-sample validity of randomization procedures, enabling a robust and efficient way to detect complex treatment effects in experimental settings. We demonstrate this combined benefit both theoretically and empirically through applied examples.

Machine Learning and Causal Inference

A Unifying Weighting Perspective on Causal Machine Learning: Kernel Methods, Gaussian Processes, and Bayesian Tree Models Jared Murray* Jared Murray, Avi Feller,

Causal machine learning methods are powerful tools for estimating heterogeneous treatment effects but are often opaque and difficult to assess in practice. However, many methods –including outcome models with or without augmentation – produce estimates that are linear in the observed outcomes, making them weighting estimators. These model-implied weights can be interpreted as estimates of the Riesz representer for the target estimand, or, equivalently, as balancing weights.

In this paper we derive the implied weights of kernel ridge regression and many Bayesian nonparametric regression models (via their representation as conditional Gaussian process regressions); examples include BART, Bayesian causal forests, and Bayesian neural networks, as well as generic Gaussian process models. We then present a common weighting framework that connects these models to kernel methods like the R-learner and kernel mean matching. We use this framework to present new guidance for specifying semiparametric outcome models (or, more generally, choosing or learning kernels). In particular, we show that under mild conditions the Robinson regression-on-residuals parameterization of an outcome model produces implied weights that approximately balance broad classes of functions between treated and control groups for the average treatment effect in any target population. We therefore connect the desirable properties of Robinson transform and its corresponding Neyman orthogonal score to the balancin

Machine Learning and Causal Inference

On the Role of Surrogates in Conformal Inference of Individual Causal Effects Larry Han*
Larry Han, Chenyin Gao, Peter Gilbert,

Learning the Individual Treatment Effect (ITE) is essential for personalized decision-making, yet causal inference has traditionally focused on aggregated treatment effects. While integrating conformal prediction with causal inference can provide valid uncertainty quantification for ITEs, the resulting prediction intervals are often excessively wide, limiting their practical utility. To address this limitation, we introduce Surrogate-assisted Conformal Inference for Efficient INdividual Causal Effects (SCIENCE), a framework designed to construct more efficient prediction intervals for ITEs. SCIENCE accommodates covariate shifts between source data containing primary outcomes, and target data containing only surrogate outcomes or covariates. Leveraging semi-parametric efficiency theory, SCIENCE produces rate double-robust prediction intervals under mild rate convergence conditions, permitting the use of flexible non-parametric models to estimate nuisance functions. We quantify efficiency gains by comparing semi-parametric efficiency bounds with and without the incorporation of surrogates. Simulation studies demonstrate that our surrogate-assisted intervals offer substantial efficiency improvements over existing methods while maintaining valid group-conditional coverage. Applied to the phase 3 Moderna COVID-19 vaccine trial, SCIENCE illustrates how multiple surrogates can be leveraged to generate more efficient prediction intervals.

Machine Learning and Causal Inference

Causal Inference on Outcomes Learned from Text Amar Venugopal* Amar Venugopal, Iman Modarressi, Jann Spiess,

We propose a machine-learning tool that yields causal inference on text in randomized trials. Based on a simple econometric framework in which text may capture outcomes of interest, our procedure addresses three questions: First, is the text affected by the treatment? Second, which outcomes is the effect on? And third, how complete is our description of causal effects? To answer all three questions, our approach uses large language models that suggest systematic differences across documents that are reflective of the effect of the intervention and then provides valid inference based on costly validation. Specifically, we highlight the need for sample splitting to allow for statistical validation of LLM outputs, as well as the need for human labelling to validate substantive claims about which outcomes effects are on. We illustrate the tool in a proof-of-concept application using abstracts of academic manuscripts.

Policy Learning

Automatic Double Reinforcement Learning in Semiparametric Markov Decision Processes with Applications to Long-Term Causal Inference Lars van der Laan* Lars van der Laan, David Hubbard, Allen Tran, Nathan Kallus, Aurelien Bibaut,

Double reinforcement learning (DRL) (Kallus and Uehara, 2020) enables statistically efficient inference on the value of one policy in a nonparametric Markov Decision Process (MDP) given trajectories generated by another policy, but this necessarily requires stringent overlap between the state distributions, which is often violated in practice. To relax this and extend DRL, we study efficient inference on linear functionals of the Q-function (of which policy value is a special case) in infinite-horizon time-invariant MDPs under semiparametric restrictions on the Q-function. These restrictions can relax the overlap requirement and lower the efficiency bound, yielding more precise estimates. As an important example we study evaluation of long-term value under domain adaptation given a few short trajectories from the new domain and restrictions on the difference between the domains, which can be used for long-term causal inference. Our method combines flexible estimates of the Q-function and of the Riesz representer of the functional of interest (e.g., stationary state density ratio for policy value) and is automatic in that we do not need to know the form of the latter, only the functional we care about. To address potential model misspecification bias, we extend the adaptive debiased machine learning (ADML) framework of van der Laan et al. (2023) to construct nonparametrically valid and superefficient estimators that are adaptive to the functional form of

Matching, Weighting

A Deep Learning Approach to Nonparametric Propensity Score Estimation with Optimized Covariate Balance Liang Li* Liang Li, Maosen Peng, Chong Wu, Yan Li,

This paper addresses some key challenges in causal inference using a case study on the effect of erythrocyte-to-platelet ratio (EPR) changes on sepsis outcomes based on the MIMIC-IV electronic health records database. Observed inconsistencies across existing propensity score methods highlight issues such as model misspecification, poor overlap, and inadequate covariate balance. To overcome these limitations, we propose a novel propensity score weighting method based on two sufficient and necessary conditions: “local balance,” ensuring conditional independence of covariates and treatment assignment across a dense grid of balancing scores, and “local calibration,” guaranteeing that the balancing scores correspond to the true propensity scores. Using a neural network, we develop a nonparametric propensity score model that satisfies these conditions, effectively optimizing covariate balance, minimizing bias, and stabilizing inverse probability of treatment weights. Extensive numerical studies demonstrate that the proposed method successfully addresses these challenges, providing robust treatment effect estimation. In the case study, high EPR changes were associated with a significant 16% increased risk of 28-day mortality (HR: 1.16, 95% CI: 1.04–1.29). Our method offers a practical solution for causal inference in complex observational studies, with broad implications for improving clinical decision-making in critical care settings.

Matching, Weighting

Mixing Samples to Address Weak Overlap in Causal Inference Suehyun Kim* Jaehyuk Jang, Jaehyuk Jang, Kwonsang Lee,

In observational studies, the assumption of sufficient overlap (positivity) is fundamental for the identification and estimation of causal effects. Failing to account for this assumption yields inaccurate and potentially infeasible estimators. To address this issue, we introduce a simple yet novel approach, Mixing, which mitigates overlap violations by constructing a synthetic treated group that combines treated and control units. Our strategy offers three key advantages. First, it improves estimator accuracy by preserving unbiasedness while reducing variance. The benefit is particularly significant in settings with weak overlap, though the method remains effective regardless of the overlap level. This phenomenon results from the shrinkage of propensity scores in the mixed sample, which enhances robustness to poor overlap. Second, it enables direct estimation of the target estimand without discarding extreme observations or modifying the target population, thus facilitating straightforward interpretation of the results. Third, the mixing approach is highly adaptable to various weighting schemes, including contemporary methods such as Entropy Balancing. The estimation of the Mixed IPW (MIPW) estimator is done via M-estimation, and the method extends to a broader class of weighting estimators through a resampling algorithm. We illustrate the mixing approach through extensive simulation studies and provide practical guidance with a real-data analysis.

Matching, Weighting

Inference in Matching: A Novel Variance Estimation Approach for Overlapping Matched Sets Xiang Meng* Xiang Meng, Aaron Smith, Natesh Pillai, Luke Miratrix,

Matching estimators are widely used in causal inference to estimate treatment effects by pairing treated units with similar control units. While their asymptotic properties are well-studied, finite-sample inference remains challenging, particularly when control units are reused across multiple matches. Existing methods, including the wild bootstrap procedure, can produce unreliable inference when there is substantial overlap in matched samples—a common scenario when the number of treated units is of similar order to the number of control units.

We address this challenge by developing a novel variance estimator that remains valid even under substantial overlap in matched samples. Our approach leverages a more general theoretical framework based on a derivative control condition that improves upon traditional Lipschitz assumptions. This allows for broader applicability in settings where the derivative grows moderately but not uniformly. We prove the consistency of our variance estimator under weaker conditions than existing methods and establish its asymptotic normality.

Through extensive simulation studies, we demonstrate that our method significantly outperforms the state-of-the-art wild bootstrap approach in settings with extensive control unit reuse. While the wild bootstrap achieves only 61% coverage with artificially short confidence intervals in high-overlap scenarios, our method maintains near-nominal coverage rates.

Matching, Weighting

Estimating Average Treatment Effect via Marginal Outcome Density Ratio Linying Yang*
Linying Yang, Robin Evans,

Doubly robust estimators, such as AIPW, offer the advantage of providing `two chances' to perform estimation correctly and still obtain a consistent estimator. However, due to inverse probability weighting by the propensity score, these estimators can suffer from practical positivity violation, where some covariates predict the treatment so well that our weights become extremely large; this inflates the efficiency bound and estimation variance. This leads to the concept of the marginal density ratio. Instead of manually evaluating propensity scores or selecting features in the pre-treatment covariate space, we shift our focus to the outcome space, allowing the observed outcomes to determine which information should be included. In this paper, we introduce the Marginal outcome density Ratio estimator (MR) and the Augmented Marginal outcome density Ratio estimator (AMR); we demonstrate the advantages of these estimators in filtering necessary information to obtain treatment effects. We argue in this paper that, using this information filtering, MR and AMR are able to estimate the average treatment effect more effectively in small samples than their direct counterparts, IPW and AIPW. We also give examples on its contribution to sparsity condition of ATE estimation in the high-dimensional context.

Machine Learning and Causal Inference

Generalized Propensity Score Estimation of Multiple Binary Treatments via Localized Adversarial Lasso Yan Chen* Yan Chen, Alexandre Belloni, Matthew Harding,

We address the challenge of estimating generalized propensity scores for multiple (possibly dependent) binary treatments, a core problem in applications like multi-level treatments in causal inference and demand modeling for bundled products. The probability distribution of an M-dimensional binary treatment vector, without further assumptions, requires $O(2^M)$ parameters, leading to high-dimensional complexity. Understanding the (in)dependence structure among treatments enables a factorization of the probability distribution, substantially improving estimation.

To estimate this distribution, we utilize a Bahadur representation, linking sparsity in coefficients to independence across components. We propose regularized and adversarial regularized estimators to obtain an estimator adapted to the dependence structure, allowing for rates of convergence to depend on this intrinsic (lower) dimension. We address challenges like nuisance parameter estimation, covariate incorporation, and nonseparable moment conditions and provide pointwise rates of convergence for our locally penalized estimator which is computationally tractable.

Our framework is applied to estimating average treatment effects (ATE) under multiple binary treatments. Simulations validate our theoretical findings, showcasing improved finite-sample performance and superior coverage ratios for true ATEs, underscoring the practical and computational benefits of our approach.

Sensitivity Analysis

Reconciling Overt Bias and Hidden Bias in Sensitivity Analysis for Matched Observational Studies Siyu Heng* Siyu Heng, Yanxin Shen, Pengyun Wang,

Matching is one of the most commonly used causal inference study designs in observational studies. However, post-matching confounding bias typically exists, including overt bias due to inexact matching and hidden bias due to unmeasured confounding. Therefore, in matched observational studies, researchers routinely adopt the Rosenbaum sensitivity analysis framework to assess the impacts of post-matching confounding bias (either overt or hidden) on causal conclusions. In this work, we point out that this routine practice can be overly conservative because solving the Rosenbaum sensitivity model often renders the allocations of hypothetical hidden bias in sensitivity analysis contradict the overt bias observable from the matched dataset. To remove this contradiction, we propose an iterative convex programming approach to conduct a more powerful sensitivity analysis by adapting the solution space of hidden bias in sensitivity analysis to be compatible with the overt bias observed from the matched dataset. Our approach is asymptotically valid and uniformly more powerful than the conventional Rosenbaum sensitivity analysis framework and does not require any modeling assumptions on either the treatment or outcome variable. Our approach has been evaluated through extensive simulation studies and applied to an observational study on educational research.

Sensitivity Analysis

Controlling the False Discovery Proportion in Observational Studies with Hidden Bias
Mengqi Lin* Mengqi Lin, Colin Fogarty,

We propose an approach to exploratory data analysis in matched observational studies. We consider the setting where a single intervention is thought to potentially impact multiple outcomes, and the researcher would like to investigate which of these causal hypotheses come to bear while accounting not only for the possibility of false discoveries, but also the possibility that the study is plagued by unmeasured confounding. For any candidate set of hypotheses, our method provides sensitivity intervals for the false discovery proportion (FDP). For a set cR, the method describes how much unmeasured confounding would need to exist for us to believe that the proportion of true hypotheses is $0/|cR|$,$1/|cR|$,...,$|cR|/|cR|$. Moreover, the resulting confidence statements intervals are valid simultaneously over all possible choices for the rejected set, allowing the researcher to look in an ad hoc manner for promising subsets of outcomes that maintain a large estimated fraction of correct discoveries even if a large degree of unmeasured confounding is present. The approach is particularly well suited to sensitivity analysis, as conclusions that some fraction of outcomes were affected by the treatment exhibit larger robustness to unmeasured confounding than the conclusion that any particular outcome was affected. While the method involves solving quadratically constrained integer programs, we demonstrate that they can be efficiently handled or typically bypassed in large samples.

Sensitivity Analysis

Nonparametric Sensitivity Analysis for Unobserved Confounding with Survival Outcomes
Rui Hu* Rui Hu, Ted Westling,

In observational studies, the observed association between an exposure and outcome of interest may be distorted by unobserved confounding. Causal sensitivity analysis is often used to assess the robustness of observed associations to potential unobserved confounding. For time-to-event outcomes, existing sensitivity analysis methods rely on parametric assumptions on the structure of the unobserved confounders and Cox proportional hazards models for the outcome regression. If these assumptions fail to hold, it is unclear whether the conclusions of the sensitivity analysis remain valid. Additionally, causal interpretation of the hazard ratio is challenging. To address these limitations, in this paper we develop a nonparametric sensitivity analysis framework for time-to-event data. Specifically, we derive nonparametric bounds for the difference between the observed and counterfactual survival curves and propose estimators and inference for these bounds using semiparametric efficiency theory. We also provide nonparametric bounds and inference for the difference between the observed and counterfactual restricted mean survival times. We demonstrate the performance of our proposed methods using numerical studies and an analysis of the causal effect of physical activity on respiratory disease mortality among former smokers.

Sensitivity Analysis

Robustness of Proximal Inference TBD TBD* Cory McCartan, Melody Huang,

Proximal inference has been proposed as an alternative identification approach to relaxing traditional selection-on-observables (SOO) assumptions (i.e., no unobserved confounding) in observational causal inference. Instead of assuming researchers measure all relevant confounders, proximal inference assumes researchers have access to two informative proxies: a treatment proxy and an outcome proxy, which satisfy certain conditional independence assumptions. We formalize the trade-offs made between using a traditional SOO identification strategy in contrast to the proximal assumptions and derive the necessary scope conditions for proximal inference to provide more robust estimates than SOO. We consider the realistic setting in which both SOO and proximal assumptions are violated, finding that under even small violations of selection-on-observables, small violations in the exclusion restriction can amplify the resulting bias from proximal inference. We extend classical results from the instrumental variables literature to the proximal inference setting, and find that weak proxies can exacerbate both efficiency loss and potential bias. We compare the different approaches on a re-analysis of the impact of vote share shifts on legislative behavior.

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

Effect of Paid Sick Leave Policies on Child Health Lila Basnet* Lila Basnet, Kanika Arora, George Wehby,

In the absence of the national paid leave policies for workers, paid sick leave and paid family leave are labor policies implemented by fifteen states and are being considered by several other states. These policies allow employees to take time off when required to take care of themselves or the health of family members without incurring income loss or job loss. The importance of paid leave is high considering the corresponding needs of children as even healthy children have substantial health needs for regular care. The utilization of preventive child health care services requires parents' time to arrange visits to the service site. To examine the effect of paid sick leave policies on the utilization of health services, we identify the paid sick leave policies in each state and analyze data from the National Survey of Child Health from 2016 to 2023 for this study. Our outcome is the utilization of health services including dental visits, disease screening, and outpatient visits. Due to the heterogeneity of timing when policies were adopted, we used the Callaway-Sant'Anna staggered difference-in-differences approach. We estimate the effect of paid sick leave policies on the utilization of child health services and consider the validity of the causal assumptions involved in the model. This study will examine the effect of paid sick leave on the utilization of child health services and provide insights into the design and consideration of paid sick leave.

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

Mosaic inference on panel data Asher Spector* Asher Spector, Emmanuel Candes, Rina Foygel Barber,

The analysis of panel data via linear regression is ubiquitous in applied causal inference. However, such analyses typically assume that the residuals for different observations are cluster-independent; in general, standard confidence intervals may be invalid if this assumption is violated. This paper introduces a method called the mosaic permutation test that can be used to (a) test this assumption and (b) weaken it. We elaborate on these two contributions below.

Testing: Our method allows analysts to use nearly any machine learning technique to detect violations of the cluster-independence assumption while exactly controlling the false positive rate under a mild “local exchangeability” condition. To illustrate our method, we conduct a large-scale review of the literature and survey whether cluster-independence assumptions are accurate.

Inference: Our method produces confidence intervals for linear models that are (i) finite-sample valid under a local exchangeability assumption and (ii) asymptotically valid under the conventional cluster-independence assumption. In short, our method is valid under assumptions that are strictly weaker than classical methods. In experiments on real, randomly selected datasets from the literature, we find that many existing methods produce standard errors that are up to ten times too small, whereas mosaic methods produce reliable results.

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

Factorial Difference-in-Differences Yiqing Xu* Yiqing Xu, Anqi Zhao, Peng Ding,

In many panel data settings, researchers apply the difference-in-differences (DID) estimator, exploiting cross-sectional variation in a baseline factor and temporal variation in exposure to an event affecting all units. However, the exact estimand is often unspecified and the justification for this method remains unclear. This paper formalizes this empirical approach, which we term factorial DID (FDID), as a research design including its data structure, estimands, and identifying assumptions. We frame it as a factorial design with two factors—the baseline factor G and exposure level Z, and define effect modification and causal moderation as the associative and causal effects of G on the effect of Z, respectively. We show that under standard assumptions, including no anticipation and parallel trends, the DID estimator identifies effect modification but not causal moderation. To identify the latter, we propose an additional factorial parallel trends assumption. Moreover, we reconcile canonical DID as a special case of FDID with an additional exclusion restriction and link causal moderation to G's conditional effect with another exclusion restriction. We extend our framework to conditionally valid assumptions, clarify regression-based approaches, and illustrate our findings with an empirical example. We offer practical recommendations for FDID applications.

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

Smoking heat Gun: estimating the effect of heatwaves on gun violence through a distance augmented synthetic control method Giulio Grossi* Giulio Grossi, Falco Johannes Bargagli Stoffi, Leo Vanciu,

Gun violence in the United States has become an increasingly pressing issue in recent years. The upward trend is alarming, and policymakers are focusing their efforts on reducing the death toll. In this study, we examine gun violence in a changing world. Our primary focus is the relationship between heatwaves—direct consequences of climate change in intensity and frequency—and gun violence. We investigate these relationships within a potential outcomes framework, defining causal effects for areas that have experienced a heatwave and those potentially subject to spillover effects. We estimate the causal effect by introducing a spatially augmented version of the synthetic control method, leveraging the spatial information in the data to improve the interpretability of our estimates and reduce their variability. We employ a Bayesian regression approach to penalize the selection of more distant control units, within a semiparametric framework that balances unobserved spatial confounding.
Our findings contribute substantively by clarifying how environmental factors relate to gun violence, and methodologically by naturally extending the synthetic control method to spatial data.

Instrumental Variables

Regularized DeepIV with Model Selection Hui Lan* Hui Lan, Zihao Li, Vasilis Syrgkanis, Mengdi Wang, Masatoshi Uehara,

In this paper, we study nonparametric estimation of instrumental variable (IV) regressions. While recent advancements in machine learning have introduced flexible methods for IV estimation, they often encounter one or more of the following limitations: (1) restricting the IV regression to be uniquely identified; (2) requiring minimax computation oracle, which is highly unstable in practice; (3) absence of model selection procedure. In this paper, we analyze a Tikhonov-regularized variant of the seminal DeepIV method, called Regularized DeepIV (RDIV) regression, that can converge to the least-norm IV solution, and overcome all three limitations. RDIV consists of two stages: first, we learn the conditional distribution of covariates, and by utilizing the learned distribution, we learn the estimator by minimizing a Tikhonov-regularized loss function. We further show that RDIV allows model selection procedures that can achieve the oracle rates in the misspecified regime. When extended to an iterative estimator, we prove that RDIV matches the current state-of-the-art convergence rate. Furthermore, we conducted numerical experiments to justify the efficiency of RDIV empirically. Our results provide the first rigorous guarantees for the empirically well-established DeepIV method, showcasing the importance of regularization which was absent from the original work.

Instrumental Variables

Two-Stage Machine Learning for Instrumental Variable Regression David Bruns-Smith* David Bruns-Smith,

Instrumental variable (IV) regression with two-stage least squares (2SLS) is a widely-used estimator for observational causal inference. Unfortunately, modern non-linear machine learning (ML) models cannot be directly integrated into 2SLS, a fact sometimes called the “forbidden regression” problem. Existing extensions like nonparametric IV (NPIV) and kernel IV (KIV) have limited performance in practice, and newer approaches with trees or neural networks require iterative or adversarial training, introducing thorny convergence issues.

We propose two-stage machine learning (2SML), a simple two-stage IV procedure compatible with arbitrary ML models, and implementable in a few lines of code. Our key insight is that the second stage of 2SLS can be reformulated as a least squares problem where the predictions are projected onto a certain feature expansion, obtained via a first-stage regression of the outcome on the instruments. For kernel methods, 2SML is numerically equivalent to NPIV/KIV, but with a markedly different interpretation of the first stage. However, our framework also seamlessly generalizes to support tree ensembles and neural networks.

We establish L2-convergence, and evaluate 2SML on a novel, challenging IV benchmark, demonstrating significant performance gains over existing approaches. Additionally, we develop a complementary debiasing procedure that provides valid confidence intervals for linear functional estimands.

Instrumental Variables

Local Effects of Continuous Instruments without Positivity Luke Keele* Luke Keele, Prabrisha Rakshit, Alex Levis,

Instrumental variables are a popular study design for the estimation of treatment effects in the presence of unobserved confounders. In the canonical instrumental variables design, the instrument is a binary variable. In many settings, however, the instrument is continuous. Standard estimation methods can be applied with continuous instruments, but they require strong assumptions. While recent work has introduced more flexible estimation approaches, these methods require a positivity assumption that is implausible in many applications. We derive a novel family of causal estimands using stochastic dynamic interventions that allows a range of intervention distributions that are continuous with respect to the observed distribution of the instrument. These estimands focus on a specific local effect but do not require a positivity assumption. Next, we develop doubly robust estimators for these estimands that allow for estimation of the nuisance functions via nonparametric estimators. We use empirical process theory and sample splitting to derive asymptotic properties of the proposed estimators under weak conditions. In addition, we derive methods for profiling the principal strata as well as a method of sensitivity analysis. We evaluate our methods via simulation and demonstrate their feasibility using an application on the effectiveness of surgery for specific emergency conditions.

Instrumental Variables

An Instrumental Variable Approach under a Multiplicative Selection Model for Data Missing Not-at-Random Yunshu Zhang* Yunshu Zhang, Eric Tchetgen Tchetgen,

Instrumental variable (IV) methods offer a valuable approach to account for outcome data missing not-at-random. A valid missing data instrument is a measured factor which (i) predicts the nonresponse process and (ii) is independent of the outcome in the underlying population. For point identification, all existing IV methods for missing data including the celebrated Heckman selection model, a priori restrict the extent of selection bias on the outcome scale, therefore potentially understating uncertainty due to missing data. In this work, we introduce an IV framework which allows the degree of selection bias on the outcome scale to remain completely unrestricted. The new approach instead relies for identification on (iii) a key multiplicative selection model, which posits that the instrument and any hidden correlate of both selection and the outcome, impact the selection mechanism independently on the multiplicative scale. Interestingly, we establish that any regular statistical functional of the missing outcome is nonparametrically identified under (i)-(iii) via a modified Wald ratio estimand reminiscent of the standard Wald ratio estimand in causal inference. For estimation and inference, we characterize the efficient influence function for any functional defined on a nonparametric full data model, which we leverage to develop semiparametric efficient multiply robust IV estimators. Several extensions of the methods are also considered, including the important practical

Instrumental Variables

Identification and Estimation with Deconfounded Instruments under Index Sufficiency
Christian Tien* Christian Tien,

This paper extends a novel methodology for identifying and estimating the causal effects of endogenous (treatment) variables on an outcome variable with partially endogenous instrumental variables. A crucial estimation step called “deconfounding” recovers variation in the instruments, which is unassociated with some observed variables called proxies, and consequently with any unobserved variables that explain the association between the instruments and proxies. These unobserved variables are called common confounders of the instruments and proxies. This paper explores the role of index sufficiency as a minimal parametric assumption, which naturally fits in with the deconfounding approach and permits the identification and estimation of causal effects in the common confounding setup. Novel semiparametric estimation theory is provided. Root-n-estimation of causal effects under index sufficiency in otherwise nonparametric models is shown to be possible, combining ideas from debiasing with respect to sequentially dependent nuisance functions [Singh, 2021, Chernozhukov et al., 2022] with recent results on strong identification subject to nuisance functions, which are defined as solutions to possibly ill-posed inverse problems [Bennett et al., 2022]. An empirical application on the returns to education with NLS97 data demonstrates the appeal of this approach in practical settings with partially endogenous instruments.

Design of Experiments

Your LLM Is Too Big for Causal Inference with Text Srikar Katta* Srikar Katta, Graham Tierney, Chris Bail, Sunshine Hillygus, Alexander Volfovsky,

Many modern social science questions ask how linguistic properties causally affect human behavior. Because text properties are often interlinked (e.g., angry reviews often use profane language), analysts must control for latent confounding to isolate causal effects. Recent literature proposes adaptations of large language models, transformers, and other deep learning techniques to learn latent representations in text that successfully predict both treatment status and the outcome. However, because the treatment is encoded in the text, these deep learning methods run the risk of learning representations that actually encode the treatment itself, inducing an overlap bias. Rather than depending on post-hoc adjustments of the text, we introduce a new experimental design that allows scientists to handle latent confounding, avoid the overlap issue, and unbiasedly estimate treatment effects. We run an experiment that showcases this experimental design to study how expressing humility in political statements affects readers' beliefs. We leverage the ground truth information in our collected experimental data to demonstrate the failures in current language model approaches for causal inference in text. We then return to our implemented study and discover novel relationships between expressing humility and the perceived persuasiveness of political statements, offering important insights for social media platforms and social scientists.

Design of Experiments

Evaluate Power and Sample Size in Efficient Randomized Control Trial Design Lauren Liao*
Lauren Liao, Emilie Højbjerre-Frandsen, Alejandro Schuler,

To design a prospective study, power and sample size calculations are necessary to ensure a successful randomized trial targeting an effect of interest. Traditionally, calculations from unadjusted estimators are most commonly used in application. While semi-parametrically efficient estimators have been proved to reduce the variance while yielding more robust results, prospective studies leveraging efficient estimators are limited. To encourage a more efficient trial design, Schuler (2021) proposed a new formula for power calculation to leverage historical data to estimate sample size needed for semi-parametrically efficient estimators. However, this formula can overestimate the power using targeted minimum loss based estimation (TMLE) as suggested for the difference-in-means estimator. Instead, we recommend leveraging prognostic covariate adjustment as suggested in Liao et al. (2023) in combination with TMLE to minimize the likelihood of having an underpowered study. We showcase the practical usage of the power calculation developed by Schuler (2021) in a simulation to consider different machine learning algorithms and data generating distributions. We also illustrate the power calculation and validation with a case study using data from Novo Nordisk A/S.

Design of Experiments

Dynamic Assortment Optimization via Optimal Experimental Design Aurelien Bibaut*
Aurelien Bibaut, Guy Aridor, Nathan Kallus,

Online streaming platforms hope to offer catalogs of items (music, video, etc.) that are valuable to users. We consider evaluating catalogs in terms of user choice among the included items and an outside option and study the problem of designing batched adaptive experiments that would be maximally informative about a parameter of interest in terms of this unknown choice model, such as the value of a counterfactual catalog. Our discrete choice model is parametrized by a high-dimensional matrix of utility parameters with an underlying low rank structure, capturing heterogeneity in preferences with manageable statistical complexity without a priori clusterings. We derive the semiparametric efficiency bound for parameter of interest given data from different experimental designs, yielding an interpretable criterion for selecting a design. A key technical challenge arises because the optimal experimental design depends on unknown nuisance parameters, which must be learned adaptively. We show that our approach achieves, up to a negligible term, the same regret as an oracle that knows these nuisance parameters in advance. We demonstrate the value of our approach in a real-world application to a large streaming company.

Design of Experiments

Sequential Rerandomization Under Ellipsoidal and Rectangular Constraints Kyle Schindl*
Kyle Schindl, Zach Branson,

When designing a randomized experiment, one way to ensure treatment and control groups exhibit similar covariate distributions is to randomize treatment until some prespecified level of covariate balance is satisfied; this strategy is known as rerandomization. Most rerandomization methods utilize balance metrics based on ellipsoidal or rectangular constraints on the covariate mean-differences. In this work, we derive general results for treatment-versus-control rerandomization schemes that employ either type of constraint for covariate balance. In addition to allowing researchers to quickly derive properties of rerandomization schemes not previously considered, our theoretical results provide guidance on how to choose the constraint in practice. After deriving optimal choices of constraints for covariate balance and variance reduction, we extend this framework to sequential experiments. There, we develop an optimal updating strategy such that experimenters can change their rerandomization strategy after observing the covariates (or outcomes) from previous groups. This leads to substantial efficiency gains over traditional sequential experimental design strategies.

Design of Experiments

Maximizing Gains from Existing Information: An Adaptive Pairing and Stratification Procedure for Experimental Designs Zikai Li* Zikai Li,

In social science experiments, given a fixed budget and/or a limited number of participants, researchers need to optimize statistical efficiency. To this end, they often use stratification. However, conventional practices of stratification are agnostic with respect to the predictive relationship between the covariates and the outcomes, even though it is this predictive relationship that motivates stratification. Such practices thus fail to take advantage of all available information when some data is available for the experimental outcome and the covariates. This paper introduces an adaptive pairing and stratification procedure for running an experiment in batches. This approach builds upon recent work that demonstrates the theoretical optimality of pairing units based on the expected sum of potential outcomes. The method incorporates information about the relationship between covariates and potential outcomes when pairing or stratifying units. It uses data from earlier batches not just to inform pairing decisions but also to rematch observations across different batches without compromising the validity of inference under the superpopulation framework. In experimental settings where sequential treatment assignment and outcome collection are feasible, this approach can improve the efficiency of treatment assignments relative to its alternatives. My simulations demonstrate such gains can be substantial.

Applications in Health and Biology

Unveiling Heterogeneous Treatment Effects in the BEST-CLI Trial: Harnessing Causal Machine Learning to Personalize Care for Chronic Limb-Threatening Ischemia Tien Tran*
Tien Tran, Alik Farber, Matthew Menard, Kenneth Rosenfield, Niteesh Choudhry, Zafar Zafari,

Randomized controlled trials (RCTs) are foundational for determining the efficacy of treatments, yet their insights often overlook heterogeneous treatment effects (HTEs) across subpopulations. Machine learning (ML) methods, such as causal forests, have emerged as powerful tools to identify these HTEs, enhancing healthcare decision-making by personalizing treatment and optimizing resource allocation. This study applies causal forests to the BEST-CLI trial, a multicenter RCT comparing open surgical bypass (OSB) with endovascular therapy (EVT) for patients with chronic limb-threatening ischemia.

Leveraging individual-level data from the trial, we estimated subgroup-specific treatment effects to explore variability in clinical outcomes. Our analysis revealed significant heterogeneity in the benefits of OSB versus EVT based on key patient characteristics, including age, comorbidities, and anatomical considerations. These findings underscore the potential of causal machine learning to refine clinical guidelines by tailoring treatment recommendations to individual patient profiles. By incorporating ML-derived insights into RCT analyses, this study demonstrates a methodological advance in translating trial data into actionable, patient-centered care strategies.

Applications in Health and Biology

Overcoming an extreme positivity violation to distinguish the causal effects of surgery and anesthesia via a separable effects model Caleb Miles* Caleb Miles, Amy Pitts, Caleb Ing, Ling Guo,

The U.S. Food and Drug Administration has cautioned that prenatal exposure to anesthetic drugs during the third trimester may have neurotoxic effects; however, there is limited clinical evidence available to substantiate this recommendation. One major scientific question of interest is whether such neurotoxic effects might be due to surgery, anesthesia, or both. Isolating the effects of these two exposures is challenging because they are observationally equivalent, thereby inducing an extreme positivity violation. To overcome this, we adopt the separable effects framework of Robins and Richardson (2010) with a novel perspective, shifting the focus from starting with a known intermediate variable of interest to starting with known separable components of treatment that are of interest. In particular, under this framework, surgery and anesthesia are the known separable components of the exposure, and we identify the effect of anesthesia (alone) by blocking effects through intermediate variables that are assumed to completely mediate the causal pathway from surgery to the outcome. We apply this approach to data from the nationwide Medicaid Analytic eXtract (MAX) from 1999 through 2013, which linked 16,778,281 deliveries to mothers enrolled in Medicaid during pregnancy. Furthermore, we assess the sensitivity of our results to violations of our key identification assumptions.

Applications in Health and Biology

Causal Inference with Protein Sequences: Biologically Plausible Dimension Reduction Using AlphaFold Matthew Laws* Matthew Laws, Rohit Bhattacharya,

Identifying disease-causing mutations is crucial for designing targeted treatments. However, randomized experiments for this purpose are often prohibitively slow, expensive, and typically conducted in model organisms, limiting scalability and generalizability. This work explores how observational causal inference can address this challenge. The primary obstacle is the high dimensionality of the treatment variable: the mutation space of a protein sequence. While high-dimensional confounding is well-studied in causal inference, handling high-dimensional treatments remains relatively under explored.

We introduce a novel, tailored approach that utilizes protein folding as a biologically plausible method for dimensionality reduction. Using AlphaFold3 (Abramson et al., 2024), we fold mutated sequences into their 3D protein structures and align them with healthy counterparts to generate a continuous misalignment score. Then, applying continuous treatment methods, we estimate the causal effects of misaligned protein structures on disease development. Assuming mutations disrupt protein alignment, our method provides bounds on the effects of genetic mutations on disease progression.

We evaluated our methodology using both real and semi-synthetic datasets, with a particular focus on BRCA1 mutations—a gene strongly associated with breast cancer development. We also compare with more generic approaches of dealing with high dimensional treatments and show that ours is more desirable.

Applications in Health and Biology

Positive and negative control outcomes to inform target trial emulations with observational data: an application to diabetes medications in the BESTMED Consortium Lucia Petito*
Lucia Petito, Emma Hegermiller, Indhumathy Chelliah, Golnaz Loftian, Ryan Carnahan, Satyender Goel, Alan Kaul, Andrea DeVries, Cecilia Lansang, Marie McDonnell, Vinit Nair, Elisa Priest, Vincent Willey, Alexander Turchin, Miguel Hernan,

Although emulating a target trial precludes many design-induced biases, confounding due to noncomparability of treatment groups on unmeasured factors remains a concern. Employing control outcomes—outcomes for which the magnitude of the treatment effect is known and the confounding structure is similar to the one for the main outcome of interest—may help identify whether residual confounding exists after adjustment for measured factors. Here, we describe the use of control outcomes for confounding assessment in a study of the effect of second-line treatments for type 2 diabetes (dipeptidyl peptidase 4 inhibitors; glucagon-like peptide-1 receptor agonists; sodium-glucose cotransporter-2 inhibitors; sulfonylureas [SU]; basal insulin) on the 3-year risk of cardiovascular events using data on 57,910 individuals from the BESTMED Consortium (http://www.bestmed.org). We study 2 control outcomes: 3-month risk of cardiovascular events (negative control outcome) and 12-month change in hemoglobin A1c (positive control outcome). An outcome that does not share the same confounding structure, 3-year risk of herpes zoster, is considered as it has been used in previous falsification analyses. We conclude that selection of control outcomes with similar confounding structures is critically important to identify the presence of substantial confounding after adjustment for the information available in the observational database.

Heterogeneous Treatment Effects

Combining Observational and Experimental Data to Learn Interpretable Subgroups with Heterogeneous Treatment Effects Rahul Ladhania* Rahul Ladhania, Amelia Haviland,

In this paper, we propose and evaluate a two study approach to combine large “noisy” observational data with “clean” smaller experimental data to learn interpretable population subgroups with heterogeneous treatment effects. In the first study, we use an observational dataset, potentially with unobserved confounding, to identify sub-groups exhibiting the most distinctive treatment-outcome relationships. Our method employs a non-parametric approach, validated through a three-stage sample-splitting process to minimize overfitting and to ensure robustness. While Study 1 reduces noise, it remains susceptible to bias from unobserved confounding. Study 2 leverages an experimental design, applying the sub-group definitions learned in Study 1 to estimate treatment effects within each group, thereby testing the causal hypotheses generated in Study 1. We demonstrate the strengths and limitations of our approach through a simulation setting which varies the degree and direction of unobserved confounding. Additionally, we apply our method to data from the Women’s Health Initiative, a landmark 1991 study investigating the health effects of hormone replacement therapy on postmenopausal women.

Heterogeneous Treatment Effects

Distilling Causal Effects: Stable subgroup estimation via distillation trees in causal inference Ana Kenney* Melody Huang, Tiffany Tang, Ana Kenney,

Recent methodological developments have introduced new black-box approaches to better estimate heterogeneous treatment effects; however, these methods fall short of providing interpretable characterizations of the underlying individuals who may be most at risk or benefit most from receiving the treatment, thereby limiting their practical utility. In this work, we introduce a novel method, causal distillation trees (CDT), to estimate interpretable subgroups. CDT allows researchers to fit any machine learning model of their choice to estimate the individual-level treatment effect, and then leverages a simple, second-stage tree-based model to “distill” the estimated treatment effect into meaningful subgroups. As a result, CDT inherits the improvements in predictive performance from black-box machine learning models while preserving the interpretability of a simple decision tree. We theoretically characterize the stability of CDT in estimating substantively meaningful subgroups and provide stability-driven diagnostics for researchers to evaluate the quality of the estimated subgroups. We illustrate our proposed method on a randomized controlled trial of antiretroviral treatment for HIV from the AIDS Clinical Trials Group Study 175 and show that CDT out-performs state-of-the-art approaches in constructing stable, clinically relevant subgroups.

Generalizability/Transportability

Minimax Regret Estimation for Generalizing Heterogeneous Treatment Effects with Multisite Data Yi Zhang* Yi Zhang, Melody Huang, Kosuke Imai,

To test scientific theories and develop individualized treatment rules, researchers often wish to learn heterogeneous treatment effects that can be consistently found across diverse populations and contexts. We consider the problem of generalizing heterogeneous treatment effects (HTE) based on data from multiple sites. A key challenge is that a target population may differ from the source sites in unknown and unobservable ways. This means that the estimates from site-specific models lack external validity, and a simple pooled analysis risks bias. We develop a robust CATE estimation methodology with multisite data from heterogeneous populations. We propose a minimax-regret framework that learns a generalizable CATE model by minimizing the worst-case regret over a class of target populations whose CATE can be represented as convex combinations of site-specific CATEs. Using robust optimization, the proposed methodology accounts for distribution shifts in both individual covariates and treatment effect heterogeneity across sites. We show that the resulting CATE model has an interpretable closed-form solution, expressed as a weighted average of site-specific CATE models. Thus, researchers can utilize a flexible CATE estimation method within each site and aggregate site-specific estimates to produce the final model. Through simulations and a real-world application, we show that the proposed methodology improves the robustness and generalizability of existing approaches.

Dynamic Treatment Regimes

Efficient estimation of optimal treatment rules with fused randomized trials and missing covariates Nicholas Williams* Nicholas Williams, Kara Rudolph, Iván Díaz,

A fundamental principle of clinical medicine is that a treatment should only be administered to those patients who would benefit from it. Treatment strategies that assign treatment to patients as a function of their individual characteristics are known as dynamic treatment rules. Randomized clinical trials are considered the gold standard for estimating the marginal causal effect of a treatment on an outcome; they are often not powered to detect heterogenous treatment effects. The availability of multiple trials presents an opportunity for combining data from multiple randomized trials, often called data-fusion, to better estimate dynamic treatment rules using the combined data-set than either data-set alone. However, there may be a mismatch in the set of patient covariates measured between trials. We address this problem here; namely, we propose a nonparametric estimator for the optimal dynamic treatment rule that leverages information from the set of randomized trials with missing covariates to estimate the conditional average treatment effect. We show, under certain conditions, that the proposed estimator is more efficient than an estimator for the conditional average treatment effect that only uses the set of trials that measure all covariates. We apply the estimator to fused randomized trials of medications for the treatment of opioid use disorder to estimate a treatment rule that would match patient subgroups with the medication that would minimize risk of relapse.

Heterogeneous Treatment Effects

Trustworthy assessment of heterogeneous treatment effect estimator via analysis of relative error Zijun Gao* Zijun Gao,

Accurate heterogeneous treatment effect (HTE) estimation is essential for personalized recommendations, making it important to evaluate and compare HTE estimators. Traditional assessment methods are inapplicable due to missing counterfactuals. Current HTE evaluation methods rely on additional estimation or matching on test data, often ignoring the uncertainty introduced and potentially leading to incorrect conclusions. We propose incorporating uncertainty quantification into HTE estimator comparisons. In addition, we suggest shifting the focus to the estimation and inference of the relative error between methods rather than their absolute errors. Methodology-wise, we develop a relative error estimator based on the efficient influence function and establish its asymptotic distribution for inference. Compared to absolute error-based methods, the relative error estimator (1) is less sensitive to the error of nuisance function estimators, satisfying a “global double robustness” property, and (2) its confidence intervals are often narrower, making it more powerful for determining the more accurate HTE estimator. Through extensive empirical study of the ACIC challenge benchmark datasets, we show that the relative error-based method more effectively identifies the better HTE estimator with statistical confidence, even with a moderately large test dataset or inaccurate nuisance estimators.

Design-Based Causal Inference

Inference for Group Interaction Experiments Cyrus Samii* Ye Wang, Cyrus Samii, Jiawei Fu,

A common experimental research design is one in which individuals are put into groups and then interact within the groups under different group-level treatment conditions. We present methods for design-based inference for such “group interaction” experiments. A key consideration is that group interaction implies potential interference, which yields dependencies that should be accounted for when making inferential claims. We show that when interference is present, standard cluster robust inference is super-population consistent in accounting for such dependencies for inference on marginalized causal effects that account for interference. When interference is not present but groups are formed through individual random assignment, individual-level heteroskedasticity robust inference is consistent for inference on the usual average treatment effect. We prove the consistency and asymptotic normality of the difference-in-means estimator for group interaction experiments and extend our framework to cases with restricted group compositions and covariate-conditional effects. Finally, we validate our theoretical propositions through simulation exercises and a replication study of such experiments in social science.

Causal Inference in Networks

Agnostic Characterization of Interference in Randomized Experiments David Choi* David Choi,

We give an approach for characterizing interference by lower bounding the number of units whose outcome depends on selected groups of treated individuals, such as depending on the treatment of others, or others who are at least a certain distance away. The approach is applicable to randomized experiments with binary-valued outcomes. Asymptotically conservative point estimates and one-sided confidence intervals may be constructed with no assumptions beyond the known randomization design, allowing the approach to be used when interference is poorly understood, or when an observed network might only be a crude proxy for the underlying social mechanisms. Point estimates are equal to Hajek-weighted comparisons of units with differing levels of treatment exposure. Empirically, we find that the width of our interval estimates is competitive with (and often smaller than) those of the EATE, an assumption-lean treatment effect, suggesting that the proposed estimands may be intrinsically easier to estimate than treatment effects.

https://arxiv.org/abs/2410.13142

Design-Based Causal Inference

Causal Interpretation of Regressions With Ranks Lihua Lei* Lihua Lei,

In studies of educational production functions or intergenerational mobility, it is common to transform the key variables into percentile ranks. Yet, it remains unclear what the regression coefficient estimates with ranks of the outcome or the treatment. In this paper, we derive effective causal estimands for a broad class of commonly-used regression methods, including the ordinary least squares (OLS), two-stage least squares (2SLS), difference-in-differences (DiD), and regression discontinuity designs (RDD). Specifically, we introduce a novel primitive causal estimand, the Rank Average Treatment Effect (rank-ATE), and prove that it serves as the building block of the effective estimands of all the aforementioned econometrics methods. For 2SLS, DiD, and RDD, we show that direct applications to outcome ranks identify parameters that are difficult to interpret. To address this issue, we develop alternative methods to identify more interpretable causal parameters.

Design of Experiments

The Conflict Graph Design: Estimating Causal Effects under Arbitrary Neighborhood Interference Christopher Harshaw* Christopher Harshaw, Vardis Kandiros, Charilaos Pipis, Constantinos Daskalakis,

A fundamental problem in network experiments is selecting an appropriate experimental design in order to precisely estimate a given causal effect of interest. In fact, optimal rates of estimation remain unknown for essentially all causal effects in network experiments. In this work, we propose a general approach for constructing experiment designs under network interference with the goal of precisely estimating a pre-specified causal effect. A central aspect of our approach is the notion of a conflict graph, which captures the fundamental unobservability associated with the casual effect and the underlying network. We refer to our experimental design as the Conflict Graph Design. In order to estimate effects, we propose a modified Horvitz–Thompson estimator. We show that its variance under the Conflict Graph Design is bounded as $O(\lambda(H)/n)$, where $\lambda(H)$ is the largest eigenvalue of the adjacency matrix of the conflict graph. These rates depend on both the underlying network and the particular causal effect under investigation. Not only does this yield the best known rates of estimation for several well-studied causal effects (e.g. the global and direct effects) but it also provides new methods for effects which have received less attention from the perspective of experiment design (e.g. spill-over effects). In addition to point estimation, we construct asymptotically valid confidence intervals for the casual effect of interest.

Causal Inference and Common Support Violations

Prognostic scores and representation learning for causal effect estimation with weak overlap David Bruns-Smith* Oscar Clivio, Alexander D'Amour, Alexander Franks, Oscar Clivio, Avi Feller, Chris Holmes,

Overlap, also known as positivity, is a key condition for modern causal machine learning. Many popular estimators suffer from high variance and become brittle when features strongly differ across treatment groups. This is especially challenging in high dimensions: the curse of dimensionality can make overlap implausible. Modern causalML methods typically address this issue only indirectly, leveraging dimension reduction or other representation learning that does not account for overlap. In the limit, such methods reduce features to a scalar, such as the prognostic score (i.e., the conditional counterfactual mean). Building on a venerable empirical literature, we argue that the prognostic score is an unreasonably effective dimension reduction approach, and is a promising default in otherwise complex settings. To show this, we first propose a class of feature representations called deconfounding scores, which preserve both identification and the target of estimation while also improving overlap; the propensity and prognostic scores are two special cases. We characterize the corresponding optimization problem in terms of controlling overlap under an unconfoundedness constraint. We then derive closed-form expressions for overlap-optimal representations under a broad family of generalized linear models with Gaussian covariates and show that this coincides with the prognostic score. We conduct extensive experiments to assess this behavior empirically.

Applicants in Social Sciences

Examining Racial Disparities in Healthcare Expenditures via Causal Path-Specific Effects
Xiaxian Ou* Xiaxian Ou, Xinwei He, Razieh Nabi,

Racial disparities in healthcare expenditures have been widely documented, yet the underlying drivers remain complex and require further exploration. This study employs causal path-specific effects to assess how different factors contribute to the observed differences. Leveraging data from the Medical Expenditures Panel Survey, we examine the roles of socioeconomic status, insurance access, health behaviors, and health status. Our pathway-specific framework provides detailed assessments of how expenditures would change if specific mediating factors were counterfactually aligned across racial groups, offering a structured approach to quantifying sources of disparities. We also address several challenges in measuring the pathway-specific disparities. The relationships between race, healthcare spending, and mediating factors are complex, necessitating robustness against model misspecification. Furthermore, healthcare expenditures are characterized by zero-inflation and right skewness, requiring specialized modeling approaches. In addition, for reliable inference, it is essential to quantify uncertainty, ensuring that our estimators exhibit desirable statistical properties such as asymptotic normality and $sqrt{n}$-consistency. To address these complexities, we analyze the MEPS data using robust influence function-based estimators and integrate flexible statistical and machine learning methods, including super learners and a two-part model for zero-inflated skewed expenditures.

Applicants in Social Sciences

Assessing the Causal Impact of Early Tracking Postponement on Inequality of Opportunity: Evidence from the Italian Single Middle School Reform Kevin Taglialatela Scafati* Kevin Taglialatela Scafati, Paolo Brunori, Fabrizia Mealli,

Over the past two decades a growing body of empirical research has sought to evaluate the causal impact of school tracking on inequality. This paper contributes to this literature by exploiting the case of the 1963 Italian Single Middle School reform to investigate the causal effect of an early tracking postponement on inequality of opportunity (IOp). Our primary outcome of interest is a long-term well-being measure, namely an estimate of the individual permanent income. By exploiting the reform's innovations and their implementation timeline, we identify two groups of students that are exposed to educational systems that differ solely due to the presence of early tracking. We assume these groups overlap in terms of key observed ascribed characteristics – such as birthplace, parental education and occupation, sex – which serve as potential sources of inequality and confounding. Building on established social science literature, we employ inequality indices (e.g., Gini, MLD) applied to predicted individual well-being based on ascribed factors as measures of IOp. Our primary estimands contrast these measures across the two selected exposure groups, providing insights into the causal effect of interest. The nature of these novel estimands presents challenges for estimation; we adopt a Bayesian approach to inference which allows to naturally quantify the uncertainty of the targeted quantities and to flexibly specify the outcome model through a non-parametric BART specification.

Applications in Health and Biology

A new design for observational studies applied to the study of the effects of high school football on cognition late in life Katherine Brumberg* Katherine Brumberg, Dylan Small, Paul Rosenbaum,

Do the impacts that occur when playing high school football have concussive effects that accelerate cognitive decline late in life? We examine this possibility using newly available cognitive data describing people in 2020 who graduated high school in 1957. Someone who was 18 in 1957 would be 81 in 2020. For this comparison we develop a new design for an observational study, called a triples design, and discuss its advantages and construction. A triples design consists of M blocks of size 3, where a block contains either one treated individual and two controls or two treated individuals and one control. A triples design is the simplest design that uses weights, with just two weights. Like full matching, a triples design can match more people than can matched pairs, yet have smaller within-block covariate distances. Unlike full matching, there are no matched pairs. Like matching with multiple controls, a triples design will have a larger design sensitivity than a design which includes matched pairs, under simple models for continuous outcomes; that is, in favorable situations the design is expected to report greater insensitivity to unmeasured biases. Because there are just two weights, it is easy to construct weighted graphics for exploratory displays from triples designs. A heuristic algorithm containing network optimization constructs the design.

Applications in Health and Biology

The effect of long-term adherence to physical activity recommendations in midlife on plasma proteins associated with frailty in the Atherosclerosis Risk in Communities (ARIC) study Fangyu Liu* Fangyu Liu, Jennifer A. Schrack, Keenan A. Walker, Jeremy Walston, Rasika A. Mathias, Michael E. Griswold, Priya Palta, B. Gwen Windham, John W. Jackson,

Clinical trials have shown favorable effects of exercise on frailty, supporting physical activity (PA) as a treatment and prevention strategy. However, less is known about the biological mechanisms underlying PA's benefits. To date, proteomics studies that examined the effects of PA on proteins, some of which may function as molecules in the biological processes underlying frailty, have focused on structured exercise programs over a short term. To better understand the benefits of long-term, less structured PA in free living, we emulated a target trial that assigned 14,898 middle-aged adults to either (i) achieve and maintain the recommended PA level (≥150 minutes/week of moderate-to-vigorous physical activity [MVPA]) through 6 (+/-0.3) years of follow-up or (ii) follow a "natural course" strategy, where all individuals engage in their habitual MVPA. We estimated the population-level difference between (i) and (ii) on 45 previously identified frailty-associated proteins (standardized) at the end of the follow-up using inverse probability of weighting (IPW) and iterative conditional expectations (ICE). We found that long-term adherence to recommended MVPA improved the population levels of many frailty-associated proteins (ranging from 0.04 to 0.11 standard deviation); the greatest benefits were seen in proteins involved in the nervous system and inflammation. Sensitivity analyses suggested that the results were robust to unmeasured confounding and left truncation due to death.

Applications in Health and Biology

The Impact of Unconditional Cash Transfer Payment Frequency on Dietary Choices: A Multivalued Treatment Model Nicolas Guzman-Tordecilla* Nicolas Guzman-Tordecilla, Antonio Trujilla, Andres Vecino-Ortiz, Shu Wen Ng,

Low-income elderly populations face significant challenges in managing resources for dietary health, yet they remain understudied in public financial assistance (PFA) research. This study investigates how the payment frequency of an unconditional cash transfer program influences dietary behaviors among this group in Colombia. Using nationally representative data, we evaluated the causal effects of monthly versus bi-monthly payments on food expenditures, dietary diversity, and diet quality. We employed a Multivalued Treatment Model (MTM) with Augmented Inverse Probability Weighting—a novel approach that advances propensity score methods through doubly robust estimation, ensuring unbiased results when either the propensity score or outcome model is correctly specified. This method enables precise estimation of causal effects across multiple intervention levels, such as payment frequencies, while addressing selection bias. Despite its robustness, the MTM remains underutilized in empirical research, making this study one of the first to apply it to consumer behavior and nutrition outcomes, bridging a critical methodological gap. Findings demonstrate that monthly payments significantly enhance dietary diversity and quality by supporting healthier food purchases compared to bi-monthly payments. By advancing causal inference methods and linking income timing to health outcomes, this study offers insights for optimizing PFA programs in fostering healthier consumer behavior.

Applications in Health and Biology

Estimating subgroup effects of prenatal opioid exposure across levels of baseline risk factors Andy Shen* Andy Shen, Sherian X. Li, Michael W. Kuzniewicz, Lena S. Sun, Samuel D. Pimentel,

Prenatal opioid exposure (POE) is associated with adverse health outcomes and neurodevelopmental conditions later in childhood, including ADHD. Since ADHD risk is partially genetic and hereditary, its relationship with POE may be modified by maternal neurodevelopment history. In this study, we explore how effects of POE on ADHD risk vary with maternal ADHD and depression. Using a birth cohort from an integrated healthcare system with 15 facilities, we utilize cardinality matching to match each POE subject to five unexposed controls. We conducted a match for the entire POE group, separate cardinality matches for the POE group with maternal ADHD and the remaining POE group without maternal ADHD (matching to unexposed subjects with or without maternal ADHD), and separate cardinality matches for subgroups defined by maternal depression. We did not observe a significant effect of POE on ADHD overall, but effect estimates differed substantially for subjects with and without maternal ADHD, and were significant for the latter group. Our results suggest POE may have a smaller impact on babies at elevated risk for ADHD due specifically to maternal ADHD, although no such effect appears present for maternal depression.

This project is supported by the FDA.

Applications in Health and Biology

Causal Impact of Visit Frequency and Type on Lung Function in Cystic Fibrosis Patients: A Target Trial Emulation Approach Alexandra Hinton* Alexandra Hinton, Louisa Smith, Jonathan Zuckerman, Edmund Sears,

Background: The Cystic Fibrosis Foundation’s recommendation of at least four annual clinic visits is now two decades old. Since it was issued, life expectancy for pwCF has increased by 28 years as a result of treatment advances. The optimal frequency of clinic visits for CF patients needs reevaluation.

Methods: The study aims to estimate the causal effect of different clinic visit frequencies and types on lung function in stable adult CF patients. Using data from the Cystic Fibrosis Foundation Patient Registry (2022-2023: ~20,000 patients), the study will emulate randomization of patients to 12 treatment strategies, combining three care types (in-person, telehealth, hybrid) and four visit frequencies (quarterly, triennial, biannual, annual; subject to positivity constraints).

Planned Analysis: We will use doubly robust methods to estimate the effect of visit patterns on lung function. Sensitivity analyses will be conducted to address potential biases from missing data, missed visits, and unmeasured confounding.

Expected Results: This study will provide evidence on the causal impact of different clinic visit frequencies on lung function in stable adult CF patients. Results are pending and will be available for presentation at the conference.

Conclusions: Findings from this study may inform updates to clinical guidelines for CF care, potentially optimizing resource utilization while maintaining or improving patient outcomes.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

Interference in Causal Inference into Household Mortgage Default Decisions Walter Torous*
Walter Torous, William Torous,

Strategic default theory argues that homeowners have an economic incentive to default when their home's value falls below the amount owed on the property. Empirical evidence suggests that strategic defaults by homeowners were common during the Global Financial Crisis.
Recently, Ganong and Noel (GN) put forward a causal attribution methodology to investigate household mortgage default decisions and conclude that no defaults can be accounted for by strategic theory. Rather, all defaults are explained by homeowners' inability to pay stemming from negative life events. These results have important policy implications as they suggest that any debt restructuring programs should focus on liquidity provision rather than principal reduction.
GN, however, ignore the possibility of interference in their causal attribution methodology. There is ample empirical evidence that observing others in foreclosure is an important determinant when considering the option to default yourself.
This paper explores the implications of contagion on GN's causal attribution methodology. We do so by embedding a network model of neighborhoods that allows endogenous interactions among nearby neighbors to influence a homeowner's likelihood of default.
We find that GN's estimate of the share of defaults attributable to strategic theory is downward biased in the presence of contagion. The magnitude of the bias depends critically on neighborhood characteristics and can be significant.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

E-Processes for Sequential Tests of Inhomogeneous Poisson Point Processes Michael Lindon* Michael Lindon, Nathan Kallus,

Motivated by monitoring the arrival of incoming adverse events such as customer support calls or crash reports from users exposed to an experimental product change, we consider sequential hypothesis testing of continuous-time inhomogeneous Poisson point processes. Specifically, we construct a continuous-timed closed-form e-process for testing the equality of arrival rates with a time-uniform Type-I error guarantee at a nominal α. We characterize the asymptotic growth rate of the proposed e-process under the alternative and show that it has power 1 when the average rates of the two Poisson process differ in the limit. We complement our testing procedure by providing multivariate confidence sequences for inference on the cumulative rates and observe an interesting relationship to the universal inference e-process for testing composite null hypotheses.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

How Environmental Factors Drive Convective Cloud Detrainment Heights: An Application of LiNGAM Dié Wang* Dié Wang, Simon Lee, Tao Zhang,

This study investigates how environmental factors influence the level of maximum detrainment (LMD) in deep convective clouds (DCCs). Through an application of the Linear Non-Gaussian Acyclic Model (LiNGAM), we discover causal structures between environmental variables and LMD, observed at six tropical sites operated by the Atmospheric Radiation Measurement (ARM) user facility. LiNGAM effectively identifies causal directions among variables of interest, revealing robust relationships such as those among the lifting condensation level (LCL), level of free convection (LFC), and convective inhibition (CIN), aligning with prior knowledge. Relative humidity is shown to directly influence LMD; however, this relationship exhibits strong nonlinearity and becomes difficult to detect when the contrast between oceanic and continental environments is excluded from the analysis.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

Impact of Metacognitive Conversations In Large Scale LLM based Intelligent Tutoring System Duy Pham* Duy Pham, Kirk Vanacore, Zeyneb Sena Sarioglan, Owen Henkel, Ryan Baker,

In an effort to revolutionize math education in underserved regions, our project leverages advanced AI technology to enhance both student engagement and performance in Rori, an open-source Intelligent Tutoring system. Rori is an AI-powered math tutor accessible via WhatsApp that delivers over 500 micro-lessons based on the Global Proficiency Framework, providing scaffolded practice questions and interactive feedback through natural language processing. This project aims to enhance student engagement and academic performance on Rori by integrating two LLM-based metacognitive conversational modules—one that normalizes making mistakes and another that promotes a growth mindset—and by employing propensity score matching to balance key covariates such as student age, the level of difficulty of content they start in the platform, and their reasons for using Rori, thereby isolating the causal impact of these interventions as we track usage metrics and math performance outcomes.

Bayesian Causal Inference

Bayesian inference for direct and spillover effects for two-stage randomized experiments affected by noncompliance Claudia Mastrogiacomo* Claudia Mastrogiacomo, Laura Forastiere, Joshua Warren,

Interference occurs when one individual's outcome depends on the treatments of other individuals. Two-stage randomized experiments, where clusters are first randomly assigned to a certain proportion of treated individuals (dosage) and individuals are then randomly assigned to a treatment, have been used to estimate direct and spillover effects in the presence of interference. However, estimation of causal effects in two-stage randomized trials can be complicated by noncompliance since, in a setting with interference, an individual's potential treatment uptake and potential outcomes may be influenced by the treatment assignment and uptake of others.

Under the principal stratification framework, we define principal strata based on potential treatment uptake and on whether and at which dosage an individual would switch their compliance behavior. We develop a Bayesian inference method to impute compliance types and estimate direct and spillover effects within strata, modeling the outcomes as a Gaussian process with the covariance matrix depending on the switching thresholds. We apply our novel method to data from a two-stage randomized experiment conducted in villages in rural Honduras to estimate direct and spillover effects of a maternal and child health intervention.

Causal Discovery

The Impact of the COVID-19 Pandemic on PhD Education: An Empirical Analysis of a Large Comprehensive European University Maxime François* Maxime François, Kristof De Witte,

The COVID-19 pandemic profoundly disrupted higher education, yet its specific effects on PhD students remain underexplored. This study uses a quasi-experimental identification strategy, employing difference-in-differences and supervisor fixed-effects models, to causally analyze the pandemic's impact on doctoral education outcomes, including time to graduation and graduation probabilities. Drawing on a panel dataset from 2010 to 2023, covering 19,565 PhD students and 2,126 supervisors at a top-50 European university, we find that the probability of graduation decreased by 3.5% in 2020, rebounded in 2021, but declined by 4.6% in 2022. Additionally, the average time to graduation extended by 2 months in 2021 and 1.4 months in 2022, with signs of recovery by 2023. These findings highlight significant disruptions in doctoral timelines due to the pandemic. Interestingly, scholarship discontinuations dropped by 18.5% in 2020, attributed to university closures enabling professors to focus on research, which led to increased successful funding applications. This stability in funding contributed to steady PhD enrollment in 2020 and a 19% surge in 2021. Field-specific analysis revealed that Biomedical Sciences were the most affected, with graduates experiencing an average delay of 4 months compared to other research fields. These results underscore the pandemic's heterogeneous impacts on doctoral education and emphasize the need for tailored interventions to mitigate future disruptions.

Causal Discovery

Integrating Time Series Analysis and Causal Discovery for Enhanced Sequential Data Understanding Suat Babayigit* Suat Babayigit,

Understanding sequential data and uncovering causal relationships are critical challenges in modern data analysis. This study integrates time series analysis with causal discovery methods to address limitations such as non-stationarity, irregular sampling, and multivariate causality in existing approaches. A novel method combining Fourier analysis and Granger causality is proposed, enhancing the strengths of each technique while mitigating their weaknesses. The method was evaluated using real-world datasets across domains, yielding robust performance in identifying causal structures within complex time series data. Results highlight the potential for improved accuracy and interpretability in sequential data modeling. This work contributes to advancing the fields of time series analysis and causal inference, offering practical insights for applications in healthcare, finance, and environmental science.

Causal Discovery

A Weighted Estimation Approach for Combining Causal Effect Estimates from Multiple Testable Causal Models Ina Ocelli* Ina Ocelli, Ted Westling, Rohit Bhattacharya,

In observational studies, causal inference frequently requires choosing among multiple plausible causal models. Determining which models, if any, are correctly specified poses a challenge which can compromise the reliability of inferences. Recently, researchers have explored empirical tests for validating causal models, such as the testability of backdoor models (Entner et al., 2013) and front-door models (Bhattacharya and Nabi, 2022). While these developments enable valid pre-tests, they also introduce post-selection inference issues.
To address these limitations, we propose a novel approach that unifies model testing and causal effect estimation. Unlike existing methods, which often treat testing and estimation as separate tasks, our method computes both an association measure indicating model correctness and the causal effect estimates within a single framework. The association measures and effect estimates are plugged into a weighted estimator. We show that this weighted estimator is consistent and accounts for the variability in both testing and estimation procedures, as long as at least one model and the empirical test associated with it are correct. This design accounts for issues related to post-selection inference while ensuring robust causal effect estimation by taking advantage of model testability. Through simulation studies, we demonstrate the effectiveness of our method in enhancing the reliability of causal inference in challenging observational study settings.

Causal Discovery

Understanding the Comorbidity of Epilepsy and Depression on Sleep Disorder by Analyzing National Health and Nutrition Examination Survey (NHANES) data Vasundhara Acharya*
Vasundhara Acharya, Bulent Yener, Madeline C Fields, Lara Marcuse,

Epilepsy, sleep disorders (SD), and major depressive disorder (MDD) often co-occur, complicating care. MDD, common in epilepsy, may mediate its impact on sleep, while some antiseizure medications (ASMs) can worsen mood or sleep disturbances. Age, race, and Body Mass Index (BMI) further modulate these effects. Traditional logistic regression and pairwise analyses cannot capture such complexities. We propose a Structural Causal Model (SCM) learned from NHANES 2015–2020 data. Categorical variables are encoded using a Variational Autoencoder to improve on one-hot encoding. Bootstrapping is applied to each causal structure learning algorithm to compute edge confidence (EC) scores. Flipped-edge conflicts are resolved via EC comparison and significance tests (z-test). Aggregated edges and scores form an ensemble graph representing part of SCM, benchmarked against domain knowledge. Our causal estimations assume causal sufficiency (no unobserved confounders). Our initial analysis shows that MDD is a major driver of SD (EC=0.52), and epilepsy also contributes (EC=0.34). Age increases the likelihood of ASM use, with an overall average treatment effect (ATE) of 0.0217 ($p<0.001$). Conditional estimates reveal that in the highest BMI category (≥25), the ATE of age on medication intake is 0.02135 for Black patients versus 0.02942 for White patients, warranting longitudinal validation. MDD’s ATE of 0.205 on SD ($p<0.001$) underscores the need to address depression to improve sleep outcomes.

Causal Fairness, and Bias/Discrimination

General Methods for Fair Classification by Constraining Path-Specific Causal Effects León Segovia* León Segovia, Herbert Susmann, Iván Diaz,

Machine learning (ML) for efficient decision making has proliferated across domains, typically in matters concerning resource allocation and policy implementation. It is well documented that these algorithms can codify sociostructural bias present in training data; in fact, perpetuating unequal treatment of minoritized groups. To overcome this issue, researchers must be able to formalize the underlying relationships between discrimination and existing disparities. We use causal mediation analysis —a promising way to characterize fairness in ML—to achieve this goal by creating classification methods to improve fair outcomes. Building on existing work, our methods operationalize fairness as the optimization of the expected loss under constraints of path-specific effects (PSE). This project extends existing methods for constrained risk minimization in two ways: using causal parameters developed in prior work to address problems where constraint paths involve intermediate confounding; as well as where the exposure is non-binary. We implement recently developed causal parameters as alternative definitions of PSE to appropriately account for interventions on information shared between variables; enabling a more rigorous analysis of the complex dynamics between discrimination and disparities mediated by socioeconomic factors. We establish the relevance of our novel methods by applying them to case studies in public health research.

Causal Fairness, and Bias/Discrimination

A Systems Thinking Approach to Algorithmic Fairness Chris Lam* Chris Lam,

Systems thinking provides us with a way to model the algorithmic fairness problem by allowing us to encode prior knowledge and assumptions about where we believe bias might exist in the data generating process. We can then encode these beliefs as a series of causal graphs, enabling us to link AI/ML systems to politics and the law. This allows us to combine techniques from machine learning, causal inference, and system dynamics in order to capture different emergent aspects of the fairness problem. We can use systems thinking to help policymakers on both sides of the political aisle to understand the complex trade-offs that exist from different types of fairness policies, providing a sociotechnical foundation for designing AI policy that is aligned to their political agendas and with society’s values.

Causal Fairness, and Bias/Discrimination

A Systems Thinking Approach to Algorithmic Fairness * Chris Lam, Chris Lam,

Systems thinking provides us with a way to model the algorithmic fairness problem by allowing us to encode prior knowledge and assumptions about where we believe bias might exist in the data generating process. We can then encode these beliefs as a series of causal graphs, enabling us to link AI/ML systems to politics and the law. This allows us to combine techniques from machine learning, causal inference, and system dynamics in order to capture different emergent aspects of the fairness problem. We can use systems thinking to help policymakers on both sides of the political aisle to understand the complex trade-offs that exist from different types of fairness policies, providing a sociotechnical foundation for designing AI policy that is aligned to their political agendas and with society’s values.

Causal Inference and SUTVA/Consistencies Violations

Scaling-Up Experiments in Centralized Markets Wisse Rutgers* Wisse Rutgers, Dmitry Arkhangelsky,

This paper focuses on identification of the marginal policy effect (MPE, Wager and Xu 2021) in centralized markets. Hu, Li, and Wager (2021) show that the MPE can be decomposed in a direct and indirect effect. Where the direct effect is simply equal to a traditional average treatment effect, we show how to leverage knowledge about the market structure in centralized markets to identify the usually hard-to-estimate indirect effect. Furthermore, we show that in certain centralized markets where the mechanism contains a random component we can identify a causal effect not only for the MPE, but for a range of different treatment probabilities. We consider centralized markets similar to that of Munro (2024), where products are allocated by a centralized mechanism based on reports submitted by agents. When agents' preferences in the market are directly observed from their reports the framework can be applied straightforwardly. Furthermore, when preferences are not directly observed but can be derived from observables the framework still applies.

Causal Inference Education

Formation of a University's Causal Inference Collaboratory Emily Roberts* Emily Roberts,

A collaboratory is a creative group process designed to solve complex problems that brings the opportunity for new organizational networks to form. This year the Institute for Public Health Practice and Research Policy funded our proposal to establish the Causal Inference Collaboratory at the University of Iowa. This initiative aims to foster collaboration and methodological advancements, positioning our budding group as a resource for researchers to advance causal inference research at our university.

Our group has three primary aims: 1. To conduct a review of how causal theory and methods can provide innovative insights into public health research broadly and at our university. 2. Develop a training program through workshops and collaborative projects with a Graduate Research Assistant. Our workshops showcase various research methods, equipping participants with the skills necessary to conduct their own research and contribute to ongoing causal projects. 3. Create a platform for collaboration and continuous learning through working groups. This component emphasizes collaboration on competitive grants related to causal inference research.

This talk will showcase our successes and challenges of working toward these aims, highlighting the outcomes achieved through new collaborations and impactful research. Findings will include characterization of our ongoing research and teaching activities related to causal inference across our College of Public Health and beyond.

Causal Inference Education

The Influence of Nudging on Summer School Attendance Intentions: A RCT in Flanders
Diogo Vieira Nunes da Conceição* Diogo Vieira Nunes da Conceição, Kristof De Witte,

This study explores how behavioural nudges and anchoring biases affect primary school students' decisions regarding summer school participation in the Flemish region of Belgium. Our randomized controlled trials implemented on the last two weeks of the 2023/24 academic year revealed that students shown a low irrelevant anchor reported to be able to remember, on average, two words less out of a list of 15 words than their peers (N = 492). Respondents in the fourth grade or above with a low socioeconomic status exposed to a high anchor reported to be able to remember 1.7 words more than their control counterparts. Nudges emphasizing the trade-off inherent to attending a summer school reduced willingness to participate by 0.52 standard deviations for lower SES students (N = 567). While some nudges decreased participation intentions, they improved students' perceptions of friendships and of their school, highlighting the need for tailored interventions in educational policymaking.

Causal Inference in Networks

Design-based weighted regression estimators for conditional spillover effects Fei Fang* Fei Fang, Edoardo Airoldi, Laura Forastiere,

In a clustered interference setting, with networks collected within clusters and no interference between clusters, we introduce a general causal estimand for conditional spillover effects, offering flexible ways of integrating unit-to-unit spillover effects. In particular, we define spillover effects from the treatment received by one neighbor, averaged over the distribution of the cluster treatment, and conditional on the characteristics of the treated unit. Such definition enables to access the heterogeneity of a unit's spillover effect on their neighbors with respect to the unit's characteristics. Two weighted regression-based estimators are proposed: i) at the individual level, taking neighbors' averages either in the outcomes or in the treatments within weights; and ii) at the dyadic level, where the outcome of one unit is regressed on the treatment of each neighbor. When covariates driving the heterogeneity are categorical, we prove the equivalence of the two regression-based estimators to the non-parametric Hajek estimator. For continuous covariates, we demonstrate that both estimators consistently estimate the proposed estimands. Under a design-based perspective, we derive HAC variance estimators and establish the central limit theorem. Simulations are conducted to compare the performance of our estimators. Finally, we apply our methods to a randomized experiment conducted in Honduras to evaluate the spillover effect of a behavioral intervention.

Causal Inference in Networks

Peer effects in the linear-in-means model may be inestimable even when identified Alex Hayes* Alex Hayes, Keith Levin,

Linear-in-means models are widely used to investigate peer effects. Identifying peer effects in these models is challenging, but conditions for identification are well-known. However, even when peer effects are identified, they may not be estimable, due to an asymptotic colinearity issue: as sample size increases, peer effects become more and more linearly dependent. We show that asymptotic colinearity occurs whenever nodal covariates are independent of the network and the minimum degree of the network is growing. Asymptotic colinearity can cause estimators to be inconsistent or to converge at slower than expected rates. We also demonstrate that dependence between nodal covariates and network structure can alleviate colinearity issues in random dot product graphs. These results suggest that linear-in-means models are less reliable for studying peer influence than previously believed.

Causal Inference in Networks

Estimation of Causal Effects Under K-Nearest Neighbors Interference Michael Higgins*
Michael Higgins, Samirah Alzubaidi,

Treatment interference occurs when the treatment status of one unit may affect the response of another unit. Such settings are becoming increasingly common, especially for experiments conducted on social networks. We consider a model of treatment interference, the K-nearest neighbors interference model (KNNIM), for which the response of one unit depends not only on the treatment status given to that unit, but also the treatment status of its K closest neighbors. We derive causal estimands under KNNIM in a way that allows us to identify how each of the K-nearest neighbors contributes to the indirect effect of treatment. We propose unbiased estimators for these estimands and derive conservative variance estimates for these unbiased estimators. We then consider extensions of these estimators under an assumption of no weak interaction between direct and indirect effects. We apply our methodology to an experiment designed to assess the impact of a conflict-reducing program in middle schools in New Jersey, and we give evidence that the effect of treatment propagates primarily through a unit’s closest connection.

Causal Inference in Networks

Causal Inference in Dynamic Networks Peem Lerdputtipongporn* Peem Lerdputtipongporn, David Choi, Nynke Niezink,

For longitudinal network settings, this study introduces a novel methodology for evaluating the effects of dyadic interventions (i.e., intervening on a network connection). The approach can be applied to either dyadic-level, such as whether the presence of a connection causes the tie to persist or be reciprocated in subsequent periods, or to individual-level outcomes, such as whether the presence of a connection between two individuals causes their behaviors in subsequent periods to become more similar. The methodology employs a doubly-robust estimation technique that combines low-rank matrix approximation with longitudinal network models, for which we establish unbiasedness and consistency under certain conditions.

Causal Inference with Dependent Data

Non-parametric Counterfactual Regression with Applications in Causal Inference with Dependent Data Prabrisha Rakshit* Prabrisha Rakshit, Arun Kumar Kuchibhotla, Eric Tchetgen Tchetgen,

Series regression estimates the conditional mean of a response by regressing on features derived from basis functions. OLS-based series estimators achieve minimax optimality but rely on restrictive assumptions about basis functions. To address this, [1] introduced the Forster-Warmuth (FW) learner, which extends to counterfactual regression using a unified pseudo-outcome approach, reducing bias from nuisance function estimation and achieving minimax rates under mild conditions. However, these results depend on an i.i.d. sample condition. We extend the FW-learner framework to dependent data settings, such as time series and spatial structures. Under certain dependence conditions, the FW learner's l_2 error rate matches the i.i.d. case, preserving minimax optimality. This extension is computationally efficient and broadens the FW framework to dependent contexts. We demonstrate its utility by estimating dose-response curves for continuous treatments under both unconfoundedness and unmeasured confounding scenarios. As an application, we estimate air pollution's immediate effects on heart attack rates, using same-day pollution fluctuations to reveal actionable health patterns. This work highlights how computational advancements in nonparametric regression can address dependencies while maintaining theoretical guarantees.
[1]Y Yang, A K Kuchibhotla, and E J. Tchetgen Tchetgen. Forster–Warmuth Counterfactual Regression: A Unified Learning Approach. (2023)

Design of Experiments

Adaptive Experimental Design Using Shrinkage Estimators Evan Rosenman* Evan Rosenman, Kristen Hunter,

In multi-armed trials, adaptive designs are widely used to improve estimation efficiency, identify optimal treatments, or maximize rewards. Recent studies have explored using adaptive trials to achieve better simultaneous estimates of the effects of K active treatments relative to a control arm. One approach is to employ either batch or sequential variants of Neyman allocation, typically using Horvitz-Thompson-style estimators to produce causal estimates at the trial's conclusion. However, this approach may be inefficient in that it fails to borrow information across the treatment arms.

In this paper, we consider adaptivity when using a Stein-like shrinkage estimator for heteroscedastic data to estimate the causal effects in multi-arm trials. This estimator pools information across treatment effect estimates, provably reducing the expected squared error loss compared to independent estimation of treatment effects. We demonstrate that the expected loss can be expressed as a Gaussian quadratic form, enabling efficient computation via numerical integration. This result paves the way for sequential adaptivity, allowing treatments to be assigned to minimize the shrinker loss. Through simulations, we demonstrate that this approach can yield meaningful reductions in estimation error and characterize how our adaptive algorithm assigns treatments differently than would a sequential Neyman allocation.

Design of Experiments

Designing Experiments to Evaluate Algorithm-Assisted Decision-Making in the Presence of Behavior Adaptation Maggie Wang* Maggie Wang, Michael Baiocchi,

Algorithm-based decision support tools have the potential to enhance decision quality in a range of applications, including in medicine and in criminal justice. Algorithms are commonly evaluated on retrospective data using performance metrics like accuracy, but these metrics do not necessarily reflect how an algorithm-based tool would change decision-making if it were integrated into a workflow where the human user has the final say. Not only might decision-makers be overly trustful of or averse to tool output, but decision-making behavior may actually adapt with repeated exposure to the tool, e.g. due to gradual building of trust. While randomized experiments can provide better evidence on the impact of decision support tools than retrospective studies, traditional experimental designs and effect estimands still may not correctly capture the impact of the tool because they ignore behavior adaptation. If unaccounted for, behavior adaptation can lead to biased estimates of tool effectiveness and incorrect conclusions about whether the tool should be deployed. In this work, we define time-varying effect estimands that either account for or explicitly measure the impact of behavior adaptation by contrasting different sequences of tool exposure. We then propose experimental designs that target these effect estimands. We demonstrate the utility of our designs with simulations based on a real-world pilot study of a clinical decision support tool that predicts patient deterioration.

Design of Experiments

Towards Efficient Statistical Inference and Optimal Design in Adaptive Experiments Wenxin Zhang* Wenxin Zhang, Mark van der Laan,

Adaptive experiments play a crucial role in clinical trials and online A/B testing. Unlike static, non-adaptive trial designs, adaptive experimental designs dynamically adjust treatment randomization probabilities and other key design elements in response to data collected sequentially during the experiment. These designs are useful for achieving different objectives, such as reducing uncertainty in causal estimand estimation or improving benefit of participants within the experiment. Despite their advantages, the adaptive nature of these designs and the time-dependent nature of the data introduce significant challenges in making unbiased statistical inferences from non-i.i.d. data. Building upon the Targeted Maximum Likelihood Estimator (TMLE) literature that has provided valid statistical inference tailored to adaptive experimental settings using inverse weighting strategies tailored for adaptive experiment settings, we propose a new TMLE that further improves the efficiency for estimating causal estimands under adaptive designs.
Beyond efficient statistical inference, we further introduce a general framework for implementing optimal adaptive designs, customized to achieve various objectives efficiently. The performance of our proposed estimators and adaptive designs is demonstrated through theoretical analysis and extensive simulations.

Design of Experiments

Causal Inference for Ordinal Outcomes with Temporal Structure in Randomized Experiments Rituparna Dey* Rituparna Dey, Tirthankar Dasgupta,

Randomized experiments involving sequentially collected ordinal outcomes over time, have been largely unexplored in causal inference. While most causal inference studies focus on continuous outcomes, standard causal parameters like the average treatment effect (ATE) lose their meaning when applied to ordinal data. This study introduces novel nonparametric causal estimands for randomized experiments, addressing both the ordinal nature and temporal structure of the data. A motivating industrial experiment illustrates these challenges, where treatment effects evolve over multiple time points. Our approach extends the potential outcomes framework and the distributional causal estimands and their sharp bounds within the temporal context. Furthermore, we propose time-adjusted estimands utilizing past outcomes for sharper inference. Estimation is performed using a matching-based imputation approach, ensuring that the missing potential outcomes are imputed while preserving the ordinal structure. A simulation study, employing a latent variable model, demonstrates the advantages of our approach, showing improved estimation accuracy over traditional methods. This work has broad applicability in fields such as clinical trials, user experience research, and social sciences.

Design of Experiments

Causal Inference for Ordinal Outcomes with Temporal Structure in Randomized Experiments *
Rituparna Dey, Tirthankar Dasgupta, Rituparna Dey,

Randomized experiments involving sequentially collected ordinal outcomes over time, have been largely unexplored in causal inference. While most causal inference studies focus on continuous outcomes, standard causal parameters like the average treatment effect (ATE) lose their meaning when applied to ordinal data. This study introduces novel nonparametric causal estimands for randomized experiments, addressing both the ordinal nature and temporal structure of the data. A motivating industrial experiment illustrates these challenges, where treatment effects evolve over multiple time points. Our approach extends the potential outcomes framework and the distributional causal estimands and their sharp bounds within the temporal context. Furthermore, we propose time-adjusted estimands utilizing past outcomes for sharper inference. Estimation is performed using a matching-based imputation approach, ensuring that the missing potential outcomes are imputed while preserving the ordinal structure. A simulation study, employing a latent variable model, demonstrates the advantages of our approach, showing improved estimation accuracy over traditional methods. This work has broad applicability in fields such as clinical trials, user experience research, and social sciences.

Design-Based Causal Inference

Sensitivity Analysis for Attributable Effects in Case^2 Studies Kan Chen* Kan Chen, Ting Ye, Dylan Small,

The case^2 study, also referred to as the case-case study design, is a valuable approach for conducting inference for treatment effects. Unlike traditional case-control studies, the case^2 design compares treatment in two types of cases with the same disease. A key quantity of interest is the attributable effect, which is the number of cases of disease among treated units which are caused by the treatment. Two key assumptions that are usually made for making inferences about the attributable effect in case^2 studies are 1.) treatment does not cause the second type of case, and 2.) the treatment does not alter an individual's case type. However, these assumptions are not realistic in many real-data applications. In this article, we present a sensitivity analysis framework to scrutinize the impact of deviations from these assumptions on obtained results. We also include sensitivity analyses related to the assumption of unmeasured confounding, recognizing the potential bias introduced by unobserved covariates. The proposed methodology is exemplified through an investigation into whether having violent behavior in the last year of life increases suicide risk via 1993 National Mortality Followback Survey dataset.

Design-Based Causal Inference

Improving efficiency in double-sampling strategies for informative missing data in treatment effect estimation Shuo (Mila) Sun* Shuo (Mila) Sun, Alex Levis, Rajarshi Mukherjee, Rui Wang, Sebastien Haneuse,

Missing or incomplete data is a widespread challenge in observational studies, especially when the data at hand were not originally collected for research purposes. These data may be particularly susceptible to being missing-not-at-random (MNAR). To mitigate bias due to MNAR data, we propose to use a double-sampling strategy, through which the otherwise missing data are ascertained on a sub-sample of study units. We generalize the nonparametric estimation results to the case where the data are initially subject to arbitrary coarsening, and develop nonparametric efficient estimators of any smooth full data functional of interest. Since the double-sampling strategy can be planned from the beginning, it provides an opportunity to allocate resources effectively within a fixed budget. Motivated by this, we derive the optimal sampling rule that minimizes semiparametric efficiency bound, subject to a budget constraint. The optimal double-sampling rules generally depend on the unknown full data distribution. To address this, we conduct a pilot study to estimate unknown quantities and investigate asymptotic properties, using average treatment effects (ATEs) as an example, considering both fixed pilot sample sizes and cases where the sample size approaches zero at a specific rate. Two simulation studies, assuming Hŏlder smooth functions and sparsity functions, respectively, verify the efficiency of the proposed optimal sampling rules in finite samples.

Design-Based Causal Inference

Randomization Inference with Sample Selection Zeyang Yu* Zeyang Yu, Xinran Li, Peizan Sheng,

Randomization inference (RI) is widely used in scientific fields due to its two key advantages: it makes no distributional assumptions and leverages the structure of the experimental design. Most literature focuses on the ideal scenario with no missing outcomes, but sample attrition is common in experiments. Current work often relies on strong assumptions, like missing at random or sharp missingness, to validate RI. When this assumption fails, it can cause severe size distortion in the RI procedure. We first show that when testing a sharp null hypothesis, we can obtain a valid p-value by using the worst-case p-value under arbitrary missingness mechanisms. Finding the worst-case p-value boils down to finding an imputation of the missing outcomes that minimizes a distribution-free test statistic. We then extend the worst-case inferential approach to test the quantiles of the individual effect. This involves utilizing the worst-case imputation for missing outcomes and the imputation procedure in Caughey et al. (2023), simultaneously. Furthermore, we show that the conservative test for testing a sharp null and treatment effect quantile can be improved by incorporating additional assumptions on missingness mechanisms, namely, monotone missingness, sharp missingness, and missing at random. We illustrate our methods with simulations and an application.

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

Granular Synthetic Control Yuhang Zhang* Yuhang Zhang, Haitian Xie,

Unlike cross-sectional or short-panel causal inference methods such as unconfoundedness, instrumental variables, and difference-in-differences, which delineates nonparametric identification and semiparametric estimation stages, the classic synthetic control method estimates the counterfactual outcome directly through weighted combinations of control units, suppressing a distinct nonparametric identification stage. In this paper, we introduce a refined synthetic control method that, with access to granular-level data, develops nonparametric identification and then proceeds to semiparametric estimation. Leveraging micro-level data allows for more flexible, covariate-specific weighting, and requires only two time periods for identification. We explore this methodology in both panel and repeated cross-sectional settings, developing doubly robust identification and proposing semiparametric estimation based on double/debiased machine learning methods. The effectiveness of our approach is demonstrated through an empirical application.

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

Estimating effects of longitudinal modified treatment policies (LMTPs) on rates of change in health outcomes with repeated measures data Daniel Malinsky* Anja Shahu, Daniel Malinsky,

Longitudinal modified treatment policies (LMTPs) quantify the effects of interventions that depend on the natural value of exposure, generalizing “stochastic” and “shift” interventions as well as other policy-relevant quantities. The current LMTP estimation approach yields effects on outcomes measured at the end of a study; however, repeated measures data often contains time-varying outcomes measured at each visit and interest may lie in estimating effects on the rate of change in these outcomes over time. For example, one may wish to quantify the effect of an LMTP on the rate of progression of a disease. We extend the LMTP approach to estimate the effect on change in a time-varying outcome over time and propose a hypothesis testing framework to formally test whether there is a difference in change in the outcome over time under an LMTP versus the natural outcome trajectory (or versus a different LMTP). Repeated measures data also frequently has unique data complications that must be considered. One such complication is that of irregular visit times, where the visit timing varies among individuals from some pre-specified time. We propose an extension to our work that permits effect estimation and hypothesis testing for an LMTP in a setting with irregular visit times. We present results from a simulation study which shows that ignoring irregular visit times may lead to bias, and we illustrate our hypothesis testing framework in both regular and irregular visit time settings.

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

Cluster-robust inference with a single treated cluster using the t-test Xinran Li* Xinran Li, Chun Pong Lau,

We consider the situation where treatment is assigned at the cluster level and unobserved dependencies exist among units within each cluster. This situation often occurs in difference-in-differences estimation, where a single treated cluster is compared to a finite number of control clusters. We assume the availability of asymptotically Gaussian cluster-level estimators, albeit with asymptotic variances that are unknown and challenging to estimate due to dependencies within clusters. Inference for treatment effects in this context is equivalent to a two-sample testing problem, where (i) one group comprises a single observation while the other includes a finite number of observations with a common mean, and (ii) all observations follow independent Gaussian distributions with potentially heteroskedastic and unknown variances. We propose exact t-tests tailored to this problem, incorporating constraints on relative heterogeneity of variances across groups. We illustrate the advantage of the proposed method through both simulations and empirical applications.

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

Parallel Trends in an Unparalleled Pandemic: Difference-in-differences for infectious disease policy evaluation Alyssa Bilinski* Alyssa Bilinski, Shuo Feng,

Researchers frequently employ difference-in-differences (DiD) to study infectious disease policy. DiD assumes that treatment and comparison groups would have moved in parallel in expectation, absent the intervention (“parallel trends assumption”). Our work formalizes often unaddressed epidemiological assumptions required for common DiD specifications, assuming an underlying Susceptible-Infectious-Recovered (SIR) data-generating process, and proposes more robust specifications. We first demonstrate that popular specifications can encode strict assumptions: DiD modeling incident infections or rates will produce biased treatment effect estimates unless untreated potential outcomes for both groups come from a data-generating process with the same initial infection and transmission rates. Modeling log incidence or growth allows for different initial infection rates, but invokes conditions on transmission parameters. We propose alternative specifications based on epidemiological parameters — the effective reproduction number and the effective contact rate — that are both more robust to differences between groups and can be extended to more complex transmission dynamics. In power analyses, we highlight minimal differences between incidence and log incidence models; our alternative specifications have lower power than incidence or log incidence, but higher power than log growth. We illustrate practical implications re-analyzing published studies of COVID-19 mask policy.

Dynamic Treatment Regimes

Dynamic Local Average Treatment Effects Ravi Sojitra* Ravi Sojitra, Vasilis Syrgkanis,

We enable identification, estimation, and inference for Local Average Treatment Effect (LATE) estimands in multiple time period settings with noncompliance and allow for treatment dynamics. Dynamics occur when treatment is encouraged in each time period depending on previous encouragements, treatments, and states (e.g. short term outcomes and time varying covariates). Although one may hope to leverage estimates of the effects of switching from one sequence of treatments to another, such quantities are not estimable under noncompliance and standard identifying assumptions for dynamic settings and LATEs. We introduce two conditions to enable identification of Dynamic LATEs that quantify effects for subpopulations who would comply with encouragements. First, we show that One Sided Noncompliance enables identification of all Dynamic LATEs corresponding to treating in a single period only. Second, further adding Staggered Adoption enables identification of effects of treating in multiple time periods. In general, the second result holds when the treatment effect of not continuing to comply is uncorrelated with whether one continues complying. We also show that a sequential extension of Monotonicity in Imbens and Angrist (1994) is not sufficient for the first result and an additional assumption is necessary for the second result. Finally, we use the automatic debiased machine learning framework to perform plug-in estimation and inference based on our identification results.

Generalizability/Transportability

Efficient Estimation of Causal Effects Under Two-Phase Sampling with Error-Prone Outcome and Treatment Measurements Keith Barnatchez* Keith Barnatchez, Kevin Josey, Nima Hejazi, Bryan Shepherd, Giovanni Giovanni, Rachel Nethery,

In causal inference studies using electronic health record (EHR) data, clinical outcomes and treatments are commonly recorded with significant error. In practice, researchers can often validate error-prone measurements for a small, randomly selected subset of the full EHR dataset — a special case of two-phase sampling, where easily measured variables are collected for all subjects, and expensive-to-measure variables for a random subset. To improve efficiency, researchers frequently implement biased sampling designs, where validation probabilities depend on patients' initial error-prone measurements. In this work, we address the specific challenge of causal inference with error-prone outcome and treatment measurements under biased validation sampling designs, and the broader problem of causal inference under two-phase sampling. We highlight two asymptotically equivalent approaches to constructing nonparametric doubly-robust estimators of counterfactual means under general two-phase sampling designs. We argue these approaches can yield estimators with meaningfully different behavior in finite samples. For our specific measurement error problem, we construct novel doubly-robust estimators through each approach, and propose modifications to improve one approach's finite sample efficiency. Through simulation studies and data from the Vanderbilt Comprehensive Care Clinic, we demonstrate the efficiency gains our proposed methods can provide over current leading methods.

Generalizability/Transportability

Invariant Risk Minimization for Large Language Models Marko Veljanovski* Marko Veljanovski, Zach Wood-Doughty,

Invariant Risk Minimization (IRM) is a leading approach for out-of-distribution (OOD) generalization, framing optimization as finding a data representation such that the optimal classifier on top matches for every environment. While IRM has been extensively tested with image data, text-based datasets remain underexplored, despite OOD generalization being crucial for LLM performance. Accordingly, we aim to evaluate the effectiveness of directly applying IRM to text-based datasets, exploring optimization adjustments to enhance IRM's compatibility with LLM predictors. In particular, we adapt a data-generating process (DGP) by Wood-Doughty et al. (2021), originally created for evaluating causal inference methods, to create synthetic text to thoroughly evaluate IRM against Empirical Risk Minimization (ERM). Our DGP utilizes two parameters: tau, controlling the ordering correlation, and delta, controlling the ordering preference strength within a modified Zipfian distribution. Correspondingly, we compare IRM and ERM over varying values of tau and delta, revealing specific environments in which the otherwise optimal invariant predictor fails to achieve strong performance.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2021. Generating synthetic text data to evaluate causal inference methods. arXiv preprint arXiv:2102.05638.

Heterogeneous Treatment Effects

Evaluating Finite-Sample Properties of Machine Learning Approaches for Assessing Heterogeneity of Treatment Effect in Clinical Trials Lisa Levoir* Lisa Levoir, Bryan Blette, Andrew Spieker,

Inferring heterogeneity of treatment effect is a popular secondary aim of clinical trials. Recently, many trial analyses have moved from traditional subgroup analyses to more modern assessments of heterogeneity using machine learning. While there are several such methods available to estimate conditional average treatment effects (CATEs) in clinical trials, these methods are often applied in trial settings that have lower sample sizes than were considered in the simulations of corresponding seminal methodological work, making the validity of inference in these settings unclear. To provide guidance to practitioners, we conducted a simulation study to evaluate the performance of different regression and machine learning estimators for the CATE, including ordinary least squares (OLS) and causal forests, in a variety of settings across a range of sample sizes. We evaluated 95% confidence interval (CI) coverage, bias, and variance under linear and non-linear data generating mechanisms (DGM) in the presence of 0 to 100 nuisance covariates and 0 to 16 effect modifying covariates. We found that while tree-based ensembles like causal forests can be quite flexible to linear or nonlinear settings, they can have meaningfully impaired coverage in many settings at sample sizes which constitute most trial applications. As expected, OLS has superior performance under linear DGMs but has poor performance under nonlinear DGMs. We conclude with recommendations for practitioners.

Heterogeneous Treatment Effects

Guidance on Individualized Treatment Rule Estimation in High Dimensions Philippe Boileau* Philippe Boileau, Ning Leng, Sandrine Dudoit,

Individualized treatment rules, cornerstones of precision medicine, inform patient treatment decisions with the goal of optimizing patient outcomes. These rules are generally unknown functions of patients' pre-treatment covariates, meaning they must be estimated from clinical or observational study data. Myriad methods have been developed to learn these rules, and these procedures are demonstrably successful in traditional asymptotic settings with a moderate number of covariates. The finite-sample performance of these methods in high-dimensional covariate settings, which are increasingly the norm in modern clinical trials, has not been well characterized, however. We perform a comprehensive comparison of state-of-the-art individualized treatment rule estimators, assessing performance on the basis of the estimators' accuracy, interpretability, and computational efficacy. Sixteen data-generating processes with continuous outcomes and binary treatment assignments are considered, reflecting a diversity of randomized and observational studies. We summarize our findings and provide succinct advice to practitioners needing to estimate individualized treatment rules in high dimensions. All code is made publicly available, facilitating modifications and extensions to our simulation study. A novel pre-treatment covariate filtering procedure is also proposed and is shown to improve estimators' accuracy and interpretability.

Heterogeneous Treatment Effects

Hierarchical Approximations to the Universal Path for Efficient Targeted Maximum Likelihood Estimation Kaiwen HOU* Kaiwen HOU, Mark van der Laan,

TMLE promises efficiency by constructing paths in the statistical model space, solving the efficient score equation in as few updates as possible. In principle, the universal least favorable path (UFLP) achieves single-step convergence by exactly matching the canonical gradient at every measure along the path, but for many important causal parameters—such as the variance of CATE—this path is intractable to construct.

We propose a hierarchy of approximate paths derived from standard perturbation expansions of the ULFP's defining PDE. The local path in standard TMLE enforces a first-order condition only at the path's initial measure and solves the efficient score equation up to O(initial rate^2). Moving to second order dramatically reduces the remainder: we show that imposing an expectation-based second-order condition yields a path whose explicit construction minimizes a KL divergence with a natural interpretation of maximizing the Cramér–Rao lower bound. Moreover, this path is highly computable via off-the-shelf convex-optimization routines. An alternative second-order path enforces pointwise second-order conditions, achieving O(initial rate^3) with little additional computational cost.

Extending to higher-order local paths provides an increasingly refined approximation of the ULFP, theoretically approaching single-step TMLE. Simulations show that these refined paths are computationally feasible and yield better finite-sample performance than the basic first-order path.

Heterogeneous Treatment Effects

A systematic comparison of machine learning methods for estimating heterogeneous treatment effects in large-scale randomized trials David Selby* Pei Zhu, Luke Miratrix, Richard Dorsett, David Selby, Polina Polskaia, Nicholas Commins,

Analysts often seek to understand how treatment effects vary across individuals, and machine learning offers a flexible framework for exploring this heterogeneity. Despite various proposed methods, there is limited guidance on which approach to use in experimental evaluations. Our study investigates the robustness of these techniques against diverse impact-generating mechanisms, ranging from simple to complex scenarios, emphasizing the need for methods that perform well across various situations.

Using an empirical Monte Carlo approach, we analyze two US education trials, generating datasets with covariates and untreated outcomes similar to the originals before applying treatment effects to yield outcomes of the treatment group. We evaluate various methods in repeated train-test scenarios to assess their ability to capture impact heterogeneity against various impact-generating processes.

Our findings reveal that, in the absence of impact variation, all estimators show low bias, with regularized methods being more precise. For straightforward heterogeneity, most methods outperform the average treatment effect (ATE), with stable Lasso models typically achieving lower root mean square error (RMSE). With complex heterogeneity, tree-based methods and double machine learning approaches yield lower bias. Generally, methods optimally capture heterogeneity similar to their design, though the Causal Forest is notably strong across contexts.

Instrumental Variables

Discrete Nonparametric Instrumental Variables Aurelien Bibaut* Aurelien Bibaut, Apoorva L Lal, Lars van der Laan, Nathan Kallus,

We consider the discrete Nonparametric Instrumental Variables (NPIV) problem. Inference on identified functionals of solutions to the conditional moment restriction (CMR) is not robust to small perturbations of the identification condition, even if these perturbations vanish asymptotically. In this paper, we remedy this situation by introducing always-defined, that is, not reliant on identification conditions, functionals of pseudo-inverses, inducing an approximating sequence to a linear functional of the structural parameter. In doing so, we relax the very stringent identification condition under a finite number of discrete instruments to a much milder asymptotic identification condition. We derive a novel efficiency theory under the many-weak-instruments asymptotic regime. We propose a novel split-IV estimator for the minimum norm solution to the CMR, inspired by the Jackknife Instrumental Variable Estimator (JIVE), from which we construct a many-weak-instruments-efficient estimator of the functional of the structural parameter. We propose an application to the construction of confounding-robust surrogate indices in the context of experimentation in large online platforms.

Instrumental Variables

A novel multivariable Mendelian randomization framework to disentangle highly correlated exposures with application to metabolomics Lap Sum Chan* Lap Sum Chan, Mykhaylo Malakhov, Wei Pan,

Mendelian randomization (MR) utilizes genome-wide association study (GWAS) summary data to infer causal relationships between exposures and outcomes, offering a valuable tool for identifying disease risk factors. Multivariable MR (MVMR) estimates the direct effects of multiple exposures on an outcome. This study tackles the issue of highly correlated exposures commonly observed in metabolomic data, a situation where existing MVMR methods often face reduced statistical power due to multicollinearity. We propose a robust extension of the MVMR framework that leverages constrained maximum likelihood (cML) and employs a Bayesian approach for identifying independent clusters of exposure signals. Applying our method to the UK Biobank metabolomic data for the largest Alzheimer disease (AD) cohort through a two-sample MR approach, we identified two independent signal clusters for AD: glutamine and lipids, with posterior inclusion probabilities (PIPs) of 95.0% and 81.5%, respectively. Our findings corroborate the hypothesized roles of glutamate and lipids in AD, providing quantitative support for their potential involvement.

Interference and Consistency Violations

Low-rank weighting estimators for causal inference with interference Souhardya Sengupta*
Souhardya Sengupta, Kosuke Imai, Georgia Papadogeorgou,

A primary challenge in observational studies with interference lies in the high dimensionality of the treatment space, arising from the potential for the outcome of a unit to be influenced by treatments applied to other units. In practice, this implies that informative causal inference requires an assumption about the structure of interference patterns. In this paper, we develop a general statistical framework for obtaining causal effect estimators that leverage such a structural assumption. We first show that under an arbitrary interference pattern, the standard inverse probability weighting (IPW) estimator is the only uniformly unbiased weighting estimator. Next, we consider a class of assumptions about interference patterns that can be represented as a low-rank structure of potential outcomes. We then derive an unbiased weighting estimator under such an assumption by minimizing the norm of weights, resulting in a drastic improvement in efficiency over the IPW estimator. When the propensity score is known, the resulting optimal weights are projections of IPW weights onto a subspace that incorporates the low-rank structure. We study the asymptotics of the proposed estimator in the partial interference setting and analyze how misspecifying the low-rank structure affects the proposed estimator, along with a data-driven approach to selecting a low-rank structure among several. Extensive simulations and a real-world application outline the effectiveness of our method.

Machine Learning and Causal Inference

Doubly Robust Estimation of Causal Excursion Effects in Micro-Randomized Trials with Missing Longitudinal Outcomes Jiaxin Yu* Jiaxin Yu, Tianchen Qian,

Micro-randomized trials (MRTs) are increasingly utilized for optimizing mobile health interventions, with the causal excursion effect (CEE) as a central quantity for evaluating interventions under policies that deviate from the experimental policy. However, MRT often contains missing data due to reasons such as missed self-reports or participants not wearing sensors, which can bias CEE estimation. In this paper, we propose a two-stage, doubly robust estimator for CEE in MRTs when longitudinal outcomes are missing at random, accommodating continuous, binary, and count outcomes. Our two-stage approach allows for both parametric and nonparametric modeling options for two nuisance parameters: the missingness model and the outcome regression. We demonstrate that our estimator is doubly robust, achieving consistency and asymptotic normality if either the missingness or the outcome regression model is correctly specified. Simulation studies further validate the estimator's desirable finite-sample performance. We apply the method to HeartSteps, an MRT for developing mobile health interventions that promote physical activity.

Machine Learning and Causal Inference

RieszBoost: Gradient Boosting for Riesz Regression Kaitlyn Lee* Kaitlyn Lee, Alejandro Schuler,

Answering causal questions often involves estimating linear functionals of conditional expectations, such as the average treatment effect or the effect of a longitudinal modified treatment policy. By the Riesz representation theorem, these functionals can be expressed as the expected product of the conditional expectation of the outcome and the Riesz representer, a key component in doubly robust estimation methods. Traditionally, the Riesz representer is estimated indirectly by deriving its explicit analytical form, estimating its components, and substituting these estimates into the known form (e.g., the inverse propensity score). However, deriving or estimating the analytical form can be challenging, and substitution methods are often sensitive to practical positivity violations, leading to higher variance and wider confidence intervals. In this paper, we propose a novel gradient boosting algorithm to directly estimate the Riesz representer without requiring its explicit analytical form. This method is particularly suited for tabular data, offering a flexible, nonparametric, and computationally efficient alternative to existing methods for Riesz regression. Through simulation studies, we demonstrate that our algorithm performs on par with or better than indirect estimation techniques across a range of functionals, providing a user-friendly and robust solution for estimating causal quantities.

Machine Learning and Causal Inference

Differentially Private Two-Stage Empirical Risk Minimization Joowon Lee* Joowon Lee, Guanhua Chen,

We propose a differentially private algorithm for two-stage empirical risk minimization (ERM), designed to balance privacy, utility, and efficiency. In the first stage, we compute data-dependent sample weights to balance covariate distributions between treatment groups, ensuring control of confounding factors. In the second stage, the resulting weighted ERM is solved with objective perturbation to obtain a private optimal model parameter.

Our main application is individualized treatment rules (ITRs) with privacy guarantees. For this application, it is important to use weights that balance covariate distributions between treatment groups to control confounding factors, so the first stage cannot be privatized. This precludes the standard composition method of privatizing multi-stage pipelines for private ITR. Our proposed method requires privatizing only the second stage and uses deterministic perturbation analysis for the first stage. We establish guarantees for efficient differential privacy and utility of our method. Our general framework applies to a wide range of ITR problems with inverse probability weights and distributional covariate balancing weights.

Machine Learning and Causal Inference

Enhancing Causal Inference: Reducing Uncertainty in Causal Forest Models through Cross-Validation Yufan Ji* Yufan Ji, Abdollah Shafieezadeh, Noah Dormady,

In data-driven decision-making, understanding causal relationships is essential. Machine learning, especially causal forest models, has transformed causal inference by estimating the Conditional Average Treatment Effect (CATE), improving personalized strategies across sectors like healthcare, education, and energy. Causal forests, an adaptation of random forests, use the potential outcomes framework to assess heterogeneous treatment effects. However, today's research has primarily focused on accuracy while overlooking the crucial impact of uncertainty on decision making. This is important because large epistemic uncertainties can lead to skewed resource allocation, misinformed prioritization, and suboptimal results in critical areas. This research presents a new method that reduces predictive uncertainty in causal forest models through hyperparameter optimization via Bayesian search and introduces a cross-validation layer to mitigate overfitting. The approach is validated with both synthetic and real data, demonstrating its ability to outperform traditional tools through uncertainty reduction and enhancing decision accuracy. This improvement not only strengthens the reliability of causal estimates but also optimizes decision-making, enabling more effective resource use based on solid causal evidence. These advancements in addressing uncertainty mark a significant contribution to causal inference, with wide-reaching implications for various applications.

Machine Learning and Causal Inference

Robust and Adaptive Causal Null Hypothesis Tests Under Model Uncertainty Junhui Yang*
Junhui Yang, He Bai, Rohit Bhattacharya, Ted Westling,

In observational studies, analysts may posit multiple plausible causal models for a single dataset. As these models typically rely on untestable assumptions, it is often unknown which (if any) of them are correctly specified. Yang et al. [2023] introduced a method of testing a common causal null hypothesis in such settings by combining semiparametric theory with ideas from evidence factors. Their test is asymptotically valid if at least one of the proposed causal models is correct, and exact if exactly one of them is correct. However, the test is conservative when more than one model is valid, resulting in reduced power for small, but non-zero, effects, which may be the case in many real-world applications. In this talk, we propose a test that adapts to the number of correctly specified causal models. We do this by proposing a consistent estimator of the number of identified functionals that are zero under the null and adapting the downstream test accordingly. Under regularity conditions, our test remains asymptotically exact even when more than one causal model is correctly specified. Our adaptive test demonstrates significantly improved power over existing methods. In simulations, its performance is similar to an oracle that already knows the number of correctly specified models. Our adaptive procedures thus enhance power while preserving robustness to causal model misspecification, offering a flexible and practical solution for hypothesis testing under model uncertainty.

Machine Learning and Causal Inference

Automatic doubly robust inference for linear functionals via calibrated debiased machine learning Lars van der Laan* Lars van der Laan, Alex Luedtke, Marco Carone,

In causal inference, many estimands of interest can be expressed as a linear functional of the outcome regression function; this includes, for example, average causal effects of static, dynamic and stochastic interventions. For learning such estimands, in this work, we propose novel debiased machine learning estimators that are doubly robust asymptotically linear, thus providing not only doubly robust consistency but also facilitating doubly robust inference (e.g., confidence intervals and hypothesis tests). To do so, we first establish a key link between calibration, a machine learning technique typically used in prediction and classification tasks, and the conditions needed to achieve doubly robust asymptotic linearity. We then introduce calibrated debiased machine learning (C-DML), a unified framework for doubly robust inference, and propose a specific C-DML estimator that integrates cross-fitting, isotonic calibration, and debiased machine learning estimation. A C-DML estimator maintains asymptotic linearity when either the outcome regression or the Riesz representer of the linear functional is estimated sufficiently well, allowing the other to be estimated at arbitrarily slow rates or even inconsistently. We propose a simple bootstrap-assisted approach for constructing doubly robust confidence intervals. Our theoretical and empirical results support the use of C-DML to mitigate bias arising from the inconsistent or slow estimation of nuisance functions.

Machine Learning and Causal Inference

Automatic Debiased Machine Learning for Smooth Functionals of Nonparametric M-Estimands Lars van der Laan* Lars van der Laan, Aurelien Bibaut, Nathan Kallus, Alex Luedkte, Lars van der Laan,

We propose a unified framework for automatic debiased machine learning (autoDML) to perform inference on smooth functionals of infinite-dimensional M-estimands, defined as population risk minimizers over Hilbert spaces. By automating debiased estimation and inference procedures in causal inference and semiparametric statistics, our framework enables practitioners to construct valid estimators for complex parameters without requiring specialized expertise. The framework supports a generic models parameterized by Hilbert spaces, Neyman-orthogonal loss functions with unknown nuisance components, and nonlinear functionals of multiple M-estimands. We formalize the class of parameters efficiently estimable by autoDML as a novel class of nonparametric projection parameters, defined via orthogonal minimum loss objectives. We introduce three autoDML estimators based on one-step estimation, targeted minimum loss-based estimation, and the method of sieves. For data-driven model selection, we derive a novel decomposition of model approximation error for smooth functionals of M-estimands and propose adaptive debiased machine learning estimators that are superefficient and adaptive to the functional form of the M-estimand. Finally, we illustrate the flexibility of our framework by constructing autoDML estimators for the long-term survival under a beta-geometric model and the average treatment effect under a heterogeneous treatment effect model.

Machine Learning and Causal Inference

Anytime-Valid Inference for Double/Debiased Machine Learning of Causal Parameters
Abhinandan Dalal* Abhinandan Dalal, Patrick Blöbaum, Shiva Kasiviswanathan, Aaditya Ramdas,

Double (debiased) machine learning (DML) is widely used for learning causal/structural parameters due to its flexibility with high-dimensional nuisance functions and its ability to avoid bias from regularization or overfitting. However, the classic double-debiased framework is valid only for a fixed sample size, limiting its ability to either collect more data for sharper inference or stop early when stable estimates arise. This poses concerns in large-scale experiments with high costs or life-or-death decisions, and in observational studies where confidence intervals may fail to shrink even with more data because of partial identifiability.

We propose time-uniform counterparts to asymptotic DML, allowing valid inference and confidence intervals at any (possibly data-dependent) stopping time. Our conditions are only slightly stronger than standard DML requirements but guarantee anytime-valid inference. These results let any existing DML procedure become anytime-valid with minimal changes, making it highly adaptable. We demonstrate this with two examples: (a) local average treatment effect in online experiments with non-compliance, and (b) partial identification of average treatment effect in observational studies with unmeasured confounding.

Machine Learning and Causal Inference

Average Causal Effect Estimation and Efficiency Gains via Verma Constraints in the Causal Napkin Graph Anna Guo* Anna Guo, Razieh Nabi, David Benkeser,

In causal inference, certain patterns of unmeasured confounding between treatment and outcome can invalidate conventional methods like the g-formula, or equivalently, back-door and front-door functionals. This work explores the “Napkin graph,” a causal structure that combines key features of M-bias, instrumental variables, and back-door/front-door criteria. By leveraging pre-treatment “trapdoor” variables, which influence the outcome exclusively through the treatment, nonparametric identification of the average causal effect is achieved via a ratio of two g-formulas, addressing limitations of traditional methods.

In this work, we propose novel estimators for the Napkin functional, including doubly robust one-step and targeted minimum loss-based estimators that achieve asymptotic linearity under flexible convergence rates for nuisance parameter estimation. A central innovation of our work is the principled utilization of Verma constraints—generalized independence restrictions between observable variables in graphical models with unmeasured variables. Specifically, we leverage a Verma constraint between the trapdoor variable and the outcome to achieve significant efficiency gains, advancing semiparametric causal effect estimation in such models.

The methods are validated through simulations and real data analysis using an HR dataset to estimate the causal effect of employee skill on performance. We also introduce the texttt{napkintmle} R package for easy implementation.

Machine Learning and Causal Inference

Causal estimands and estimators for evaluating the impact of preventive interventions on long term outcomes Allison Codi* David Benkeser, Razieh Nabi, Elizabeth Rogawski-McQuade, Allison Codi, Mats Stensrud,

Establishing the long-term effects of interventions aimed at preventing intermediate outcomes poses significant challenges. For example, vaccines designed to prevent diarrhea caused by Shigella bacteria in children may also positively impact long-term growth, as Shigella-induced diarrhea is a known cause of growth faltering. However, given the relatively low frequency of Shigella-related diarrhea, the vaccine's marginal causal effect on growth may be too small to detect in a typical randomized controlled trial. Nevertheless, clinicians and policymakers are highly interested in demonstrating the broader benefits of vaccination on growth outcomes.

To address this challenge, we propose alternative causal estimands that enjoy improved power for detecting effects on long-term outcomes in realistic trial settings. Specifically, we introduce estimands based on principal stratification and interventional causal frameworks and demonstrate that both approaches yield the same identifying functional under different assumptions. Notably, the principal stratification approach relies on cross-world independence assumptions, whereas the interventional estimand does not.

We further derive nonparametric efficient, and doubly robust estimators for these estimands, leveraging machine learning techniques for nuisance parameter estimation. Through realistic simulations, we illustrate that these estimators can provide clinically meaningful inferences even within the constraints of practical Shi

Machine Learning and Causal Inference

Doubly robust conformal prediction for missing data Manit Paul* Manit Paul, Arun Kumar Kuchibhotla, Eric J. Tchetgen Tchetgen,

Conformal Prediction (CP) has seen growing attention in recent years, providing new tools for tackling missing data problems. However most of these applications of CP lack robustness as they remain largely disconnected from modern semi-parametric efficiency theory. In this paper we consider the general problem of obtaining distribution-free valid prediction regions for the outcome based on a coarsened version of the complete data. We do this by deriving the efficient influence function of the quantile of the outcome under a given semi-parametric model and then performing a conformal risk control procedure. We employ modern non-parametric methods (random forests etc.) to learn the underlying nuisance functions of the semi-parametric model. This general theory has several consequences — (i) Covariate-shift problem: provides the required coverage guarantee (without any O(root-n) slack) if at-least one of the nuisance functions (propensity score and conditional distribution of the outcome) is estimated exactly — an improvement over the earlier work by Yang et al. [2022] (ii) Monotone missingness: provides multiply robust prediction set for the outcome under the Missing at Random (MAR) assumption — this to our knowledge is one of the first results of this kind. Our theory also enables the construction of robust prediction regions for non-monotone missing data under MAR assumption. We further illustrate the performance of our methods through various simulation studies.

Machine Learning and Causal Inference

Optimal Design-based Combination of Expert and Imperfect Annotations for Robust Causal Inference from Text Data Angela Zhou* Angela Zhou, Ezinne Nwankwo,

Decision-makers often have rich unstructured text data with additional contextual information on events studied in the framework of causal inference, for example, clinical notes in medicine. Human experts can extract relevant information from rich unstructured text data, but the volume of the data is too overwhelming for complete annotation. NLP tools are available and can annotate at scale, but without the same accuracy. Motivated by our ongoing work with a nonprofit that conducts street outreach with homelessness clients, we develop methods for design-based combination of scarce expert labels and imperfect NLP-based annotation tools (such as imperfect but cheaper LLM-based annotation). The nonprofit has hundreds of thousands of notes: which ones should be labeled by experts, vs. by an imperfect annotator? We model the presence of expert human labels as a sample selection or missing outcome problem. We optimize the asymptotic variance based on pilot estimates for a two-stage optimized design-based procedure to derive optimal allocations. We study the case of text-based outcomes or text-based treatments in a framework of missing outcomes or treatments. We show benefits of the methods in synthetic data, an IMDB dataset of movie reviews, and on client casenotes in the case of street outreach for homelessness.

Machine Learning and Causal Inference

A Model Ensemble Approach to Individual Fairness in Machine Learning Bernardo Modenesi* Bernardo Modenesi, Lucia Wang, Ameya Diwan,

Individual fairness is based on the principle that similar observations should be treated similarly by a machine learning (ML) model, addressing the limitations of group fairness methods. Despite its intuitive appeal, implementing individual fairness algorithms is challenging due to difficulties in defining a metric for similarity between individuals. In this paper, we develop a model ensemble approach inspired by individual fairness to assess ML model fairness. Leveraging results from the double/causal ML literature and ML clustering techniques, our method requires considerably fewer assumptions than previous individual fairness methods, in addition to being model-agnostic and avoiding cherry-picking decisions in fairness assessment. Our data-driven method involves: (i) removing variation in the dataset related to sensitive attributes using causal ML; (ii) clustering observations using random forests and a Bayesian network algorithm; (iii) performing within-cluster inference to test if the model treats similar observations similarly, and applying multiple hypothesis test correction to aggregate the results. We provide a single statistical p-value for the null hypothesis that the model is unbiased based on individual fairness and create a scale to measure the extent of bias against minorities, enhancing the interpretability of the p-value for decision-makers. We apply our methodology to assess bias in several contexts and provide a Python package for this methodology.

Machine Learning and Causal Inference

A Gaussian Process Framework for Survey-Based Event Studies Soonhong Cho* Soonhong Cho,

Survey-based event studies exploit salient events that occur during the study period to estimate causal effects by comparing respondents interviewed before and after the event. This design faces methodological challenges mainly from (i) demographic differences between pre- and post-event respondent groups and (ii) time series complexities including trends, seasonality, and autocorrelation. Existing approaches use reweighting methods to address demographic imbalances, but these depend on researcher decisions about model specification, covariate binning, and balance metrics—choices that can substantially affect results. We propose a Gaussian Process (GP) regression framework that estimates counterfactuals by learning outcome-covariate relationships from pre-event data and projecting them to post-event periods. We simplify the GP hyperparameter structure to enable automated estimation with minimal user discretion. The framework flexibly learns time-varying relationships, provides principled uncertainty quantification, and automatically handles practical challenges like irregular response timing and missing data. Through simulation studies and political science applications, we demonstrate that our approach provides robust causal estimates while requiring fewer modeling assumptions than existing approaches.

Matching, Weighting

Causal Interaction and Effect Modification: A Randomization-Based Approach to Inference
Zion Lee* Zion Lee, Kwonsang Lee,

Understanding causal interactions is critical in observational studies but often challenging due to confounding and methodological limitations. While these concepts have been extensively studied in randomized experiments, their application in observational data remains limited, particularly in settings requiring stratification to evaluate interactions. We propose a novel randomization-based inference framework utilizing matching methods to investigate causal interactions. Additionally, we provide a comprehensive review of causal interaction, explaining its unique focus on joint causal effects and distinguishing it from statistical interaction and effect modification. Using a real-world dataset, we analyze the joint effects of two treatments: residential fire safety equipment and fire response time. Our approach demonstrates how matching can mitigate confounding and identify interaction effects, offering a robust alternative to traditional methods. The findings highlight the importance of considering causal interaction in public safety interventions and provide actionable insights for fire safety policy and fire response optimization.

Matching, Weighting

Matching Methods for Difference-in-Differences with Multiple Time Periods: Evaluating the Equality of ATT Estimates Across Time Junho Jang* Junho Jang, Yitae Kwon, Kwonsang Lee,

In observational studies with multiple time points, testing for homogeneous causal effects, such as the Average Treatment effect on the Treated (ATT), is crucial for evaluating treatment efficacy over time. This paper introduces a novel testing framework that accommodates arbitrary combinations of treatment initiation and post-treatment time points. For estimation, difference-in-differences (DID) is combined with matching to focus on post-treatment periods for treated units. Our testing framework involves two main steps. First, a confidence set for the common treatment effect is constructed to narrow the range of plausible parameters. Second, a randomization-based test is conducted within this confidence set to assess the equality of treatment effects. This approach extends multivariate location testing to partially matched sets. Furthermore, the relationship between matched set structure and test power is theoretically explored, providing insights to guide matching design in practice. To illustrate its application, we use this framework on Health and Retirement Study (HRS) data, testing and summarizing treatment equality across time periods.

Matching, Weighting

Time-Series Matching: Analyzing the Causal Impact of Netflix Subscription on IPTV Viewing Behavior Yongho Yoon* Yongho Yoon, Dahai Jung, Kwonsang Lee,

Streaming services like Netflix have reshaped media consumption patterns, posing challenges and opportunities for traditional IPTV platforms such as Comcast. This study investigates the causal impact of Netflix subscriptions on IPTV viewing behavior, leveraging real-world, proprietary data not commonly accessible for analysis. A key focus is the dynamic nature of Netflix subscriptions, which fluctuate due to seasonal trends and the release of popular content such as Squid Game. To address these complexities, we extend time-series matching methodologies, capturing nuanced temporal trends that are often overlooked by existing approaches. Our analysis explores several critical questions: Does subscribing to Netflix reduce IPTV views, or does it establish an independent, non-interfering viewing pattern? Does long-term Netflix subscription accelerate the decline in IPTV usage, and does IPTV viewership rebound after subscription cancellations? Additionally, we evaluate whether viral content creates lasting effects on IPTV usage or only transient spikes. To answer these questions, we utilize a range of matching techniques, enabling a robust assessment of both short-term and long-term effects. By bridging gaps in current time-series analysis methods, this research provides actionable insights into the interplay between streaming services and traditional IPTV platforms.

Mediation Analysis, Mechanisms

Multiply Robust Estimation with Machine Learning for Causal Mediation Analysis with Clustered Data and Unmeasured Cluster-Level Confounders Cameron McCann* Cameron McCann, Xiao Liu,

Mediation analyses in behavioral research often involve clustered data; however, existing methods for causal mediation in this context are relatively limited. In this study, we extend a multiply robust estimation method to estimate causal mediation effects in clustered data while accounting for unmeasured pre-treatment confounders at the cluster level. To control for unmeasured cluster-level confounders in the nuisance model estimation, we perform cluster-mean centering and include cluster means and cluster dummies; for comparison, we also examined the performance when excluding the cluster dummies. For statistical inference, we consider within-cluster correlation in calculating standard errors and confidence intervals. Through simulations, we assess the performance for inference of the cluster-average and individual-average causal mediation effects, comparing nuisance model estimation with (1) parametric models (fixed or random-effects regressions) versus (2) machine learning prediction models (a super learner ensemble of parametric and nonparametric regressions). Finally, we illustrate the method using data from the National Longitudinal Study of Adolescent to Adult Health (Add Health; Harris & Udry, 2008).

Mediation Analysis, Mechanisms

Post-treatment problems: What can we say about the effect of a treatment among sub-groups who (would) respond in some way? Tanvi Shinkre* Tanvi Shinkre, Chad Hazlett, Nina McMurry,

Investigators are often interested in how a treatment affects an outcome for units responding to treatment in a certain way. We may wish to know the effect among units that, for example, meaningfully implemented the intervention, passed an attention check, or survived to the endpoint. Simply conditioning on the observed value of the relevant post-treatment variable introduces problematic biases. Further, assumptions such as “no unobserved confounding” (of the post-treatment mediator and the outcome) or of “no direct effect” (of treatment on outcome) required of several existing strategies are typically indefensible. We propose the Treatment Reactive Average Causal Effect (TRACE), which we define as the total effect of the treatment in the group that, if treated, would realize a particular value of the relevant post-treatment variable. Given the total effect of treatment, and by reasoning about the treatment effect among the “non-reactive” group, we can identify and estimate the range of plausible values for the TRACE. We discuss this approach and its connection to existing estimands and identification strategies, then demonstrate its use with two applications: (i) a community-policing intervention in Liberia, among locations where the project was meaningfully implemented, and (ii) a field experiment studying how in-person canvassing affects support for transgender rights, among participants whose feelings towards transgender people become more positive.

Mediation Analysis, Mechanisms

Monotonic Path-Specific Effects: Application to Estimating Educational Returns Aleksei Opacic* Aleksei Opacic,

Conventional research on educational effects typically either employs a “years of schooling” measure of education, or dichotomizes attainment as a point-in-time treatment. Yet, such a conceptualization of education is misaligned with the sequential process by which individuals make educational transitions. In this paper, I propose a causal mediation framework for the study of educational effects on outcomes such as earnings. The framework considers the effect of a given educational transition as operating indirectly, via progression through subsequent transitions, as well as directly, net of these transitions. I demonstrate that the average treatment effect (ATE) of education can be additively decomposed into mutually exclusive components that capture these direct and indirect effects. The decomposition has several special properties which distinguish it from conventional mediation decompositions of the ATE, properties which facilitate less restrictive identification assumptions as well as identification of all causal paths in the decomposition. An analysis of the returns to high school completion in the NLSY97 cohort suggests that the payoff to a high school degree stems overwhelmingly from its direct labor market returns. Mediation via college attendance, completion and graduate school attendance is small because of individuals’ low counterfactual progression rates through these subsequent transitions.

Multilevel Causal Inference

Targeted Quality Measurement of Health Care Providers Yige Li* Yige Li, José Zubizarreta, Nancy Keating, Mary Beth Landrum,

Assessing the quality of cancer care across US healthcare providers poses significant challenges, mainly due to small practice sizes, large number of practices, and diversity of patient populations. Patients vary widely in cancer type and other critical factors, making comparisons across practices complex. In this paper, we propose an approach to adjust for such patient diversity. Our framework follows recent advancements in health outcomes research, framing quality measurement as a causal inference problem. Using covariate profiles to describe patient populations, our approach combines a weighting step and a regression step, accounting for heterogeneous effects when practice-covariate interactions are present in the data. Through extensive simulations, we compare the performance of several methods in terms of point estimates and rankings of hundreds of practices. The results show our approach produces stable and robust estimates, as well as reliable rankings. For the case study, we provide the results of 600 practices for a couple of patient covariate profiles. The proposed approach is helpful for public reporting and has the potential to help individual patient make decisions.

Multilevel Causal Inference

Conformal causal inference for cluster randomized trials Bingkai Wang* Bingkai Wang, Fan Li, Mengxin Yu,

Traditional statistical inference in cluster randomized trials typically invokes the asymptotic theory that requires the number of clusters to approach infinity. In this article, we propose an alternative conformal causal inference framework for analyzing cluster randomized trials that achieves the target inferential goal in finite samples without the need for asymptotic approximations. Different from traditional inference focusing on estimating the average treatment effect, our conformal causal inference aims to provide prediction intervals for the difference of counterfactual outcomes, thereby providing a new decision-making tool for clusters and individuals in the same target population. We prove that this framework is compatible with arbitrary working outcome models—including data-adaptive machine learning methods that maximally leverage information from baseline covariates, and enjoys robustness against misspecification of working outcome models. Under our conformal causal inference framework, we develop efficient computation algorithms to construct prediction intervals for treatment effects at both the cluster and individual levels, and further extend to address inferential targets defined based on pre-specified covariate subgroups. Finally, we demonstrate the properties of our methods via simulations and a real data application based on a completed cluster randomized trial for treating chronic pain.

Policy Learning

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

In many applications, the goal is to assign treatments based on unit features, leading to personalized treatment policies. This often involves optimizing an objective function with counterfactual quantities such as the conditional ATE (CATE). However, in most real-world settings, some unit features may be measured with error and overlooking these errors can introduce systematic bias. In this work, we consider such settings with discrete unobserved features. After establishing the non-identifiability of the CATE, we propose two novel frameworks for treatment policy design using partial identification techniques, focusing on the measurement mechanism—the conditional distribution of measurements given the unobserved features. The first framework requires mild invertibility of the measurement mechanism, leading to a conservative treatment policy design based on ideas from proximal causal inference. The second framework further incorporates modeling assumptions on the measurement mechanism. We demonstrate that sharp bounds on the CATE can be obtained in this framework. This approach also leads to a novel characterization for informative measurements. We provide computationally efficient methods for obtaining bounds and quantifying uncertainty. Additionally, we show that integrating natural treatment values can further improve the bounding methods in treatment policy designs. We evaluate our methods through simulations and compare with naive methods that ignore measurement errors.

Policy Learning

Optimal Policy Learning Under Spatial Dependence With Applications to Groundwater in Wisconsin
Xindi Lin* Xindi Lin, Hyunseung Kang, Christopher Zahasky,

When installing drinking water wells, it’s well-understood that increasing well depth improves the quality of the groundwater, but also raises costs. Policymakers must therefore determine the minimum well depth needed to meet the public health standards for contaminants in groundwater, such as nitrates, a popular contaminant from fertilizers. In Wisconsin, the current approach to setting the minimum well depth is often a single, static number, which ignores the local hydrogeological characteristics. In this paper, we propose a data-driven method for estimating the Spatial Minimum Resource Threshold Policy (spMRTP), which determines the minimum treatment level needed at each location to meet the target outcome. A key feature of spMRTP is to account for spatial dependence of contaminants where high contaminants levels in one area often imply high contaminant levels in adjacent areas. We estimate spMRTP by empirical risk minimization with a novel, nonparametric, doubly robust loss function. For computation, we propose to use the Vecchia approximation to efficiently evaluate the minimizer. Our simulation results demonstrate that the proposed method outperforms competing approaches, including non-spatial methods for policy learning and indirect estimation methods. We also apply our method to water quality data collected from 2014 to 2024 in Wisconsin and generate a spatial map of optimal, minimum well depths in Wisconsin to meet the 10-ppm public health standard for nitrates.

Positivity violations

Fair comparisons of causal parameters with many treatments and positivity violations
Wenbo Wu* Alec McClean, Yiting Li, Sunjae Bae, Mara McAdams-DeMarco, Ivan Diaz,

Comparing outcomes across treatments is essential in causal inference. Researchers typically estimate a set of parameters, possibly counterfactual, with each targeting a different treatment. Treatment-specific means (TSMs) are commonly used, but identification requires a positivity assumption—that all subjects have non-zero probability of receiving each treatment—which is often implausible, especially with many treatment values. Parameters based on dynamic stochastic interventions offer robustness to positivity violations, but comparisons between the parameters may be unfair, because they can depend on outcomes under non-target treatments. To clarify when a fair comparison between two parameters targeting different treatments is possible, we propose a fairness criterion: if the conditional TSM for one treatment is greater than that for another, then the corresponding causal parameter is greater. We derive two intuitive properties equivalent to this criterion and show that only a mild positivity assumption is needed to identify fair parameters. We provide parameters that satisfy this criterion and are identifiable under the milder positivity assumption. Their non-smoothness makes standard efficiency theory inapplicable, so we propose smooth approximations of them. We then develop doubly robust-style estimators that attain parametric convergence rates under nonparametric conditions. We illustrate our methods with an analysis of dialysis providers in New York State.

Principal stratification

Semiparametric principal stratification analysis without monotonicity Jiaqi Tong* Jiaqi Tong,

Intercurrent events, common in clinical trials and observational studies, affect the existence or interpretation of final outcomes. Principal stratification addresses these challenges by defining average treatment effects within subpopulations using counterfactual intermediate outcomes as pre-treatment covariates. However, most methods rely on strong assumptions, such as monotonicity and counterfactual intermediate independence. To relax these assumptions, we consider a margin-free, and variation-independent framework for principal stratification analysis based on a conditional odds ratio sensitivity parameter. Under principal ignorability, we derive non-parametric identification formulas for principal causal effects and propose weighting and regression approaches for estimation. We further derive the efficient influence function to construct a conditionally doubly robust estimator, as well as a de-biased machine learning estimator. We use extensive simulations to illustrate the consequence of incorrectly assuming monotonicity, and the implications of misspecifying the sensitivity parameter under non-monotonicity. We apply our methods to two critical care clinical trials comparing active treatments where monotonicity is deemed questionable.

Proximal Causal Learning

Extending Proxy Methods for Causal Identification: Comparing Bridge Equations and Completeness Conditions with Eigendecomposition Approaches Helen Guo* Helen Guo, Elizabeth Ogburn, Ilya Shpitser,

Identifying causal effects in the presence of unmeasured variables is a fundamental challenge in causal inference, for which proxy variable methods have emerged as a powerful solution. This work contrasts two prominent approaches within this framework: (1) methods that use bridge equations and completeness conditions to recover identifying representations and (2) eigendecomposition approaches that identify intermediate target distributions comprising the identifying representation, up to permutation of latent state labels. The former has been developed for settings involving unmeasured confounding or mediation (Miao et al., 2018; Cui et. al., 2023; Ghassami et al., 2024). The latter approach (Kuroki & Pearl, 2014) – which may be viewed as a special case of Kruskal's uniqueness condition for the
Candecomp/Parafac decomposition (Kruskal, 1977; Stegeman & Sidiropoulos, 2007) – has been recently expanded to handle circumstances with unmeasured treatment (Zhou & Tchetgen Tchetgen, 2024). Comparing the model restrictions imposed by each approach for hidden confounding or mediation, we delineate scenarios where the models are equivalent and derive conditions for identifying the full joint distribution of the underlying causal graph. Furthermore, we extend proximal methods to simultaneously address unmeasured confounding and mediation together, and discuss assumptions under which identification is possible via different identifying representations.

Randomized Designs and Analyses

Rerandomization with Missing Data Kateryna Husar* Kateryna Husar, ,

Randomized control trials are considered the gold standard in research as they allow for high confidence in establishing cause-and-effect relationships. Randomly assigning participants to the treatment or control group ensures that any observed differences in outcomes between these groups can be attributed to the intervention rather than external factors. Yet, differences between the groups can still occur due to chance, potentially resulting in misleading results. The issue of observed covariate imbalance can be addressed in the design phase: rerandomization selects a treatment assignment from a subset of assignments that satisfy a predetermined balance criterion for pre-treatment covariates. Under rerandomization, classical estimators yield a more precise estimator and combining rerandomization with the regression adjustment can further improve inference. In practice, even in the pre-treatment stage, there may be substantial missing data, which in turn can reduce the improvements due to rerandomization and cannot be addressed by simple post-hoc regression adjustment. By introducing missing data imputation methods into the rerandomization, we recover the efficiency losses for estimating average treatment effects. We show how rerandomization that adjusts for missingness combined with regression adjustment increases the precision of the estimates compared to regression adjustment alone and recommend the use of rerandomization in the study design when missing data are present.

Randomized Designs and Analyses

Beyond Fixed Restriction Time: Adaptive RMST Methods for Non-Proportional Hazards in Clinical Trials Jinghao Sun* Jinghao Sun, Eric Tchetgen Tchetgen, Douglas Schaubel,

Restricted mean survival time (RMST) offers a compelling nonparametric alternative to hazard ratios for right-censored time-to-event data, particularly when the proportional hazards assumption is violated. By capturing the total event-free time over a specified horizon, RMST provides an intuitive and clinically meaningful measure of absolute treatment benefit. Nonetheless, selecting the restriction time L poses challenges: choosing a small L can overlook late-emerging benefits, whereas a large L may inflate variance and undermine power. We propose a novel data-driven, adaptive procedure that identifies the optimal restriction time L* from a continuous range by maximizing a criterion balancing effect size and estimation precision. Consequently, our procedure is particularly useful when the pattern of the treatment effect is unknown at the design stage. We provide a rigorous theoretical foundation that accounts for variability introduced by adaptively choosing L*. To address nonregular estimation under the null, we develop two complementary strategies: a convex-hull-based estimator, and a penalized approach that further enhances power. When restriction time candidates are defined on a discrete grid, our procedure surprisingly incurs no asymptotic penalty for selection, thus achieving near-oracle performance. In a phase III pancreatic cancer trial with transient treatment effects, our procedure uncovers clinically meaningful benefits that standard methods overlook.

Randomized Designs and Analyses

Simulation-Based Inference After Adaptive Experiments Aurelien Bibaut* Aurelien Bibaut, Brian Cho, Nathan Kallus,

In recent years, much work has been done on inference after adaptive data collection. Most of these provide methods that enforce normality of the asymptotic distribution. As noted elsewhere, control of the distribution of test statistics in adaptive experiments comes at the expense of power. In this article, we overcome the need to control the asymptotic distribution of statistics by simulating the distribution of the statistic under candidate values of the parameter of interest. Our sole focus in constructing test statistics is then to maximize power. Unlike some existing works, our framework allows for arbitrary adaptive experimental designs, and in particular does not require non-zero propensities, thereby allowing for UCB and linUCB adaptive experimental designs. We provide generic conditions for inference validity for scalar parameters in semiparametric models. Numerical experiments demonstrate drastic improvements in power and confidence interval size over all existing baselines.

Randomized Designs and Analyses

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

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 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 on sufficient condition regions to quantify individual level 'always survivor' causal effects under truncation by death. Such urn models also enable sensitivity and multiverse analysis at the individual and population level, as well as enable comparison of different identification strategies. We illustrate the practical utility of our methods using data from randomized clinical trials in oncology and laser surgery in perinatal studies. 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

Application of the Robustness of Inference to Replacement (RIR) to Differential Attrition in an RCT Kenneth Frank* Kenneth Frank,

Differential attrition is one of the most serious sources of bias in estimates of treatment effects in randomized experiments (e.g., Hewitt et al., 2010; WWC Standards Handbook 4.1, 2020). In theory, if even only one case drops out of either the treatment or control for a systematic, non-random reason, the principle of randomization as a basis for estimation and inference is compromised. While loss of one case is unlikely to overturn results in most instances (unless its outcome would have been extreme), the question generally concerns how robust an inference from an RCT is to differential, non-random attrition from treatment and control. While there are many techniques for imputing the attritted data, ultimately there will be some aspects of the attritted data that are unobservable. The purpose of this paper is to characterize the unobserved conditions in the attritted data such that, if combined with observed data, it would nullify an inference of an effect of the predictor of interest (X) on the outcome (Y). We do so for a non-parametric approach based on differences in means on the outcome in the attritted data and then for a parametric approach based on the correlation between treatment and outcome in the attritted data that uses statistical significance as a threshold.

Sensitivity Analysis

Model Selection for Causal Inference with Generalized Information Criteria Yuchen Xiao* Yuchen Xiao, Stephen Walker,

Model selection plays a critical role in ensuring reliable and reproducible statistical inference in linear regression. The use of Generalized Information Criteria (GIC) to construct regression estimators corresponds to the OLS estimator under (ell_0) constraint. One key advantage of using information criteria lies in its ability to simultaneously integrate model estimation and model evaluation, which strikes a balanced parsimony between predictive performance and model complexity. Although traditionally viewed as computationally infeasible in high-dimensional settings due to its NP-hard nature, we demonstrate that applying information criteria for model selection and estimation can be efficiently solved in polynomial time using Hopfield network optimization. This rejuvenation of information criteria for model selection in high dimensions is applied to select covariates for inclusion into the propensity score to reduce the bias and improve the statistical efficiency of propensity score estimator. Extensive simulation results are presented to validate the efficiency and effectiveness of the proposed approach with the objective of rejuvenating the application of GIC for modeling the propensity score model and the outcome model. The simulation results show that the GIC includes all true confounders and predictors of outcome while maintaining minimum false positives.

Sensitivity Analysis

Sensitivity of weighted least squares estimators to omitted variables Leonard Wainstein*
Leonard Wainstein, Chad Hazlett,

This paper introduces tools for assessing the sensitivity, to unobserved confounding, of a common estimator of the causal effect of a treatment on an outcome that employs weights: the weighted linear regression of the outcome on the treatment and observed covariates. We demonstrate through the omitted variable bias framework that the bias of this estimator is a function of two intuitive sensitivity parameters: (i) the proportion of weighted variance in the treatment that unobserved confounding explains given the covariates and (ii) the proportion of weighted variance in the outcome that unobserved confounding explains given the covariates and the treatment, i.e., two weighted partial R2 values. Following previous work, we define sensitivity statistics that lend themselves well to routine reporting, and derive formal bounds on the strength of the unobserved confounding with (a multiple of) the strength of select dimensions of the covariates, which help the user determine if unobserved confounding that would alter one's conclusions is plausible. We also develop tools for adjusted inference. The key contribution of these tools is that they apply with any non-negative weights (e.g., inverse-propensity score, matching, or covariate balancing weights). The proposed tools also refrain from distributional assumptions on the data or unobserved confounding, and can address bias from misspecification in the observed data.

Sensitivity Analysis

Multivariate one-sided testing via sample splitting in an observational study of the effect of poverty on children's physical fitness William Bekerman* William Bekerman, Dylan Small, Colin Fogarty,

When assessing the causal effect of a treatment on two or more outcomes in an observational study, a linear combination of outcomes may lessen the sensitivity of a test of the global null hypothesis to potential unmeasured biases. While all linear combinations of scored outcomes can be considered using Scheffé projections, finding the contrast that minimizes sensitivity to unmeasured biases requires corrections for multiple testing which can erode power, especially when many outcomes are of interest. To mitigate this issue, we propose splitting the sample into a planning sample to identify the optimal contrast and an analysis sample to conduct inference. We introduce a novel minimax theorem for this problem and find that the design sensitivity on the whole sample equals the design sensitivity when using split samples. We also conduct extensive simulation studies demonstrating enhanced power in finite samples. Finally, we apply our method to investigate the broad effects of low family income on children's physical activity and fitness.

Sensitivity Analysis

Robust Causal Inferences from the Sensitivity Analysis of Multiple Estimands Nathan Cheng* Nathan Cheng, Jose Zubizarreta,

When the assumption of unconfoundedness is suspect, one may reason about the robustness of a causal finding via a sensitivity analaysis. Traditional techniques in sensitivity analysis are typically designed to reason about a single estimand at a time. However, when the investigator is interested in a suite of hypotheses—answered by way of multiple estimands—and wishes to control some joint error rate, dependencies among the estimands can be leveraged to yield more precise assessments of sensitivity. In this work, we introduce an approach for the sensitivity analysis of multiple estimands in a general setting where weighting estimators are used to estimate causal quantities such as the average treatment effect (ATE). We show that useful sensitivity regions—high-dimensional sets that contain the unidentifiable ATE with high probability—can be tractably computed, and useful inferences can be extracted from them. Using simulations, we demonstrate the benefits of leveraging estimand dependence compared to naively combining one-at-a-time sensitivity analyses. Lastly, we describe some creative applications of our approach, and apply our method to real data for illustration.

Sensitivity Analysis

A Sensitivity Analysis for the Average Derivative Effect Jeffrey Zhang* Jeffrey Zhang,

In observational studies, sensitivity analysis is an important tool that can help determine the robustness of a causal conclusion to a certain level of unmeasured confounding. At the same time, exposures that arise in observational studies are often continuous rather than binary or discrete. Sensitivity analysis approaches for continuous exposures have now been proposed for several causal estimands. In this article, we focus on the average derivative effect, a classical estimand from the econometrics literature. We obtain closed-form bounds for the average derivative effect under a sensitivity model that constrains the odds ratio (at any two dose levels) of the generalized propensity score. We propose flexible, efficient estimators for the bounds, as well as point-wise and uniform confidence intervals. We examine the finite sample performance of the methods through simulations and illustrate the methods on a study assessing the effect of parental income on educational attainment.

Sensitivity Analysis

Nonparametric Sensitivity Analysis for Unobserved Confounding with Survival Outcomes * Rui Hu, Ted Westling, Rui Hu,

In observational studies, the observed association between an exposure and outcome of interest may be distorted by unobserved confounding. Causal sensitivity analysis is often used to assess the robustness of observed associations to potential unobserved confounding. For time-to-event outcomes, existing sensitivity analysis methods rely on parametric assumptions on the structure of the unobserved confounders and Cox proportional hazards models for the outcome regression. If these assumptions fail to hold, it is unclear whether the conclusions of the sensitivity analysis remain valid. Additionally, causal interpretation of the hazard ratio is challenging. To address these limitations, in this paper we develop a nonparametric sensitivity analysis framework for time-to-event data. Specifically, we derive nonparametric bounds for the difference between the observed and counterfactual survival curves and propose estimators and inference for these bounds using semiparametric efficiency theory. We also provide nonparametric bounds and inference for the difference between the observed and counterfactual restricted mean survival times. We demonstrate the performance of our proposed methods using numerical studies and an analysis of the causal effect of physical activity on respiratory disease mortality among former smokers.

Sensitivity Analysis

Applying Robustness of an Inference to Replacement (RIR) as the Sensitivity Test of Difference-in-Differences Estimator Xuesen Cheng* Xuesen Cheng,

The validity of the Difference-in-Differences (DID) estimator is affected by (1) violations of the parallel trend assumption, (2) other sources of bias in OLS estimation, and (3) the low power of parallel trend tests, which often fail to detect violations even when they exist. Robustness of an Inference to Replacement (RIR), introduced by Frank et al. (2013, 2021), provides a powerful framework for sensitivity analysis. This paper applies conditional RIR for interaction term to compute threshold values, it can be explained as that the estimated effect can be nullified or reduced below a threshold by replacing RIR*100% of the post-treatment observations from treated to the untreated status. The paper proposes PseudoRIR as a sensitivity test for the conditional parallel trend assumption, assessing its stability. This serves as an important complement to traditional parallel trend tests with low power. Additionally, this paper applies RIR to evaluate the robustness of DID estimates against potential biases. The paper also introduces HonestRIR to account for violations of the parallel trend assumption, following the approach from Ramabachan & Roth (2022). This method produces an interval of HonestRIR adjusted for trend deviations, enhancing the robustness assessment of DID estimates. Finally, the paper extends the proposed sensitivity tests to staggered DID settings and provides an application based on the framework of Callaway & Sant’Anna (2021).

Applicants in Social Sciences

Designing Realistic and Interpretable Optimal Treatment Regimes for Personalized Education Chenguang Pan* Chenguang Pan, Youmi Suk,

Optimal dynamic treatment regimes (ODTR) have gained popularity in computer science and biostatistics for personalized recommendations, but their application in personalized education has been limited. A critical aspect of real-world ODTR applications is to ensure that the estimated regime is feasible and implementable in practice. This study addresses this challenge by incorporating feasibility constraints into the ODTR estimation process. Given that pre-determined constraints may not always be available, we develop a data-driven, post-determined strategy. Specifically, we incorporate a propensity score-based constraint into the ODTR estimation procedure. Instead of using arbitrarily chosen, one-size-fits-all thresholds (e.g., 0.05), our method dynamically adjusts thresholds at each decision stage based on data. This approach accommodates both binary and multi-categorical treatments over multiple stages. Through simulation studies, we demonstrate the effectiveness of our method, coupled with Targeted Maximum Likelihood Estimation (TMLE), in producing more feasible treatment regimes, albeit with a slight trade-off in utility. Feasibility is evaluated using the coverage rate with a set of observed treatment sequences. Finally, we illustrate the trade-off between the utility and feasibility of ODTRs using data from High School Longitudinal Study of 2009 to recommend the optimal and feasible math course for each student at each stage of their high school education.

Applicants in Social Sciences

Graphical Criteria of Recoverability under Not Missing at Random Case Jiwoo Kim* Jiwoo Kim, Felix Thoemmess,

Missing data is a well-known source of bias in quantitative research, with Not Missing at Random (NMAR) cases being particularly challenging due to the dependence between missingness itself and target variables with missing values. This dependence makes it generally impossible to recover target estimands without bias. This study aims to provide a generalizable framework for determining the recoverability of statistical parameters under NMAR conditions. Specifically, we introduce criteria that classify NMAR models based on the recoverability of key parameters, such as the slope and intercept in linear models. To achieve this, we use causal graphical models and path tracing rules to establish systematic criteria for identifying different types of bias. Building on previous work by Mohan and Pearl (2021) and Thoemmes and Mohan (2015), we propose a new type of Missing-Directed Acyclic Graph (m-DAG). By applying path tracing rules in this m-DAG, we derive conditions under which path coefficients can be recovered without bias. This study contributes to the methodological literature by offering a structured approach for researchers to assess the impact of NMAR on their analyses and determine whether unbiased estimates can still be obtained under NMAR conditions.

Applicants in Social Sciences

Estimating Causal Effects of Time-Varying Treatments on Survival Outcomes: A Case-Control Approach Ying Zhou* Ying Zhou, Xiaohui Yin, Kun Chen, Avijit Mitra, Hong Yu,

Estimating the causal effect of time-varying treatments on survival outcomes in large observational studies is computationally intensive, especially for rare outcomes. The iterative conditional expectation (ICE) estimator in the g-formula framework is effective but becomes burdensome when bootstrapping for variance estimation. The rarity of outcomes at each time point can also lead to biased estimates due to class imbalance in logistic regression and other models. To address these challenges, we propose a novel case-control enhanced g-formula that integrates case-control sampling with ICE estimation. This approach reduces computational cost while preserving unbiasedness and improving estimation stability. By strategically selecting informative subsets, it mitigates class imbalance and enhances efficiency. We apply this method to a large-scale cohort study on social determinants of health and suicide risk, demonstrating its effectiveness for rare outcome modeling in longitudinal electronic health record (EHR) studies.

Application in Public Health

Estimating the Effects of Roe v. Wade Being Overturned on State-Level Abortion Rates
Flavia Jiang* Flavia Jiang, Brian Cho, Kyuseong Choi, Raaz Dwivedi, Kyra Gan,

This study investigates the causal impact of the Supreme Court’s 2022 Dobbs v. Jackson Women’s Health Organization decision, which overturned Roe v. Wade (1973) and returned abortion regulation to individual states. Using the inclusive synthetic control method, we analyze changes in yearly state-level abortion rates, defined as the number of legal abortions per 1,000 women aged 15-44 by state of residence. Outcome data were sourced from the Guttmacher Institute and the Centers for Disease Control and Prevention, carefully integrated and supplemented with seven manually selected covariates to prevent overfitting. A key challenge was identifying the treated states, given the complex interplay of direct and spillover effects resulting from the varied state abortion policies. To address this, we used the Predictability, Computability, and Stability framework to estimate effects under perturbed assumptions about the treatment group. We find that the Dobbs decision consistently decreased abortion rates in states such as Oklahoma, Texas, and Louisiana, while increasing rates in states like Nevada, Illinois, and Delaware across assumption perturbations. Further, the Dobbs decision caused the aggregate abortion rate across 45 states of interest to rise by 0.6% to 2.1% in 2023. These results highlight the nuanced public health impacts of Dobbs, though interpretation should be cautious given data and methodological limitations.

Applications in Health and Biology

Identifying causal proteomic targets for cardiovascular disease via robust Mendelian randomisation. Christopher Aldous Oldnall* Christopher Aldous Oldnall, Sjoerd Viktor Beentjes, Ava Khamseh,

Mendelian randomisation (MR), a biological application of the instrumental variable (IV) framework, is a powerful tool for identifying causal relationships in biomedical research. By using genetic variants such as single nucleotide polymorphisms (SNPs), MR bridges genome-wide association studies (GWAS) and translational medicine, enabling the discovery of causal proteins in diseases like cardiovascular disease (CVD). However, pleiotropy—violations of the exclusion restriction criterion—remains a challenge. While traditional methods like Generalised Summary-data-based Mendelian Randomisation (GSMR) attempt to address pleiotropy by excluding invalid instruments, they rely on assumptions about instrument validity. Newer novel pleiotropically robust estimators leverage potentially invalid instruments against valid ones, offering a systematic, assumption-light approach to handling pleiotropy.

Using proteomic data from the UK Biobank, we identify causal proteins across multiple CVD risk traits, including hypertension and diabetes. We evaluate the overlap and differences between proteins identified by traditional MR methods and the pleiotropy robust estimators, critically assessing their consistency and robustness. This work highlights the importance of pleiotropically robust MR in advancing causal discovery, providing a reliable framework for translating genetic insights into clinically meaningful applications.

Applications in Health and Biology

Assessing Principal Causal Effects with Outcome-dependent Sampling Using Principal Score Methods: An Application to the E3N Cohort Lisa Braito* Lisa Braito, Fabrizia Mealli, Vittorio Perduca, Gianluca Severi,

Outcome-dependent sampling designs, commonly known as case-control studies, are widely used in epidemiology to estimate the impact of an exposure on a binary outcome. These designs involve sampling individuals from a population conditional on observed outcomes. Conventional methods like logistic regression, though prevalent, may be limited for valid causal inference, especially when the goal is to understand causal mechanisms through intermediate variables. We show that directly applying principal stratification methods to case-control designs can yield biased estimates. To address this, we propose a method for estimating principal causal effects in such designs, building on principal score methods developed for observational studies and incorporating external auxiliary information about the target population. This study investigates the relationship between menopausal hormone therapy (MHT), mammographic density, and breast cancer (BC) risk using a principal stratification approach. Evidence suggests MHT increases BC risk, partially mediated by mammographic density. Data are from a nested case-control study within the French E3N cohort. To validate our methodology, we use a secondary intermediate variable, body mass index (BMI), measured across the cohort, to simulate case-control samples and compare results to full-cohort analyses. Our findings refine causal inference methods for complex sampling designs, offering a robust framework for epidemiological research.

Applications in Health and Biology

Assessing Effects of HIV Testing on HIV Incidence in Rural KwaZulu-Natal, South Africa Ke Zhang* Ke Zhang, Ashley Buchanan, Natallia Katenka, Collins Iwuji, Laura Forastiere,

Although there are effective strategies to control the HIV epidemic, it remains a significant individual and public health challenge in South Africa. HIV testing is the gateway to HIV treatment in those who have acquired HIV and HIV prevention in those who tested HIV negative. Existing studies have suggested that HIV testing has a significant effect in reducing HIV incidence. However, these studies have not fully assessed spillover effects, the effects of one's HIV testing on HIV incidence among unexposed others. Assessing spillover can provide a more complete understanding of the impact of HIV testing. The data we used is from ANRS 12249 treatment as prevention (TasP) trial, conducted in a rural region of South Africa from March 2012 to July 2016. We grouped participants by homesteads and assume partial interference limited to the homestead, estimated both the direct (i.e., the intervention effect under exposure versus no exposure while holding other factors constant) and spillover effects of altering the proportion of HIV testing in the homestead on subsequent HIV incidence. Estimation was carried out with a marginal structured model fit with time-varying inverse probability weights. On average in the study population, there were fewer new HIV cases under HIV testing exposure (i.e., direct effect) or higher proportion of HIV testing uptake in an untreated individual's homestead (i.e., spillover effect). Further research is needed to understand the underlying mechanisms.

Applications in Health and Biology

Inferring Directed Gene Regulatory Networks from Single-Cell Perturb-Seq Data and Poisson-Log Normal Models Zhongxuan Sun* Zhongxuan Sun, Hyunseung Kang, Sunduz Keles,

We present a new framework for inferring directed gene regulatory networks (GRNs) with a novel Poisson-log normal (PLN) model for single-cell Perturb-Seq data, a new type of experiment in genomics where for each cell (i.e., the study unit), CRISPR-based technology is used to either silence a particular gene (i.e., the treated cell) or not (i.e., the control cell). Unlike existing methods for discovery of GRNs that rely on acyclicity and normality assumptions, our PLN-based framework accommodates count-based, overdispersed transcriptomic data and potential feedback loops, an important feature of single-cell Perturb-Seq data. Additionally, unlike most existing methods for GRN discovery which rely on observational expressional data, the estimates from the PLN-based framework are not biased from unmeasured confounders due to the design of Perturb-Seq experiments and accurately capture regulatory directionality. We validate our inferred networks through complementary epigenomic and proteomic evidence: ChIP-seq datasets confirm the presence of regulatory interactions at key promoter and enhancer regions, and protein–protein interaction (PPI) networks that carry functional relationships among regulators and their targets. These multi-modal validations underscore the robustness of our method for reconstructing large-scale GRNs. Our results highlight the power of single-cell perturbation data, integrated with comprehensive molecular profiling, to reveal intricate regulatory circuitry

Applications in Health and Biology

Small sample size but still definitive: secondary analysis of a randomized trial on central neck dissection for papillary thyroid cancer Benjamin Cher* Benjamin Cher, Qinyun Lin, Rebecca Sippel, Kenneth Frank, Courtney Balentine,

A recent randomized controlled trial (RCT) sought to address a major controversy in endocrine surgery: whether low-risk thyroid cancer should be treated by removing only the thyroid or whether it is also necessary to routinely resect lymph nodes. The original trial found no difference in recurrence after randomizing 30 patients to standard thyroid removal (control group) and 31 to removal of both thyroid and lymph nodes (treatment group). However, the study has not noticeably changed clinical practice because of concerns about power and external validity. In this study, we use novel frequentist and Bayesian approaches to address these concerns, and we conclude that the results are robust despite small sample size and recruitment being limited to a single academic center. Using the Robustness of Inference to Replacement (RIR), we find that a large percentage of both the treatment and control groups would have to be replaced with patients who experienced similar probability of recurrence as the other group (RIR=28, requiring double switches). Additionally, in >99% of 1,000 bootstrap repetitions of the trial, we find no difference in recurrence. Finally, using Bayesian statistics, we find the study results would only change if the true effect of lymph node removal were a 100% reduction in recurrence relative to the standard surgery. We also estimate that repeating the trial would yield the same result ~70% of the time.

Applications in Health and Biology

A Large Clinical Behavioral Model: an approach to deep causal policy learning Jonas Knecht*
Jonas Knecht, Maya Petersen, Jon Kolstad,

We present a novel deep-learning approach (a “large clinical behavioral model”) to studying a wide array of dynamic behavioral regimes in diverse non-randomized health data settings. Our proposed methodology – behavioral policy learning – answers the question; “what is the distribution of optimal paths of clinical actions for a given patient at a given point in time”. We present a distinctly behavioral approach and focus on (1) causally identifying the link between providers and their patients’ outcomes, as well as (2) learning the distribution of provide-specific counterfactual clinical action paths. In this way we are able to identify the link between differences in the distribution of clinical action paths chosen by providers and their causal impact on patient outcomes.

Our identification strategy relies on conditional random assignment of providers to patients, which allows us to identify the causal link between provider assignment and patient outcomes. This identification strategy is applicable to clinical settings such as the ED and other health care delivery settings where patient encounters cross multiple provider shift changes, as well as assignments to PCPs and specialists in which common instruments are available.

The feasibility of the proposed approach is based on recent advances in generative AI and combines the unique strengths of these models, i.e. black box emulation of human behavior they have been trained on, with a formal causal inference framework.

Applications in Health and Biology

Saving patient years through innovative use of RWD Morten Medici* Morten Medici,

Real-world evidence promises large benefits for regulatory decision-making in health care, but the impact has not yet materialized within population health diseases such as cardiovascular and metabolic health, despite successes in other areas. The current pharmaceutical landscape necessitates larger randomized controlled trials which impacts more patients, delays efficacious medication and ultimately hinders drug development. Innovations within the field of casual inference, on the combination of data sources ensure protection against much of the potential bias that are endemic to purely observational data, and thus have the potential to revolutionize how large, randomized outcomes trials are conducted and analyzed in the future.
However, there is still a huge gap between the solutions in the causal inference community and the application and implementation within population health diseases. Adoption of such approaches in industry is hindered by limited regulatory experience and challenges in implementing novel causal methods into analysis plans, as well as limited acceptability by the wider clinical society, including lack of familiarity with the methods and the assumptions on which the methods rely.
Based on current ongoing activities this presentation will identify challenging gaps, indicate key opportunities and suggest a viable path to successfully implementing more causal inference methods within large organizations in health care.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

Productivity and AI Tools: a causal analysis Leandro Zanon Siqueira* Sarah Brodbeck, Rinaldo Oliveira Junior, Ciro Akiyoshi Higashi,

In real life, in many situations, when a company decides to start using a solution or tool, randomized controlled tests (RCT) cannot be done. That happens for many reasons, such as the cost of opportunity of not enabling all employees to use the tool or even a strong strategic decision of the company. In this scenario, causal inference techniques allow us to isolate the effect of other confounder variables so that we can estimate the real gain of using the tool, even without a RCT.

In this present study, we use Directed Acyclic Graphs (DAG) to estimate how much faster teams in the largest retail bank of Latin America, Itau Unibanco, can develop software by using GitHub Copilot.

Using DAGs and PyWhy library, we compared teams that used GitHub Copilot with those who didn’t (controlling by confounders and mediator variables) and it was possible to estimate that teams that used the AI tool spent about 11% less time with software development.

With this study, it was possible to understand how much it was worth the investment made in GitHub Copilot licenses by Itau Unibanco. Also, this study created a comparable framework for competitive AI tools, allowing the company to choose solutions that bring more efficiency and that present a better cost-benefit ratio.

Applications in Physical Sciences, Engineering, Environment and Miscellaneous Applications

Evaluating the impact of alternative protein in the United States Anna Thomas* Anna Thomas, Maya Mathur,

Industrial animal agriculture is a major contributor to climate change, deforestation, air and water pollution, zoonotic disease transmission, and antibiotic resistance. Substitutes for animal products ('alternative protein') are a promising approach to shift consumption in a more sustainable direction. Here, using the Nielsen Consumer Panel – a dataset of more than 1 billion grocery store transactions – we evaluate the impact of alternative protein in the United States. Specifically, we aim to assess whether alternative protein availability reduces climate impacts of food choices. We conduct our pre-registered analyses at the household, grocery store, and regional levels. In addition to descriptive analysis, we apply several causal inference techniques including synthetic control, the autoregressive distributed lag model, and interrupted time series. We compare assumptions, relative strengths, and limitations of each of these methods. We also assess sensitivity of our findings to unmeasured confounding.

Bayesian Causal Inference

The Importance of Ablation Studies for Complex Nonparametric Causal Models * Hugo Gobato Souto, Francisco Louzada Neto, Hugo Gobato Souto,

Ablation studies are essential for understanding the contribution of individual components within complex models, yet their application in nonparametric treatment effect estimation remains limited. This paper emphasizes the importance of ablation studies by examining the Bayesian Causal Forest (BCF) model, particularly the inclusion of the estimated propensity score πˆ(xi) intended to mitigate regularization-induced confounding (RIC). Through a partial ablation study utilizing a total of nine synthetic, we demonstrate that excluding πˆ(xi) does not diminish the model's performance in estimating average and conditional average treatment effects or in uncertainty quantification. Moreover, omitting πˆ(xi) reduces computational time by approximately 21%. These findings could suggest that the BCF model's inherent flexibility suffices in adjusting for confounding without explicitly incorporating the propensity score. The study advocates for the routine use of ablation studies in treatment effect estimation to ensure model components are essential and to prevent unnecessary complexity.

Causal Discovery

Nonlinear Causal Learning through Sequential Orientation of Edges in an Equivalence Class Stella Huang* Stella Huang, Qing Zhou,

While recent advances have established the identifiability of a directed acyclic graph (DAG) under assumptions on the structural causal model (SCM), many existing causal discovery methods rely heavily on strong structural and distributional assumptions, perform only bivariate comparisons for random variables, or require substantial computational time. In this work, we introduce a novel constraint-based algorithm for learning the causal DAG under the non-linear setting. Building upon the equivalence class of a DAG, our approach incorporates a novel procedure to sequentially determine the true causal direction of undirected edges. We propose a ranking procedure to determine the evaluation order of edges by formulating the pairwise Additive Noise Model (PANM) to establish the necessary conditions for edge orientation. The edges are then ordered by an associated independence measure, in which the first rank is guaranteed to fulfill the conditions for orientation. To determine the true edge direction, we employ a statistical test that compares the log-likelihoods, evaluated with respect to the competing directions, of the sub-graph consisting of the two nodes and their parents. We further establish its theoretical guarantees at the population level and consistency in the large-sample limit. Experimental results demonstrate that our method is robust and computationally efficient, both when model assumptions hold and when they are violated.

Causal Discovery

Identifying Changes in Causal Relationships for Time Series Data Stephen Savas* Stephen Savas,

In time-series data, existing causal discovery algorithms are capable of learning causal graphs from data, and newer methods have recently been extended to handle changes in causal relationships; for example, a drug may lose its effectiveness if a patient develops a tolerance. However, these models are often impractical, as they make strong assumptions about the underlying data (such as a single change point or periods of stationarity), or they require substantial computing resources. In order to adjust for a changing causal graph, it is important to understand when an existing graph no longer works, at which point users can retrain on recent data that is representative of the changed ground truth. By retesting an existing graph, users do not have to assume that the graph will change in the future — instead, they can use their preferred causal discovery algorithm that properly meets their own assumptions and time complexity needs, and only retrain if necessary to create an updated graph. This paper proposes a novel method that detects the situation where a causal graph no longer accurately represents the data. The approach works for causal discovery methods that estimate the causal effect by analyzing the perceived noise of a time series relative to the model’s output to understand when the level of noise is abnormal, at which point the model is likely no longer valid. Such a test allows for more practical applications of causal discovery relative to existing methods.

Causal Discovery

Valid post-selection inference for penalized G-estimation Ajmery Jaman* Ajmery Jaman, Ashkan Ertefaie, Michèle Bally, Renée Lévesque, Robert Platt, Mireille Schnitzer,

Understanding treatment effect heterogeneity is important for decision-making in medical and clinical practices, or handling various engineering and marketing challenges. When dealing with high-dimensional covariates or when the effect modifiers are not predefined and need to be discovered, data-adaptive selection approaches become essential. However, data-driven model selection complicates the quantification of statistical uncertainty in post-selection inference and makes it difficult to approximate the sampling distribution of the target estimator. Such model selection tends to favor models with strong effect modifiers with an associated cost of inflated type I errors. Although several frameworks and methods for valid statistical inference have been proposed for ordinary least squares regression following data-driven model selection, fewer options exist for valid inference for effect modifier discovery in causal modeling contexts. To fill this gap, we extend two different methods to develop valid inference for penalized G-estimation that investigates effect modification of proximal treatment effects within the structural nested mean model framework. In our simulation study, the proposed methods effectively controlled the false coverage rates for the target parameters, while the naive inference based on the sandwich variance estimator resulted in false coverage rates higher than the nominal level. We also illustrate these methods with a real-data application.

Causal Discovery

Valid post-selection inference for penalized G-estimation * Ajmery Jaman, Ashkan Ertefaie, Michèle Bally, Renée Lévesque, Robert Platt, Mireille Schnitzer, Ajmery Jaman,

Understanding treatment effect heterogeneity is important for decision-making in medical and clinical practices, or handling various engineering and marketing challenges. When dealing with high-dimensional covariates or when the effect modifiers are not predefined and need to be discovered, data-adaptive selection approaches become essential. However, data-driven model selection complicates the quantification of statistical uncertainty in post-selection inference and makes it difficult to approximate the sampling distribution of the target estimator. Such model selection tends to favor models with strong effect modifiers with an associated cost of inflated type I errors. Although several frameworks and methods for valid statistical inference have been proposed for ordinary least squares regression following data-driven model selection, fewer options exist for valid inference for effect modifier discovery in causal modeling contexts. To fill this gap, we extend two different methods to develop valid inference for penalized G-estimation that investigates effect modification of proximal treatment effects within the structural nested mean model framework. In our simulation study, the proposed methods effectively controlled the false coverage rates for the target parameters, while the naive inference based on the sandwich variance estimator resulted in false coverage rates higher than the nominal level. We also illustrate these methods with a real-data application.

Causal Fairness, and Bias/Discrimination

The Association between partisan and racial gerrymandering. Christiana Drake* Christiana Drake, Xiner Zhou, Bala Rajaratnam,

In the United States the members of the House of Representatives are elected every 2 years directly by the voters in their respective districts. District boundaries are drawn by the States. A candidate is elected by a simple majority of the voters in the respective district. In 1965 Congress passed the Voting Rights Act which prohibits the disenfranchisement of racial minority voters. The US Supreme Court has ruled that it is not necessarily prohibited to draw boundaries in such a way as to favor one political party. This is popularly referred to as gerrymandering. However, the Supreme Court has not ruled on a case involving gerrymandering. Some states have independent or bipartisan commissions that draw boundaries, in other states the legislature draws the boundaries. It is tempting to create congressional districts in these states that favor the party that controls the legislature. It can be shown that it is possible to create districts in such a manner that a party that has more than 50% of the vote nevertheless ends up with fewer than half the congressional representatives in the state. We use the counterfactual framework to explore the relationship between partisan gerrymandering and the potential for racial disenfranchisement. In our model the counterfactual is a measure of the value of a vote in districts that would occur in a gerrymandered district vs. a district drawn by an independent commission.

Causal Fairness, and Bias/Discrimination

Advancing Distribution Decomposition Methods Beyond Common Supports: Applications to Racial Disparities Bernardo Modenesi* Bernardo Modenesi,

I generalize and state-of-the-art approaches that decompose differences in the distribution of a variable of interest between two groups into a portion explained by covariates and a residual portion. The method I propose relaxes the overlapping supports assumption commonly imposed in causal inference methods that compare groups, such as Oaxaca-Blinder and propensity score methods, which allows groups being compared to not necessarily share exactly the same covariate support. I illustrate my method revisiting the black-white wealth gap in the U.S. as a function of labor income and other variables. Traditionally used decomposition methods would trim (or assign zero weight to) observations that lie outside the common covariate support region. On the other hand, by allowing all observations to contribute to the existing wealth gap, I find that otherwise trimmed observations contribute from 3% to 19% to the overall wealth gap, at different portions of the wealth distribution.

Causal Fairness, and Bias/Discrimination

Using Causal Inference to Unmask Racial Discrimination in Traffic Enforcement via Proxies Kai Cooper* Kai Cooper, Gregory Lanzalotto, Dean Knox, Jonathan Mummolo,

Racial bias in policing is a defining issue of modern law enforcement. Excessive use of force often dominates public attention, however, traffic enforcement, a far more common police-civilian interaction receives less scrutiny. The killing of Tyre Nichols following an unlawful traffic stop in Memphis underscores how these encounters can escalate into violence, and it is therefore of great interest to study the encounter at the first stage. Uncovering such disparities is important for building trust in law enforcement and are crucial drivers of policy debate on equitable policing. However, administrative data on policing poses formidable challenges to this goal due to (i) measurement error of key variables, (ii) sample selection, and (iii) latent confounding.

This paper presents a novel framework for measuring racial discrimination in traffic policing by using automated traffic enforcement, such as roadside traffic cameras, as a race-neutral proxy for dangerous driving. Existing methods in causal inference using proxy variables are adapted and extended to address questions of discrimination. Partial identification strategies are developed for estimands quantifying race-based policing with minimal assumptions, while explicitly accounting for challenges (i)-(iii). We illustrate the effectiveness of this approach through a simulation study and an application to a dataset of traffic citations from a large U.S. city.

Causal Inference and SUTVA/Consistencies Violations

Estimating Spillover Effects in Longitudinal Data under Unknown Interference Ye Wang* Ye Wang, Michael Jetsupphasuk,

In longitudinal data where units are situated in a space or social network, the outcome of one unit may not only be affected by its own treatment assignment history, but also the treatment assignment histories of others. The presence of interference raises a significant challenge for researchers aiming to discern the direct and spillover effects of treatments across these dimensions, given that the interference structure—how an observation's outcome is influenced by the treatment status of others—is often unknown to the researcher. In this paper, we put forward a design-based framework that combines marginal structural models with recent advancements in the literature on interference to address these complexities. We define estimands that enable researchers to separate direct effect of the treatment from spillover effects under unknown interference structures. We then develop estimators that are consistent for these estimands and asymptotically normal, assuming sequential ignorability and mild constraints on the extent of dependence caused by interference. Additionally, we introduce methods for constructing valid confidence intervals. Unlike existing approaches based on exposure mapping, our method circumvents the challenge of calculating exposure probabilities in longitudinal settings. Its effectiveness is demonstrated through simulations and replications of two empirical studies.

Causal Inference and SUTVA/Consistencies Violations

Semiparametric Proximal Causal Inference with Invalid Proxies Myeonghun Yu* Myeonghun Yu, Eric Tchetgen Tchetgen, Xu Shi,

Proximal causal inference, introduced by Miao, Geng, and Tchetgen Tchetgen [Biometrika 105 (2018) 987–993], has garnered significant attention for estimating causal effects in the presence of unmeasured confounders through the use of proxy variables. This framework has found applications in diverse fields such as longitudinal data analysis, mediation analysis, and survival analysis. However, current approaches critically assume complete knowledge of proxy validity, rendering identification results invalid when some candidate proxies are not valid. To address this limitation, we propose a novel identification framework that ensures validity as long as at least some candidate treatment-inducing proxies are valid, even without identifying the exact subset of valid proxies. Building on this framework, we develop the semiparametric theory for the average treatment effect in the presence of invalid treatment-inducing proxies and establish properties of doubly robust and locally efficient estimators. Extensive simulations validate the proposed framework, demonstrating its practical utility.

Causal Inference Education

Disparity Analysis: A Tale of Two Approaches Aleksei Opacic* Aleksei Opacic, Lai Wei, Xiang Zhou,

To understand the patterns and trends of various forms of inequality, quantitative social science research has typically relied on statistical models linking the conditional mean of an outcome variable to a set of explanatory factors, a prime example of which is the widely used Kitagawa-Oaxaca-Blinder (KOB) method. In this paper, we explicate, contrast, and extend two distinct approaches to studying group disparities, which we term the descriptive approach, as epitomized by the KOB method and its variants, and the prescriptive approach, which focuses on how a disparity of interest would change under a hypothetical intervention to one or more manipulable treatments. For the descriptive approach, we propose a generalized nonparametric KOB decomposition that considers multiple (sets of) explanatory variables sequentially. For the prescriptive approach, we introduce a variety of stylized interventions, such as lottery-type and affirmative-action-type interventions that close between-group gaps in treatment. We illustrate the two approaches by assessing the Black-White gap in college completion, how it is statistically explained by racial differences in socioeconomic background, academic performance, and college selectivity, and the extent to which it would be reduced under hypothetical reallocations of college-goers from different racial and economic backgrounds into different tiers of college – reallocations that could be targeted by race- or class-conscious admissions policies.

Causal Inference in Networks

Estimation of interference effects in networks with community structures Yuhua Zhang*
Yuhua Zhang, Ruoyu Wang, Shuo Sun,

In causal inference, the interference effect – whether an individual's outcome is affected by the treatment of its neighbors – is gaining increasing attention. The majority of existing work assumes that the observed networks represent the true underlying interference networks. In practice, this assumption is not correct and leads to the bias in the estimation of causal effects. In this work, we address the problem of whether true interference effects exist given the observed networks. In particular, our proposed framework leverages the community structures in the networks and assumes the interference effects are identically distributed for individuals in the same community. We demonstrate that our proposed model is able to identify the interference effects in theory and in simulations. We apply our proposed framework to the stroke encounter data and evaluate the potential effect of performing EVT procedures in one hospital on its neighbors.

Causal Inference in Networks

Causal Interpretation of Anterial Graphs and Constrained Confounder Selection Kai Teh*
Kai Teh, Kayvan Sadeghi, Terry Soo,

We provide a valid causal interpretation for anterial graphs, a class of graphs containing directed acyclic graphs, that are also closed under conditioning and marginalisation. In this setting, we provide a graphical procedure that returns a graph which is valid in jointly representing post-intervened variables and pre-intervened (observational) variables, thus extending single world intervention graphs, originally introduced by Richardson and Robins (2013). Using this graphical representation, we provide an element-wise procedure of selecting confounders given flexibly prescribed set constraints.

Causal Inference in Networks

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

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). I propose a new way to increase control in these estimation procedures by applying network theory models motivated by a discrete choice framework. I 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. I argue that the structure of connections in this choice-network allows the researcher to further improve controls when modeling the outcome variable. For instance, I 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, I propose several ways to enhance causal inference exercises, as well as to compute heterogeneous treatment effects.

Design of Experiments

Balancing Efficiency and Inference in Adaptive Experiments Daniel Molitor* Daniel Molitor, Ian Lundberg,

Adaptive experimental designs dramatically improve data efficiency by dynamically learning treatment effects, but they pose significant challenges for statistical inference. Standard estimators such as sample means and inverse propensity weighted estimators often yield biased treatment effect estimates, undermining a key advantage of randomized experiments. Additionally, adaptive experiments allocate samples unevenly, concentrating power on the optimal treatment(s) while leaving sub-optimal treatment arms underpowered—a critical limitation when researchers need reliable inference across all treatment arms.

We propose a framework that builds on existing methods, such as the Mixture Adaptive Design by Liang and Bojinov, which generate unbiased treatment effect estimates with anytime-valid confidence sequences. By dynamically eliminating treatment arms when their confidence sequences exclude zero, our framework harnesses the efficiency gains of adaptive experiments while maintaining valid inference and sufficient power across all treatment arms.

We validate this framework through simulations and a real-world experiment assessing how a criminal record impacts hiring decisions, moderated by resume features. We compare the results of our adaptive experiment to those of a randomized experiment, demonstrating our method’s ability to significantly reduce sample size requirements while maintaining valid inference and ensuring sufficient power across all treatment arms.

Design of Experiments

Optimal Adaptive Experimental Design for Estimating Treatment Effect Jiachun Li* Jiachun Li, David Simchi-Levi, Yunxiao Zhao,

Given n experiment subjects with potentially heterogeneous covariates and two possible treatments, namely active treatment and control, this paper addresses the question of determining the optimal accuracy in estimating the treatment effect. Furthermore, we propose an experimental design that approaches this optimal accuracy, giving a (non-asymptotic) answer to this fundamental question. The methodological contributions are listed as follows. First, we establish an idealized optimal estimator with minimal variance as benchmark, and then demonstrate that adaptive experiment is necessary to achieve near-optimal estimation accuracy. Second, by incorporating the doubly robust method into sequential experimental design, we frame the optimal estimation problem as an online bandit learning problem, bridging the two fields of statistical estimation and bandit learning. Using tools and ideas from both bandit algorithm design and adaptive statistical estimation, we propose a general low switching adaptive experiment framework, which could be a generic research paradigm for a wide range of adaptive experimental designs. Through novel lower bound techniques for non-i.i.d. data, we demonstrate the optimality of our proposed experiment. Numerical result indicates that the estimation accuracy approaches optimal with as few as two or three policy updates.

Design-Based Causal Inference

Sharp bounds on the variance of general regression adjustment in randomized experiments
Jonas Mikhaeil* Jonas Mikhaeil, Donald Green,

A growing statistical literature focuses on causal inference in the context of experiments where the target of inference is the average treatment effect in a finite population and random assignment determines which subjects are allocated to one of the experimental conditions. In this framework, variances of average treatment effect estimators remain unidentified because they depend on the covariance between treated and untreated potential outcomes, which are never jointly observed. Conventional variance estimators are upwardly biased. Aronow, Green and Lee [Ann. Statist. 42(3): 850-871 (June 2014)] provide an estimator for the variance of the difference-in-means estimator that is asymptotically sharp. In practice, researchers often use some form of covariate adjustment, such as linear regression when estimating the average treatment effect. Adapting propositions from empirical process theory, we extend the result in [Ann. Statist. 42(3): 850-871 (June 2014)], providing asymptotically sharp variance bounds for general regression adjustment. We apply these results to linear regression adjustment and show benefits both in a simulation as well as an empirical application.

Design-Based Causal Inference

Causal subgroup discovery with error control Yao Zhang* Yao Zhang, Zijun Gao,

In randomized controlled trials, researchers are often interested in understanding the heterogeneity of treatment effects through interpretable subgroup analysis. In this paper, we propose a new subgroup analysis method that leverages regression models to discover patient subgroups with significant treatment effects. This method then validates these subgroups using the randomness of treatment assignment, thereby limiting the number of false discoveries. Through experiments, we demonstrate that this new approach not only identifies promising subgroups effectively but also provides more reliable insights for developing personalized treatment plans for patients with varying conditions.

Design-Based Causal Inference

Selective inference for data-driven subgroups based on biomarkers Zijun Gao* Zijun Gao,

In randomized experiments with heterogeneous treatment effects, subgroup analysis provides significant benefits, such as personalized treatment recommendations, but poses challenges for inference when subgroups are learned from data. Motivated by the German Breast Cancer Study, where subgroups are defined using a biomarker threshold—a common practice in clinical trials—we develop a design-based inference procedure tailored to this type of subgroup selection. The validity of our method relies solely on knowledge of the randomization mechanism, requiring no assumptions about the underlying model, making it particularly suitable for complex datasets. Compared to sample-splitting based inference, our approach is deterministic and avoids the power loss associated with reduced sample size for inference. The computation of our method is often similar to a standard randomization test without selection and requires no intricate sampling procedures to approximate conditional distributions. Furthermore, when predefined biomarkers are unavailable, we extend the method by incorporating a data-driven biomarker while maintaining the desirable properties of the approach. We demonstrate the validity and efficiency of our methods through the analysis of the GBCS dataset and simulated data.

Design-Based Causal Inference

Spatial Error Models and Unmeasured Spatial Confounding: The Underlying Experiment
Sophie Woodward* Sophie Woodward, Francesca Dominici, Jose Zubizarreta,

Regression models incorporating spatially structured error terms, known as spatial error models, are commonly used to address unmeasured spatial confounding. Despite their prevalence, the extent to which these models mitigate unmeasured spatial confounding remains contentious. In this work, we examine three canonical types of spatial error models — random effects, conditional autoregressive models, and Gaussian processes — and reinterpret their generalized least squares (GLS) estimators through the lens of weighting for causal inference with a binary treatment. Our proposed framework offers new insights into how spatial error models construct contrasts between treated and control units in space. We also provide diagnostics to characterize covariate balance, study representativeness, effective sample size, sign reversal, and the use of spatial information. Crucially, we demonstrate that the design-conditional bias of GLS estimators is determined by the relative spatial smoothness of the unmeasured confounder and the treatment. Extending these insights, we propose a spatial balancing weights estimator that targets the average treatment effect for a specified population, is sample bounded, accommodates nonlinearity and treatment effect heterogeneity, and mitigates bias from spatially smooth unmeasured confounders. We evaluate the finite-sample performance of our proposed method in simulation and apply it to estimate the effect of Superfund cleanups on birth weight.

Design-Based Causal Inference

The test-negative design for the estimation of COVID-19 vaccine effectiveness: development of statistical methods in the evolving context Helen Bian* Helen Bian, Cong Jiang, Denis Talbot, Robert Platt, Mireille Schnitzer,

The test-negative design (TND) has been widely used for the rapid estimation of vaccine effectiveness against infectious diseases. The TND typically includes individuals with a common symptom profile who are receiving a laboratory test for an infection of interest. Among them, participants who test positive for the target infection are “cases” and those who test negative are “controls”. Existing statistical approaches for the TND have certain limitations in a dynamic longitudinal setting, where data can be periodically collected from different individuals over the study period.
First, time-dependent confounders may be influenced by previous vaccination and health status, such as previous infection, while also affecting the subsequent vaccination decisions. Thus, the causal relation of interest cannot be properly estimated using traditional covariate-adjusted models. Secondly, since individuals can test positive multiple times over the study period, past infections may alter immunity and create non-positivity for vaccination for the following time-period. Therefore, we propose a causal framework that accounts for the time-varying effects as well as changing risk sets, particularly in the context of the dynamic nature of infectious diseases. We propose IPTW estimators of discrete-time hazard and hazard ratios, which can be identified from the TND samples. Simulation studies are used to show the performance of these estimators.

Design-Based Causal Inference

Unifying regression-based and design-based causal inference in time-series experiments Zhexiao Lin* Zhexiao Lin, Peng Ding,

Time-series experiments, also called switchback experiments, play increasingly important roles in modern applications and are a fundamental experimental design in practice. In this paper, we examine the design-based properties of regression-based methods for estimating treatment effects in such settings. We demonstrate that the treatment effect of interest can be consistently estimated using ordinary least squares (OLS) with an appropriately specified working model. Our analysis extends to estimating a diverging number of treatment effects simultaneously, and we establish the asymptotic properties of the resulting estimators. Additionally, we show that the heteroskedasticity and autocorrelation consistent (HAC) estimator provides a conservative estimate of the variance. Importantly, while our approach relies on OLS regression, our theoretical framework accommodates misspecification of the regression model.

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

Difference-in-differences Design with Outcomes Missing Not at Random Sooahn Shin* Sooahn Shin,

This paper addresses one of the most prevalent problems encountered by political scientists working with difference-in-differences (DID) design: missingness in panel data. A common practice for handling missing data, known as complete case analysis, is to drop cases with any missing values over time. A more principled approach involves using nonparametric bounds on causal effects or applying inverse probability weighting based on baseline covariates. Yet, these methods are general remedies that often underutilize the assumptions already imposed on panel structure for causal identification. In this paper, I outline the pitfalls of complete case analysis and propose an alternative identification strategy based on principal strata. To be specific, I impose parallel trends assumption within each latent group that shares the same missingness pattern (e.g., always-respondents, if-treated-respondents) and leverage missingness rates over time to estimate the proportions of these groups. Building on this, I tailor Lee bounds, a well-known nonparametric bounds under selection bias, to partially identify the causal effect within the DID design. Unlike complete case analysis, the proposed method does not require independence between treatment selection and missingness patterns, nor does it assume homogeneous effects across these patterns.

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

Time-Varying Causal Survival Learning Xiang Meng* Xiang Meng, Iavor Bojinov,

The paper tackles a key challenge in causal inference: how to accurately estimate causal effects when the timing of treatment varies from patient to patient. Drawing on the well-known Stanford Heart Transplant study, we show how staggered adoption assumptions can be combined with survival analysis techniques to address this issue. In organ transplantation, for example, factors like donor availability and patient readiness often determine when a patient receives treatment, which can bias estimates if not handled correctly. By identifying the conditions that link staggered adoption designs to survival analysis, we demonstrate how existing survival methods can retain their causal interpretability under time-varying treatments. We further boost the precision and robustness of these estimates by incorporating double machine learning, which allows us to manage complex relationships between patient characteristics and survival outcomes. Through simulations and an analysis of heart transplant data, our approach outperforms traditional methods, reducing bias and offering theoretical guarantees for greater efficiency in survival analysis.

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

Difference-in-Differences with Non-Ignorable Attrition Javier Viviens* Javier Viviens,

Unbalanced panels are frequently used in Difference-in-Differences (DiD) applications. In this paper, I employ principal stratification analysis to highlight the potential drawbacks of the DiD research design when the outcome is missing for some units. Specifically, the conventional ATT estimand may not be well defined, and the DiD estimand cannot be interpreted causally without additional assumptions. To address these issues, I develop an identification strategy to partially identify causal effects on the set of units for which the outcome is observed and well-defined under both treatment and control. I adapt Lee bounds to the DiD setting, replacing the unconfoundedness assumption in the original trimming strategy proposed by Lee (2009) with a principal parallel trend assumption. I also explore how to leverage multiple sources of attrition to relax the monotonicity assumption, thereby allowing the four latent strata to exist, which may be of independent interest. Alongside the identification results, I present estimators and their asymptotic distributions. I illustrate the relevance of the proposed methodology by analyzing a job training program in Colombia.

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

A Meta-learner for Heterogeneous Effects in DiD and General Conditional Functionals Under Covariate Shift Hui Lan* Hui Lan, Haoge Chang, Eleanor Dillon, Vasilis Syrgkanis,

We address the problem of estimating heterogeneous treatment effects in panel data, leveraging the popular difference-in-differences (DiD) framework under the parallel trends assumption. We propose a novel doubly robust meta-learner for the conditional average treatment effect on the treated. Our framework allows for interpretable projections onto lower-dimensional subsets of interest, and can be easily implemented as a convex loss minimization problem involving a set of auxiliary models. By leveraging Neyman orthogonality, our proposed appraoch is robust to estimation errors in the auxilliary models. As a generalization to this problem, we also provide an estimation framework for general functionals under covariate shift. Additionally, we extend our methodology to handle binary instruments out of practical concerns when there is non-compliance in the treatment assignment. Empirical results demonstrate the superiority of our approach over existing baselines, and we provide a detailed discussion on the performance of our meta-learner in relations with different identifying assumptions.

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

Sensitivity Analysis for Difference-in-Differences and Related Designs Thomas Leavitt* Thomas Leavitt,

Applied researchers increasingly acknowledge that Difference-in-Difference's parallel trends assumption is unlikely to hold exactly. Hence, researchers increasingly assess the sensitivity of causal conclusions to violations of parallel trends given observable pre-treatment differences in trends. In this paper, I propose a new sensitivity analysis for the broader class of controlled pre-post designs, which nests Difference-in-Differences as a special case. To do so, I first derive a general identification framework that unifies controlled pre-post designs. In this framework, one uses models to predict untreated outcomes and then corrects the treated group's predictions using the comparison group's observable prediction errors. The point identification assumption analogous to parallel trends is that treated and comparison groups would have equal prediction errors (in expectation) under no treatment. Within this framework, I then formally ground a sensitivity analysis in the logic of multiple robustness, wherein the sensitivity (or, conversely, robustness) of a causal conclusion depends on the number of independent conditions under which it would hold. I argue that this sensitivity analysis improves upon existing alternatives: Using several real-world applications, I show that existing analyses show low sensitivity when observable data suggest it ought to be high and high sensitivity when observable data suggest it ought to be low.

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

Counterpart Statistics in the Matched Difference-in-Differences Design Sean Tomlin* Sean Tomlin,

Difference-in-differences (DiD) estimates intervention effects under the parallel trends assumption, but nuisance trends can bias estimates. Matching methods that balance pre-intervention trends have been used, yet we show they fail to adjust for latent confounders and introduce regression to the mean bias. Instead, we advocate for methods grounded in explicit causal assumptions about selection bias. We also propose a Bayesian approach to assess parallel trends, avoiding the challenges of specifying non-inferiority thresholds. We demonstrate our method using Medical Expenditure Panel Survey data to estimate the impact of health insurance on healthcare utilization.

Dynamic Treatment Regimes

Structural Nested Models in Target Trial Emulation Fuyu Guo* Fuyu Guo, Oliver Dukes, Mats Julius Stensrud, James Robins,

Target trial emulation is a popular method for estimating effects of treatment regimes from observational data. In the emulation, new trials, indexed by time, are initiated at fixed intervals. A subject participates in every trial for which eligibility criteria are met. Current methods treat each time-specific trial separately. For instance, for a trial comparing the regimes “always” versus “never” treat from initiation at t onwards, it is common to fit a hazard or risk ratio (RR) model that includes a treatment indicator and its potential confounders. Subjects are censored if they later change treatment, with inverse probability weighting to adjust for the censoring. If most subjects change treatment, the estimates will be inefficient. In this paper we propose more efficient estimators by introducing regime-specific structural nested target trial emulation models (SNTTEM). Given a regime, a SNTTEM imposes parametric models for all time-specific blip functions of the eligible subjects and leaves those for the ineligible unrestricted. A time-specific blip function quantifies on a mean scale the effect of initiating the regime at a time t versus one period later, as a function of past history. The intersection of all the earlier time-specific RR models constitutes a single SNTTEM with regime “always take the treatment that one took last time”. We show that SNTTEM can be fitted using g-estimation, a method that censors less and is more efficient than current methods.

Generalizability/Transportability

Causal machine learning for generalizing heterogeneous treatment effects Vanessa Rodriguez* Vanessa Rodriguez, Karla Diaz Ordaz, Brieuc Lehmann,

Many methods exist to generalize inferences from randomized trials to target populations, with most approaches focusing on the average treatment effect (ATE). However, the relative performance of methods for generalising conditional average treatment effects (CATEs) using machine learning (ML) meta-learners remains under explored, particularly under varying degrees of sampling bias, CATE complexity, and runtime confounding.
In this talk, we compare two approaches: the generalized T-Learner, which uses inverse probability of sampling weights (IPSW) but lacks rate robustness when using ML models, and the generalized DR-Learner, a debiased estimator that addresses these limitations. Following an extensive simulation study, we observe that the generalised DR-Learner consistently exhibits lower median mean squared error (MSE) than both the standard and generalized T-Learner in almost all cases, but especially in settings with complex sampling mechanisms and smaller sample sizes.
We also explore the use of conformal prediction to ensure valid inference, which has traditionally been a limitation of using ML methods for causal inference. We present the coverage and interval lengths obtained by using weighted conformal inference, allowing us to obtain prediction intervals for causal effects under covariate shift.
We anticipate these findings will provide practical guidance to practitioners wanting to incorporate ML methods in their analysis.

Generalizability/Transportability

Generalizing Causal Effects with Noncompliance Zhongren Chen* Zhongren Chen, Melody Huang,

Standard approaches in generalizability often focus on generalizing the intent to treat (ITT). However, in practice, a more policy-relevant quantity is the generalized impact of an intervention across compliers. While instrumental variable (IV) methods are commonly used to estimate the complier average causal effect (CACE) within samples, standard approaches cannot be applied to a target population with a different distribution from the experimental sample. This paper makes several key contributions. First, we introduce a new set of identifying assumptions in the form of a population-level exclusion restriction that allows for identification of the target complier average causal effect (T-CACE) in both randomized experiments and observational studies. This allows researchers to identify the T-CACE without relying on standard principal ignorability assumptions. Second, we propose a class of inverse-weighted estimators for the T-CACE and derive their asymptotic properties. Third, we introduce an optimization-based sensitivity analysis framework to assess the robustness of the estimator in the presence of unmeasured confounding. Our extensive simulations demonstrate that the proposed estimator yields low bias and achieves accurate coverage of confidence intervals. We illustrate our proposed method on a study evaluating the impact of deep canvassing on reducing exclusionary attitudes.

Generalizability/Transportability

Toward Personalized and Sample-bounded Meta-analyses Wenqi Shi* Wenqi Shi, José Zubizarreta,

Meta-analysis is a powerful tool for synthesizing evidence across multiple studies, often yielding more precise estimates than those of individual studies. However, traditional methodologies often lack the flexibility to tailor inferences to a well-defined target population and do not ensure that the final estimates remain sample-bounded. This paper presents a novel framework for personalized and sample-bounded meta-analyses, which extends traditional approaches by explicitly incorporating target covariate profiles and enforcing sample-boundedness, thereby enhancing both personalization and reliability. This framework includes variance estimation, allows for partial overlap between study and target populations, and supports both individual patient data (IPD) and aggregate-level data. Through personalization, this method mitigates selection bias and ensures that effect estimates apply to the intended population. By explicitly identifying study contributions and enforcing sample-boundedness, this method selects reliable study donors under appropriate assumptions, improves robustness to overlap violations, enhances transparency, and prevents extrapolation bias. We establish theoretical properties of this approach, including multiple consistency conditions and asymptotic normality, and provide diagnostics for covariate balance and study selection. Simulations demonstrate the robustness of this method to overlap violations and a case study highlights its practical application.

Goodness-of-fit

An impossibility result for evaluating the goodness-of-fit of causal models James Stratton*
James Stratton, Nicolaj Thor,

The coefficient of determination (R^2), and related measures of goodness-of-fit in a regression analysis, measure the share of variation in an outcome that can be predicted by an analyst observing a given set of independent variables. We ask whether there are corresponding measures of the share of variation causally explained by a given set of independent variables, under the assumption that the analyst has the ability to both observe and experiment with those variables. We argue that any valid measure of goodness-of-fit in a causal model should satisfy three axioms: monotonicity (the measure should not decrease as more variables are observed), completeness (the measure should equal 1 when all variables are observed), and limited information (the measure should be strictly less than 1 when the observed variables fail to fully explain variation in the outcome). Our main result is that these three requirements are incompatible. This incompatibility provides theoretical support for the view that causal inference should focus on parameter estimation rather than goodness-of-fit measurement. Motivated by this result, we develop tools for assessing the relative importance of independent variables in causing variation in any given outcome, by considering various weakenings of our axioms.

Heterogeneous Treatment Effects

The Heterogenous Treatment Effects of Compulsory Education Age Reforms – A Causal Forest Approach¥ Hannelore Nelissen* Hannelore Nelissen, Krist De Witte,

While existing literature often focuses on average treatment effects, there is limited understanding of how uniform policy measures, such as raising the compulsory education age, impact very specific subgroups of students. This paper addresses this gap by investigating the heterogeneous effects of a compulsory education age reform on school dropout rates. Using rich administrative microdata from Statistics Netherlands, we apply a causal forest model to estimate Conditional Average Treatment Effects (CATEs), revealing how policy impacts vary across individual and school characteristics. Our results show an average reduction of 1.06 percentage points in dropout rates attributable to the reform, with significant heterogeneity; approximately 29% of the estimated CATEs indicate statistically significant effects up to 4.6 percentage points. Vocational track students emerge as the most responsive group, with parental income, household composition, and school progress further influencing outcomes. Moreover, we find that only certain groups of at-risk students are most suitable targets for the policy reform, suggesting that a single policy may not address the needs of all students. The analysis advocates the need for complementary policies to better address diverse student needs. This study contributes to the literature by demonstrating the importance of nuanced, data-driven policy targeting to optimize educational outcomes across varied contexts.

Heterogeneous Treatment Effects

Randomization-Based Inference for Average Treatment Effects in Inexactly Matched Observational Studies Jianan Zhu* Jianan Zhu, Jeffrey Zhang, Zijian Guo, Siyu Heng,

Matching is a widely used causal inference study design in observational studies. Ideally, treated units are exactly matched with controls for the covariates, and randomization-based inference for the treatment effect can then be conducted as in a randomized experiment under the ignorability assumption. However, matching is typically inexact when continuous covariates or many covariates exist. Previous studies have routinely ignored inexact matching in the downstream randomization-based inference as long as some covariate balance criteria are satisfied. However, these inference methods focus on the constant treatment effect (i.e., Fisher's sharp null) and are not directly applicable to the average treatment effect (i.e., Neyman's weak null). To address this important gap, we propose a new framework — inverse post-matching probability weighting (IPPW) — for randomization-based inference for average treatment effects under inexact matching. Compared with the routinely used randomization-based inference framework based on the difference-in-means estimator for average treatment effects, our proposed IPPW framework can substantially reduce bias due to inexact matching and improve the coverage rate. Our framework can also be extended to the instrumental variable settings to simultaneously address the bias due to inexact matching and unmeasured confounding bias. We have also applied our framework to an observational study of kidney diseases among agricultural workers in Zimbabwe.

Heterogeneous Treatment Effects

Statistical Learning for Heterogeneous Treatment Effects: Pretraining, Prognosis, and Prediction Maximilian Schuessler* Maximilian Schuessler, Erik Sverdrup, Robert Tibshirani,

Robust estimation of heterogeneous treatment effects is a fundamental task for optimal decision-making in many applications from personalized medicine to educational policies. In recent years, predictive machine learning has emerged as a valuable toolbox for causal estimation, enabling more flexible and rigorous effect estimation. Despite these advances, robust conditional average treatment estimation (CATE) remains highly challenging, especially in settings with complex interactions, low signal or high dimensions. In this article, we propose a new pretraining strategy that leverages a phenomenon in real-world applications: factors that are prognostic of the outcome, are frequently also predictive of treatment effects. Drawing on the R-loss and the lasso, we develop a suite of refined model architectures of R-Learners that achieve lower error rates in settings with shared support between the mean outcome function and the treatment effect function. Intuitively, if factors associated with a good baseline prognosis are also predictive of high treatment effects, pretraining will improve the estimation of the CATE. This also offers a data-driven way to discover a potential overlap between prognostic and predictive factors. We extend this approach to nonlinear models, basis function expansion, and settings with right-censoring, which allows us to demonstrate the utility of this framework to a series of settings and medical applications.

Heterogeneous Treatment Effects

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

In our study, we introduce a novel estimator for the Conditional Average Treatment Effect (CATE), termed the parachuted estimator, notable for its double robustness. This means it remains consistent if either the propensity score model or the conditioned expected outcome model is correctly specified, and uniquely, it still performs reliably even if both models are misspecified, albeit at a slower convergence rate similar to non-parametric estimators. Our method combines parametric techniques from Augmented Direct Learning with non-parametric kernel estimation strategies, achieving optimal convergence rates and maintaining consistency.

A pivotal achievement of this research is the derivation of the asymptotic distribution for this hybrid estimator. This distribution is characterized by having its mean aligned with the estimated target—the true CATE—and its variance describable through a closed-form expression. Moreover, our investigation extends into the statistical inference concerning our parachuted estimator for CATE via established bootstrap techniques. Through rigorous theoretical analysis grounded in the work of Chatterjee and Bose (2005), we provide substantive proof affirming that these bootstrap methods yield consistent estimations within our specified framework, subject to certain regular constraints.

Heterogeneous Treatment Effects

Fisher–Rao Gradient Flows for Semiparametric Estimation of Parameters Lacking a Canonical Gradient Kaiwen HOU* Kaiwen HOU, Mark van der Laan,

We introduce a unified framework that bridges geometric gradient flow methods with semiparametric efficiency theory, specifically addressing parameters that lack a conventional canonical gradient. By generalizing the Jordan–Kinderlehrer–Otto scheme under the Hellinger distance, we equip the space of probability distributions with a Fisher–Rao geometry and construct a natural gradient flow for likelihood maximization. This approach characterizes the universal least favorable path in one-step TMLE as a Fisher–Rao gradient trajectory, thereby connecting the geometric steepest descent to classical one-step estimation. We establish well-posedness of the underlying infinite-dimensional PDE and demonstrate that the resulting one-step estimators attain asymptotic efficiency in both parametric and functional models. These results provide robust tools for efficient inference in semiparametric settings where standard efficiency theory does not apply.

Instrumental Variables

A Fully Stochastic Update to the Potential Outcome Framework: Never Too LATE Hanti Lin* Hanti Lin,

Dawid (2000) raises a philosophical objection to the potential outcome framework along with its application to the local average treatment effect (LATE). His concern is that this framework makes a very strong assumption, that every individual's potential outcomes are deterministic, which appears essential to the identification result about the LATE. I address this philosophical challenge by introducing a fully stochastic update to the potential outcome framework, which improves upon the partially stochastic account due to Small et al. (2017). Here is the idea: each individual's potential outcomes are first rendered fully stochastic, following Robins and Greenland (1991), and their probabilities—called potential probabilities—are then treated as parameters of an appropriate causal Bayes net (which comes with the causal Markov assumption). Everyone has a degree of compliance, defined as a difference between two potential probabilities; defiers are refined to be those having a negative degree of compliance. I prove that, in this fully stochastic setting, if there are no defiers, then the usual IV estimand identifies a new quantity. This new quantity reduces to the LATE in the deterministic setting, where each individual's potential probabilities are 0 or 1. I close by arguing that the proposed marriage between causal Bayes nets and the potential outcome framework is, in specific ways, superior to Pearl's (2009) nonparametric structural equation models.

Interference and Consistency Violations

Dyad-level Weighting Estimators for Causal Effects on Changes in Network Ties under Interference Qixiang Xu* Qixiang Xu, Laura Forastiere,

Public health interventions, designed to improve health outcomes, may also influence participants' social network structures. For example, in health education programs, participants who receive the intervention may become more popular and attract new social connections, while ties between untreated individuals may be weakened. We develop a novel causal inference framework to measure the effect of interventions on network change, where the treatments and outcomes are defined at the dyadic-level. We allow for dependence in the network formation between dyads by assuming the potential presence of interference from a pre-specified set of units, defined for each dyad and called interference set. Under this interference assumption, we define the direct effects of the treatment status of a dyad of units on the formation or dissolution of a directed tie between them as well as the spillover effects from the treatment of the whole interference set or a subset of it, such as common friends between two units, by changing or fixing the dyadic treatment while fixing or changing the distribution of the treatment in the interference set or in the subset of interest, respectively. We develop new Horvitz-Thomson and Hajek estimators for these network-based causal estimands and derive their asymptotic properties under dyadic-level data. We then apply our estimators to a two-stage randomized trial of a health education program in Honduras.

Interference and Consistency Violations

Doubly 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 effect identification. In this paper, we extend the notion of an exposure mapping to that of a generalized counterfactual mapping, in order to accommodate interference structures for nested counterfactuals. In particular, we allow for interference in exposure-outcome, exposure-mediator, and mediator-outcome relationships in a difference-in-differences setting. After establishing identification results, we further develop doubly robust, semi-parametric efficient estimators for 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 price changes.

Machine Learning and Causal Inference

Doubly Robust Inference on Causal Derivative Effects for Continuous Treatments Yikun Zhang* Yikun Zhang, Yen-Chi Chen,

Statistical methods for causal inference with continuous treatments mainly focus on estimating the mean potential outcome function, commonly known as the dose-response curve. However, it is often not the dose-response curve but its derivative function that signals the treatment effect. In this talk, we investigate nonparametric inference on the derivative of the dose-response curve with and without the positivity condition. Under the positivity and other regularity conditions, we propose a doubly robust (DR) inference method for estimating the derivative of the dose-response curve using kernel smoothing. When the positivity condition is violated, we demonstrate the inconsistency of conventional inverse probability weighting (IPW) and DR estimators, and introduce novel bias-corrected IPW and DR estimators. In all settings, our DR estimator achieves asymptotic normality at the standard nonparametric rate of convergence. Additionally, our approach reveals an interesting connection to nonparametric support and level set estimation problems.
Finally, we demonstrate the applicability of our proposed estimators through simulations and a case study of evaluating a job training program.

Machine Learning and Causal Inference

Structure-agnostic Optimality of Doubly Robust Learning for Treatment Effect Estimation
Jikai Jin* Jikai Jin, Vasilis Syrgkanis,

Average treatment effect estimation is the most central problem in causal inference with application to numerous disciplines. While many estimation strategies have been proposed in the literature, the statistical optimality of these methods has still remained an open area of investigation, especially in regimes where these methods do not achieve parametric rates. In this paper, we adopt the recently introduced structure-agnostic framework of statistical lower bounds, 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 prove the statistical optimality of the celebrated and widely used doubly robust estimators for both the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT), as well as weighted variants of the former, which arise in policy evaluation.

Machine Learning and Causal Inference

Berry-Esseen-Type Bound for Nonparametric Average Treatment Effect Estimator in Randomized Trials Hongxiang Qiu* Hongxiang Qiu,

We present Berry-Esseen-type bounds for Wald-confidence intervals (CIs) based on the augmented inverse probability weighted (AIPW) estimator of average treatment effect with black-box nuisance estimators, with or without cross-fitting, in a randomized controlled trial (RCT) setting with known propensity score. In this setting, it is well known that an almost arbitrary outcome regression estimator leads to a consistent asymptotically normal estimator, and a consistent outcome regression estimator leads to asymptotic efficiency. Our bound for the difference between true and nominal Wald-CI coverages vanishes at a rate related to the convergence rate of the outcome regression estimator to its limit, indicating a potential trade-off between an estimator's efficiency and the corresponding Wald-CI coverage. We also show that one possible reason for cross-fitting to improve statistical inference for AIPW estimators in RCTs, even if the Donsker condition for nuisance estimators holds, is the improved standard error estimator due to sample splitting.

Machine Learning and Causal Inference

Longitudinal Generalizations of the Average Treatment Effect on the Treated for Multi-valued and Continuous Treatments Herbert Susmann* Herbert Susmann, Nicholas Williams, Kara Rudolph, Iván Díaz,

The Average Treatment Effect on the Treated (ATT) is a common causal parameter defined as the average effect of a binary treatment among the subset of the population receiving treatment. We propose a novel family of parameters, Generalized ATTs (GATTs), that generalize the concept of the ATT to longitudinal data structures, multi-valued or continuous treatments, and conditioning on arbitrary treatment subsets. We provide a formal causal identification result that expresses the GATT in terms of sequential regressions, and derive the efficient influence function of the parameter, which defines its semi-parametric efficiency bound. Efficient semi-parametric inference of the GATT requires estimating the ratios of functions of conditional probabilities (or densities); we propose directly estimating these ratios via empirical loss minimization, drawing on the theory of Riesz representers. Simulations suggest that estimation of the density ratios using Riesz representation have better stability in finite samples. Lastly, we illustrate the use of our methods to evaluate the effect of chronic pain management strategies on the development of opioid use disorder among Medicare patients with chronic pain.

Machine Learning and Causal Inference

THE BIOPHILIA PREMIUM: VISUAL PLANT IMAGERY AND PROPERTY VALUATION Yuqian Chang* Yuqian Chang, Nathan Fong, Ning Ye, Maureen Morrin,

Homebuyers' reliance on real estate websites has increased the importance of optimizing a home's online profile. We investigate the influence of biophilic imagery (i.e., images depicting plant life) on property valuation, and demonstrate the biophilia premium with a large unstructured dataset of 40,294 sold homes with 569,828 digital images. We employ a three-step framework with multiple machine learning tools to attribute price differences to biophilic images as opposed to other correlated features of the listings, including 1) a highly accurate biophilic image classifier founded upon a two-stage deep learning algorithm (Faster R-CNN) to minimize measurement error, 2) assembling extensive covariates diagnosed as effective to address the most plausible confounders, and 3) using double machine learning to maximize the predictive power of the covariates. Homes with more biophilic images garnered more likes on the website and higher intent to visit, ultimately exhibiting transaction values 0.13% higher per biophilic image. We further reveal the heterogeneity of the effect identified, showing that the biophilia premium is stronger in areas with lower prevalence of plant imagery usage and higher property homogeneity. These areas tend to have lower socioeconomic status, property valuation, and image quality, suggesting resource constraints may be the reason behind the limited usage of biophilic imagery in certain markets, and thus providing direct implications to practitioners.

Machine Learning and Causal Inference

Isolated Causal Effects of Natural Language Victoria Lin* Victoria Lin, Louis-Philippe Morency, Eli Ben-Michael,

Recent advances in natural language processing have dramatically increased the availability of language data and models for common users. As language technologies become widespread, it is important to understand how changes in language affect reader perceptions and behaviors. For instance, as machine-generated text—including undesirable text like fake news and propaganda—proliferates in public spaces, we may wish to know whether misinformation propagated in these texts has impacts on readers' behaviors.

In this work, we formalize these impacts as the *isolated causal effect* of some *focal* language-encoded intervention on an external outcome. We show that a core challenge of estimating isolated effects is the need to approximate all *non-focal* language outside of the intervention. To address this challenge, we introduce a formal estimation framework for isolated causal effects of language and explore how different approximations of non-focal language influence effect estimates. Drawing on the principle of *omitted variable bias*, we present metrics for evaluating the quality of isolated effect estimation and non-focal language approximation along the axes of *fidelity* and *overlap*. In experiments on semi-synthetic and real-world data, we validate the ability of our framework to recover ground truth isolated effects, and we demonstrate the utility of our proposed metrics as measures of quality for both isolated effect estimates and non-focal language approximations.

Machine Learning and Causal Inference

Learning Counterfactual Distributions via Kernel Nearest Neighbors Kyuseong Choi*
Kyuseong Choi, Jacob Feitelberg, Caleb Chin, Anish Agarwal, Raaz Dwivedi,

Consider a setting with multiple units (e.g., individuals, cohorts, geographic locations) and outcomes (e.g., treatments, times, items), where the goal is to learn a multivariate distribution for each unit-outcome entry, such as the distribution of a user's weekly spend and engagement under a specific mobile app version. A common challenge is the prevalence of missing not at random data—observations are available only for certain unit-outcome combinations—where the missingness can be correlated with properties of distributions themselves, i.e., there is unobserved confounding. An additional challenge is that for any observed unit-outcome entry, we only have a finite number of samples from the underlying distribution. We tackle these two challenges by casting the problem into a novel distributional matrix completion framework and introduce a kernel-based distributional generalization of nearest neighbors to estimate the underlying distributions. By leveraging maximum mean discrepancies and a suitable factor model on the kernel mean embeddings of the underlying distributions, we establish consistent recovery of the underlying distributions even when data is missing not at random and positivity constraints are violated. Furthermore, we demonstrate that our nearest neighbors approach is robust to heteroscedastic noise, provided we have access to two or more measurements for the observed unit-outcome entries—a robustness not present in prior works on scalar nearest neighbors.

Machine Learning and Causal Inference

Just Trial Once: Ongoing Causal Validation of Updates to Machine Learning Models Jacob M Chen* Jacob M Chen, Michael Oberst,

The use of machine learning (ML) models as clinician support tools is increasing in popularity. Evaluating the causal impact of deploying such models on clinical outcomes can be done with a randomized control trial (RCT), such as a cluster-randomized trial. However, ML models are inevitably updated over time, and we often lack evidence for the impact of these updates on the same clinical outcomes. While this impact could be repeatedly validated with ongoing RCTs, in practice, such experiments are expensive, time-consuming, and difficult to run. In this work, we present an alternative solution: using only data from a prior RCT that tested other models, we give conditions under which the causal effect of an updated ML model can be precisely estimated or bounded. Our assumptions incorporate two realistic constraints: ML predictions are often deterministic, and their impact depends on clinician trust in the model. Based on our analysis, we give recommendations for cluster randomized trial designs that maximize their ability to assess future versions of an ML model. Our hope is that following our proposed trial design will save practitioners time and resources while allowing for quicker deployments of updates to ML models.

Machine Learning and Causal Inference

Advancing Real-World Evidence Studies with Multi-Agent LLM Systems for Causal Inference Rachael Phillips* Rachael Phillips, Tianyue Zhou, Mark van der Laan,

Large language models (LLMs) offer the potential to revolutionize the design and analysis of real-world evidence (RWE) studies, which leverage data from sources such as electronic health records, claims databases, and registries to inform clinical and regulatory decisions. Intuitive interfaces that engage users in natural language interactions, LLMs provide contextualized assistance and can elucidate trade-offs among different choices. In this work, we introduce a novel multi-agent LLM system tailored for RWE studies that is comprehensively grounded in causal inference frameworks and aligned with regulatory guidelines. We present applications where our customized co-pilot assists researchers in critical steps such as handling intercurrent events, defining interventions, and assessing causal identifiability. Comparative results from real-world examples illustrate the superiority of our co-pilot over proprietary solutions like ChatGPT, which often conflate critical study elements and muddle interpretations. By integrating domain-specific expertise and collaborative agent interactions, our system represents a step change in leveraging LLMs for causal inference. This presentation will conclude with a discussion of future directions and challenges, including ensuring responsible use, benchmarking performance, and fostering interdisciplinary collaboration to establish trust and governance in these AI-assisted methods.

Machine Learning and Causal Inference

Causal Inference and Adaptive Design for Evaluating Effectiveness of Medical Tests and Devices Wenxin Zhang* Wenxin Zhang, Rachael Phillips, Gene Pennello, Mark van der Laan,

Diagnostic medical devices, i.e., tests, play a critical role in detecting and monitoring diseases. However, unlike treatments, the effects of test results on health outcomes are indirect through their downstream influence on treatment decisions. This disconnection poses a challenge in evaluating the true effectiveness of tests. In this work, we propose causal estimands to evaluate effectiveness of medical tests from their explanatory and/or pragmatic utility in medical care, estimating them by the Targeted Maximum Likelihood Estimation (TMLE) method. We further propose an adaptive experiment design to better evaluate medical tests and devices. This framework is generally applicable to evaluate the effectiveness of any object (e.g. AI prediction score) whose effect on the primary outcome of interest is mediated by the consequent treatment decision. The performance of our estimators and designs is demonstrated through simulation studies.

Machine Learning and Causal Inference

Propensity Score Estimation for Generalized Treatment Spaces Alessandro Leite* Alessandro Leite, Felipe Lourenço Angelim Vieira,

Propensity scores comprise an effective strategy to control for biases when estimating causal effects based on observational data. Nevertheless, traditional methods often struggle with complex treatment structures beyond binary or discrete treatments. This work introduces a methodology for propensity score estimation that accommodates generalized treatment structures without imposing restrictive parametric assumptions. We follow a probabilistic classification model to estimate the stabilized weights and the propensity scores via density ratio estimation, making our method adaptable to various treatment forms. Experimental results show notable performance improvements in accuracy and stability across different synthetic and semi-synthetic datasets, even when the number of covariates increases. Such results suggest that our approach is flexible enough for challenging real-world applications when estimating causal effects such as healthcare, policy analysis, and personalized marketing.

Machine Learning and Causal Inference

Propensity Score Estimation for Generalized Treatment Spaces * Alessandro Leite, Felipe Lourenço Angelim Vieira, Alessandro Leite,

Propensity scores comprise an effective strategy to control for biases when estimating causal effects based on observational data. Nevertheless, traditional methods often struggle with complex treatment structures beyond binary or discrete treatments. This work introduces a methodology for propensity score estimation that accommodates generalized treatment structures without imposing restrictive parametric assumptions. We follow a probabilistic classification model to estimate the stabilized weights and the propensity scores via density ratio estimation, making our method adaptable to various treatment forms. Experimental results show notable performance improvements in accuracy and stability across different synthetic and semi-synthetic datasets, even when the number of covariates increases. Such results suggest that our approach is flexible enough for challenging real-world applications when estimating causal effects such as healthcare, policy analysis, and personalized marketing.

Machine Learning and Causal Inference

Higher-order estimators of time-varying effects in anisotropic smoothness models Matteo Bonvini* Matteo Bonvini, Edward H. Kennedy, Luke J. Keele,

The general theory of higher-order influence functions (HOIF) has been successfully applied to several pathwise differentiable parameters arising in causal inference, such as the expected conditional covariance and the treatment-specific mean. Such theory has yielded minimax optimal estimators in certain nonparametric models, e.g., those indexed by smooth nuisance parameters. More recently, minimax optimal, higher-order estimators have been derived for some non-pathwise differentiable causal parameters, an example of which is the conditional average treatment effect. In this work, we aim to extend the application of HOIF theory to causal parameters defined by a time-varying treatment. As a leading example, we consider the two-time point case g-formula functional in an anisotropic smoothness model where the nuisance functions can depend more smoothly on certain covariates. We also consider even more structured models, such as additive ones. In each setting, we design a higher-order estimator and calculate its bias and variance, and for some of them, we show that the convergence rates established are minimax optimal. We complement our theoretical findings with simulations and data analysis.

Machine Learning and Causal Inference

Causal Purification: A Causal Regularization of Latent Denoising Diffusion Models to Purify Against Spurious Adversaries Juan Castorena* Juan Castorena,

Adversarial attacks purposely deceive AI classification systems by introducing perceptually indiscernible adversarial noise, tricking classifiers into the wrong labels. Contrary to the common tendency that attacks have a general form (e.g., concentrated around high-frequency patterns), we show here that adversarial attacks can rather exploit any un-controlled spurious correlation available in the dataset. For this, we use the singular value decomposition (SVD) to represent image spectral composition and design spectral adaptive attacks (SAA's) designed to purposely create non-causal spurious correlations in specific band(s) to perturb the input. We further show that standard denoising diffusion models have difficulties in purifying some of these SAA's and propose a regularization by causal identification of a latent denoising diffusion model, operating in the singular value decomposition, that successfully removes this, and other general types of standard AA's. These findings, emphasizes the strength of regularization by causal identification of diffusion models as universal purification strategies that can be used as a pre-processing step to destroy adversarial attacks.

Matching

Comparing Coarsened Exact Matching (CEM) and Almost Matching Exactly (AME) in Real-World Settings: Insights from Cigna's Annual Value of Integration Study Aran Canes* Aran Canes,

Coarsened Exact Matching (CEM) allows for the comparison of treatment and control populations by leveraging subject matter expertise to identify confounders and determine appropriate intervals for matching. Despite its effectiveness in case/control studies, CEM often encounters issues in real-world applications. Including a variable that does not influence the outcome biases the results. The appropriate coarsening, determined by subject matter expertise, may not be accurate enough to avoid introducing confounder bias. Additionally, the selection of confounders is susceptible to errors in judgment.

Almost Matching Exactly (AME), developed at Duke University, offers alternatives to mitigate these issues. AME employs machine learning to identify relevant confounders and determine appropriate coarsening for continuous variables, reducing the need for subject matter expertise.

However, CEM has a distinctive advantage over AME. By using an outcome variable to determine binning, AME requires a unique match for each distinct outcome. In contrast, CEM can be applied once to a treatment/control population with multiple outcomes.

This paper utilizes Cigna's Annual Value of Integration Study to illustrate these advantages and disadvantages and their application in our 2024 study. By using real-world data, we aim to contribute to the causal inference community by demonstrating the utility of these methods in a non-academic setting.

Matching, Weighting

Evaluating time-specific treatment effects using randomization inference Sangjin Lee*
Sangjin Lee, Kwonsang Lee,

This study develops a systematic approach for evaluating the effect of a treatment on a time-to-event outcome in a matched-pair study. While most methods for paired right-censored outcomes allow determining an overall treatment effect over the course of follow-up, they generally lack in providing detailed insights into how the effect changes over time. To address this gap, we propose novel tests for paired right-censored outcomes using randomization inference. We further extend our tests to matched observational studies by developing corresponding sensitivity analysis methods to take into account departures from randomization. Simulations demonstrate the robustness of our approach against various non-proportional hazards alternatives, including a crossing survival curves scenario. We demonstrate the application of our methods using a matched observational study from the Korean Longitudinal Study of Aging (KLoSA) data, focusing on the effect of social engagement on survival.

Matching, Weighting

Inference for weighting-based effect estimators: Residualization produces smaller standard errors with correct coverage Arisa Sadeghpour* Arisa Sadeghpour, Erin Hartman, Chad Hazlett,

Balancing weight procedures are used in observational causal inference to adjust for covariate imbalance within the sample. Common practice for inference is to estimate robust standard errors from a weighted regression of outcome on treatment. However, it is well known that weighting can inflate variance estimates, sometimes significantly, leading to standard errors and confidence intervals that are overly conservative. Motivated by linearized standard errors from the survey literature, we instead propose using robust standard errors from a weighted regression that additionally includes the balancing covariates and their interactions with treatment. We show that these standard errors are more precise and asymptotically correct when balancing weights target exact balance, such as with entropy balancing. Gains to precision can be quite significant when the balancing weights adjust for prognostic covariates. For procedures that balance in expectation, such as inverse propensity weighting, our proposed method improves precision by reducing residuals through the parametric model. We also consider implications for other estimands such as the ATT and for approximate balancing weights. We demonstrate our approach through simulation and re-analysis of multiple empirical studies.

Matching, Weighting

High-Dimensional Matching with Genetic Algorithms Hajoung Lee* Hajoung Lee, Kwonsang Lee,

Matching in observational studies is widely used to estimate causal effects by obtaining treated and control groups with similar covariate distributions. Traditional matching methods rely on distances between observations to form pairs. However, this process often faces challenges in high-dimensional and low-sample size settings due to the curse of dimensionality, where the concentration of distances makes it difficult to distinguish between observations. To address this issue, we propose a novel matching method using genetic algorithms, shifting the focus from individual-level to group-level distances. By optimizing the similarity of the high-dimensional joint distributions of covariates between treated and control groups, our method improves causal effect estimation. This approach has key advantages: (1) it avoids dimension reduction, preserving the full scope of high-dimensional information without additional modeling, and (2) it maintains transparency by not relying on outcomes, akin to traditional matching, and (3) it performs robustly in low-sample size settings, where traditional methods may struggle. Furthermore, our results show that the proposed method is competitive with existing approaches even in low-dimensional settings. Through extensive simulations and real data applications, we validate the performance and provide practical guidance for the method, highlighting its potential as a powerful tool for causal inference in both high- and low-dimensional scenarios.

Matching, Weighting

Post-Stratification Using Bayesian Interpolation: Enhancing Causal Inference with Uncertainty-Aware Population Weighting Theo Snow* Theo Snow, Brittany Morgan Bustamante, Juliana Bartels, Simon Camponuri, Justin Remais, Alejandro Schuler, Alan Hubbard,

Accurate population inferences from stratified samples often require adjusting for biases and discrepancies between sample and population distributions, commonly through post-stratification weighting. This study investigates the limitations of deterministic post-stratification weighting and explores a Bayesian framework to address these challenges. We employ a Bayesian interpolation method to estimate post-stratification weights as posterior distributions. Using demographic data from 2013-2023 American Community Surveys and sample data from a national electronic health record system, we derive strata-specific posterior distributions and compute Bayesian weights. By specifying our priors as the expected counts of patients in each stratum if the sample had the distribution of our population of interest, we assume exchangeability between sample and population distributions while incorporating uncertainty. Bayesian weights closely resembled post-stratification weights (0.56% mean difference between the point estimates from each weighting method). However, Bayesian interpolation offers the addition of posterior confidence intervals, characterizing uncertainty in the weights. This provides a robust assessment of sampling variability and population uncertainty in weighted analyses. This work demonstrates the utility of Bayesian interpolation for causal inference by providing a principled method to control for biases while propagating uncertainty through the adjustment process.

Mediation

General targeted machine learning for modern causal mediation analysis Richard Liu*
Richard Liu, Nicholas Williams, Kara Rudolph, Iván Díaz,

Causal mediation analyses investigate the mechanisms through which causes exert their effects, and are therefore central to scientific progress. The literature on the non-parametric definition and identification of mediational effects in rigourous causal models has grown significantly in recent years. Despite great progress in the causal inference front, statistical methodology for non-parametric estimation has lagged behind, with few or no methods available for tackling non-parametric estimation in the presence of multiple, continuous, or high-dimensional mediators. In this paper we show that the identification formulas for six popular non-parametric approaches to mediation analysis proposed in recent years can be recovered from just two statistical estimands. We leverage this finding to propose an all-purpose one-step estimation algorithm that can be coupled with machine learning in any mediation study that uses any of these six definitions of mediation. The estimators have desirable properties, such as weak convergence and asymptotic normality. Estimating the first-order correction for the one-step estimator requires estimation of complex density ratios on the potentially high-dimensional mediators, a challenge that is solved using recent advancements in so-called Riesz learning.We illustrate the properties of our methods in a simulation study and real data to estimate the extent to which pain management practices mediate the total effect of having a chronic pain disorder.

Mediation Analysis, Mechanisms

Causal Mediation and Functional Outcome Analysis with Process Data Youmi Suk* Youmi Suk, Chan Park,

Over the past two decades, there has been growing interest in analyzing the effects of educational programs on outcomes using process data from computer-based testing and learning environments. However, most analyses focus on final outcomes measured at the end of a test or session, overlooking their functional nature over time. Such analyses fail to capture the dynamic causal mechanisms associated with functional outcomes. To address this limitation, this paper proposes a novel causal framework for identifying and estimating functional average treatment effects, functional natural direct effects, and functional natural indirect effects, along with their subgroup effects. Building on the literature on causal mediation and moderation, we define these effects using potential outcomes and provide nonparametric identification strategies. We then develop estimation methods using generalized additive models, a flexible and robust tool for analyzing functional data. The proposed approach is applied to examine the effects of extended time accommodations (ETA) on two functional outcomes—test scores and item access—in large-scale educational process data. In this analysis, students' disability status serves as a moderator. This application uncovers the dynamic causal mechanisms underlying the effects of ETA on outcomes and highlights when and for whom each effect works during the testing period.

Mediation Analysis, Mechanisms

Two-Stage Nuisance Function Estimation for Causal Mediation Analysis Chang Liu* Chang Liu, AmirEmad Ghassami,

When estimating the direct and indirect causal effects using the influence function-based estimator of the mediation functional, it is crucial to understand what aspects of the treatment, the mediator, and the outcome mean mechanisms should be focused on. Specifically, considering them as nuisance functions and attempting to fit these nuisance functions as accurate as possible is not necessarily the best approach to take. In this work, we propose a two-stage estimation strategy for the nuisance functions that estimates the nuisance functions based on the role they play in the structure of the bias of the influence function-based estimator of the mediation functional. We use the weighted balancing approach of Imai and Ratkovic (2014) to design the estimator of the treatment mechanism in Stage 1 and one of the nuisance functions in Stage 2. The weights in these balancing estimators are designed directly based on the bias of the final estimator of the parameter of interest. We provide parametric and nonparametric versions of the balancing estimators. The other two nuisance functions are obtained using standard parametric or nonparametric regressions. We provide robustness analysis of the proposed method, as well as sufficient conditions for consistency and asymptotic normality of the estimator of the parameter of interest. We evaluate our methods through simulations and compare with existing methods.

Mediation Analysis, Mechanisms

A Causal Mediation Model for Continuous Time Markov State Processes with application to Microbiome Data Debarghya Nandi* Debarghya Nandi, Soumya Sahu,

Chronic stress affects millions globally and is linked to elevated cortisol levels, which regulate essential functions like immune response, metabolism, and inflammation. Prolonged cortisol elevation disrupts systems, including the vaginal microbiome, vital for women's reproductive health. Categorized into community state types (CSTs), this microbiome can shift due to factors like hormonal imbalances or stress. Research indicates stress-induced cortisol destabilizes CSTs, triggering transitions from healthy to dysbiotic states linked to adverse outcomes such as bacterial vaginosis (BV). This study investigates how stress and cortisol influence CST transitions, aiming to understand stress-induced dysbiosis and its implications for reproductive health.
We developed a longitudinal causal mediation model with stress as the exposure, cortisol as the mediator, and CST transitions as the outcome. Our approach models the probability of transitioning between CSTs, estimating shifts from healthy to dysbiotic states based on stress and cortisol over time. Using a joint modeling framework, we capture correlations between cortisol dynamics and CST transitions, defining direct effects (stress impact independent of cortisol) and indirect effects (mediated via cortisol).Extensive simulations demonstrated high performance in terms of bias and coverage, validating the model for real-world application. We plan to apply it to our microbiome dataset to uncover the causal pathways.

Mediation Analysis, Mechanisms

Causal Mediation Analysis for Survival Endpoints in Longitudinal Settings with Irregular observations and Informative Cluster Size Jun Lu* Jun Lu, Ming Wang, Sanjib Basu,

Mediation analysis uncovers pathways through which an exposure A influences an outcome Y via intermediate factors M. While counterfactual-based approaches have advanced causal mediation analysis in non-longitudinal settings, real-world studies increasingly involve long-term follow-ups with irregular data structures, particularly in time-to-event outcomes. These studies reveal that exposures often act through dynamic mediator processes, rather than isolated mediator snapshots. For instance, the NACC study tracks MRI measures over time to explore how exposure of interest influences Alzheimer's disease risk through irregular MRI measures. Recent methods, including functional principal component analysis and growth curve models, address irregularities but often overlook the multi-dimensional effects of mediator processes, such as mean values, variability, trends, and informative observation times tied to patient risk profiles. We propose a latent class mediation model for irregular longitudinal data with survival endpoints. Our approach captures the mediator process pattern and models the exposure-mediator-outcome relationships jointly. Estimation and inference are conducted using Bayesian methods, with sensitivity analyses addressing time-varying confounding. We demonstrate our method through simulations and an application to the NACC dataset.

Mediation Analysis, Mechanisms

The Blessings of Multiple Mediators: Removing Unmeasured Confounding Bias via Factor Analysis
Kan Chen* Kan Chen, Ruoyu Wang, Zhonghua Liu, Xihong Lin,

Multiple mediation analysis aims to evaluate the indirect effect of an exposure on outcomes through mediators, as well as the direct effect through other pathways. Traditional methods for estimating mediation effects require the strong assumption of no unmeasured confounding between the outcome and the set of mediators. However, when the exposure and mediators are not randomized, unmeasured confounding among the exposure, mediators, and outcome can lead to biased estimates. In this work, we introduce a novel framework called FAMA (Factor Analysis-based Mediation Analysis) to address unmeasured confounding in multiple mediation analysis within a linear model setting. FAMA combines an omitted-variable bias approach with factor analysis to estimate natural indirect effects in the presence of unmeasured confounders. We validate the framework through theoretical analysis and simulation studies, demonstrating its effectiveness and robustness. Additionally, we applied FAMA to data from the U.S. Department of Veterans Affairs Normative Aging Study to detect DNA methylation CpG sites that mediate the effect of smoking on lung function. Our analysis identified multiple DNA methylation CpG sites that may mediate the effect of smoking on lung function and robust to unmeasured confounding bias. Notably, we observed effect sizes ranging from –0.18 to –0.79, with a false discovery rate controlled at 0.05. This includes CpG sites in the genes AHRR and F2RL3 in the presence of unmeasured bias

Missing Data

A Complete Multiple Imputation Algorithm for Missing Data Graphs Trung Phung* Trung Phung, Ilya Shpitser, Rohit Bhattacharya,

Imputation is one of the most popular methods for analyzing data with missing values. However, the most widely used methods, such as Multiple Imputation with Chained Equations (MICE), operate under the Missing At Random (MAR) assumption, which may be incorrect in many real-world settings. Recently, much progress has been made in identification theory for Missing Not At Random (MNAR) models that can be represented graphically—missing data graphs provide an intuitive causal interpretation of missingness mechanisms and a concise representation of the statistical model. These results, however, have seen limited use in practice, in part due to the complexity of the identifying functionals for the propensity score and the existence of only a few bespoke estimation strategies. We remedy this issue by proposing a new imputation method that can be applied to any missing data graphical model whose full data law is identified. The algorithm is recursive—imputation for a data row uses all other rows whose missing variables are subsets of the current one, which is a direct consequence of the sound and complete identification theory for the full law. In contrast, MICE treats all rows as equal while performing Gibbs sampling due to its MAR assumption. We further show how computational and statistical efficiency of our method can be improved by employing graph sparsity. We evaluate our method against MICE, showing comparable results under MAR and superior, less biased results under MNAR.

Multilevel Causal Inference

Examining Longitudinal Treatment Effects: Contrasts of Treatment Regimes Accounting for Time-varying Confounders and Clustering Hanna Kim* Hanna Kim, Jee-Seon Kim,

In longitudinal social science research, treatments offered at different time points often share a common objective but vary in implementation. When individuals follow different treatment patterns, the effects can be conceptualized as contrasts of potential outcomes under static treatment regimes. Studying these causal estimands provides evidence on the effectiveness of competing strategies within a unified framework.

This study evaluates methods for estimating static treatment regime effects, addressing key challenges in social science applications: time-varying confounders and clustered data. For example, the impact of participating in Head Start from ages three to four on children's vocabulary development may depend on intermediate vocabulary levels and cluster-specific variations in curricula. We compare longitudinal inverse probability of treatment weighting (l-IPTW), g-computation, and targeted maximum likelihood estimation (TMLE), adapting models to include cluster fixed effects and robust standard errors. Additionally, we integrate TMLE with the SuperLearner algorithm for flexible functional forms.

Results from a real data analysis of Head Start effectiveness and a small-scale simulation study highlight differences among approaches and provide practical guidance for identifying and interpreting longitudinal treatment effects under static regimes, a novel framework in social science research.

Policy Learning

Policy Learning through Cooperative Bargaining Eli Ben-Michael* Eli Ben-Michael,

Algorithmic decision making is increasing in importance in high stakes settings. Classical methods for data driven decision making and individualized treatment rules take a utilitarian, top down approach; if one decision leads to a higher expected utility than all others, then that decision is optimal. Under this approach if few individuals benefit greatly, that can outweigh many individuals being slightly harmed. In this paper, we consider learning policies with a bottom up approach, using ideas from public choice theory. We learn policies using an objective that finds a Nash equilibrium, meaning that any departure from the policy would lead to a worse outcome for some individuals. We show that the solution is a randomized policy that assigns an action according to the proportion of individuals that benefit from that action, irrespective of the scale of the benefit. For binary decisions, the objective results in a logistic loss function for a model that predicts wether or not individual treatment effects are positive. However, individual treatment effects are not identifiable and so it is not possible to estimate such policies even with unlimited data. To address this, we partially identify the objective function and derive the maximin and minimax optimal policies, and show how to estimate such policies empirically with data using flexible models for the treatment effect. We characterize the properties of such policies and demonstrate this approach via applications

Randomized Designs and Analyses

Randomization Tests for Distributions of Individual Treatment Effects using Multiple Rank Statistics Jake Bowers* David Kim, Yongchang Su, Jake Bowers, Xinran Li,

In this paper we study quantiles of individual treatment effect. Recent developments on randomization-based approaches provides finite-sample valid inference for quantiles in both completely randomized and stratified randomized experiments. However, since previous methods are using Stephenson's Rank Statistic, where the most powerful hyperparameter choice is not fixed, there exists a burden of choices to use them. We propose combining polynomial rank sum statistic to enhance the power of existing approaches. Combining rank scores offers much more power to detect rare treatment effects or common treatment effects in long-tailed outcomes than single rank scores-based testing procedures such as the Wilcoxon Rank test and the Stephenson's Rank Test. Moreover, using Polynomial functions of rank scores further increases power and eases application of this approach from completely randomized experiments to strata-randomized experiments.

Randomized Designs and Analyses

Modern causal inference approaches to improve power for subgroup analysis in randomized clinical trials Antonio D'Alessandro* Antonio D'Alessandro, Michele Santacatterina, Samrachana Adhikari, Jiyu Kim, Falco Bargagli-Stoffi, Donald Goff,

In randomized clinical trials (RCTs), subgroup analysis is often planned to evaluate the heterogeneity of treatment effects within pre-specified subgroups of interest. However, these analyses frequently have smaller sample sizes, reducing the power to detect heterogeneous effects. A way to increase power is borrowing external data from similar RCTs or observational studies. In this project, we target the conditional average treatment effect (CATE) in the original RCT as the estimand of interest, provide identification assumptions, and propose a doubly robust estimator that uses machine learning and nonparametric Bayesian techniques. Borrowing data, however, may present the additional challenge of practical violations of the positivity assumption—the conditional probability of receiving treatment in the external data source may be small, leading to large inverse weights and erroneous inferences—thus negating the potential power gains from borrowing external data. To overcome this challenge, we also propose a covariate balancing approach, an automated debiased machine learning (DML) estimator, and a calibrated DML estimator. We show improved power in various simulations and offer practical recommendations for the application of the proposed methods. Finally, we apply them to evaluate the effectiveness of citalopram for negative symptoms in first-episode schizophrenia patients across subgroups defined by duration of untreated psychosis, using data from two RCTs and an observatio

Randomized Designs and Analyses

On the nonparametric identification of the proportion of always survivors Veronica Ballerini*
Veronica Ballerini, Alessandra Mattei, Fabrizia Mealli,

We often aim to evaluate the treatment effect on an outcome variable that may be censored by a time-to-event intermediate variable, e.g., the effect of a new therapy on patients' quality of life two years after treatment; such evaluation is possible only if the patients are still alive at the time of measurement. A widely used approach in such scenarios is the principal stratification, which makes it possible to define the Survivor Average Causal Effect (SACE) estimand—i.e., a contrast between potential outcomes for the subgroup of individuals who would have survived under both treatment and control. Recently, time-varying SACE-like estimands have been proposed, focusing on the subgroup of those who would have survived at least until a given time t. For identification, the monotonicity assumption is typically invoked. In this work, we prove that it is always possible to nonparametrically identify the proportion of "always survivors" for at least one time t—and, in continuous time, for an infinite number of times in a given interval—under minimal assumptions and relaxing monotonicity. Specifically, it is possible to prove the identifiability in the time interval bounded by the minima of the potential intermediate time-to-event variables, simply assuming these minima differ. We apply this theoretical framework to a case of safety assessment in RCTs to highlight its practical relevance. Furthermore, we introduce a time-varying estimand for safety analysis in this context.

Randomized Designs and Analyses

Bayesian Estimation of the Survivor Average Causal Effect for Cluster-Randomized Crossover Trials Dane Isenberg* Dane Isenberg, Nandita Mitra, Fan Li, Michael Harhay,

In cluster-randomized crossover (CRXO) trials, groups of individuals are assigned to one of two sequences of alternating treatments. Since clusters act as their own control, the CRXO design is typically more statistically efficient than the usual parallel-arm cluster-randomized trial. CRXO trials are increasingly popular in critical care studies where the number of available clusters is generally limited. In trials among severely ill patients, researchers often want to assess the effect of treatments on secondary non-terminal outcomes, but there may be several patients who do not survive to have these measurements fully recorded. To this end, we provide a causal inference framework for addressing truncation by death in the setting of CRXO trials. We target the survivor average causal effect (SACE) estimand, a well-defined subgroup treatment effect represented via principal stratification. We propose structural and standard modeling assumptions to enable SACE identification and estimation within a Bayesian paradigm. We evaluate the small-sample performance of our proposed Bayesian approach for estimation of SACE using CRXO trial data through a simulation study. We apply our methods to a two-period cross-sectional CRXO study examining the impact of proton pump inhibitors as compared to histamine-2 receptor blockers on certain non-mortality outcomes among adults requiring invasive mechanical ventilation.

Sensitivity Analysis

Robustness of Proximal Inference TBD TBD* Cory McCartan, Melody Huang,

Proximal inference has been proposed as an alternative identification approach to relaxing traditional selection-on-observables (SOO) assumptions (i.e., no unobserved confounding) in observational causal inference. Instead of assuming researchers measure all relevant confounders, proximal inference assumes researchers have access to two informative proxies: a treatment proxy and an outcome proxy, which satisfy certain conditional independence assumptions. We formalize the trade-offs made between using a traditional SOO identification strategy in contrast to the proximal assumptions and derive the necessary scope conditions for proximal inference to provide more robust estimates than SOO. We consider the realistic setting in which both SOO and proximal assumptions are violated, finding that under even small violations of selection-on-observables, small violations in the exclusion restriction can amplify the resulting bias from proximal inference. We extend classical results from the instrumental variables literature to the proximal inference setting, and find that weak proxies can exacerbate both efficiency loss and potential bias. We compare the different approaches on a re-analysis of the impact of vote share shifts on legislative behavior.

Sensitivity Analysis

Quantifying the Robustness of Inferences to Replacement of Data: Applications to Main Effects and Moderators Kenneth Frank* Kenneth Frank, Ran Xu, Qinyun Lin, Spiro Maroulis,

One of the most important factors affecting the use of evidence for policy or practice is uncertainty of study results. Conventional and new methods attempt to mitigate against sources of uncertainty due to sampling error, systematic bias, or heterogeneous treatment effects. Here we acknowledge that while there have been improvements, there will always be uncertainty regarding an inference. The issue then is to discuss inferences in clear precise terms. Therefore, we characterize uncertainty by quantifying how much the data would have to change to nullify an inference. Specifically, we present the Robustness of Inference to Replacement (RIR) and extend it to interaction or moderating effects. To nullify the inference of an effect of the Early Vocabulary Tier 2 Intervention (EVI) from Coyne et al.'s randomized experiment, one would expect to have to replace 88% (roughly 1264) cases with children for whom the intervention had no effect. To nullify the inference of an interaction of the intervention with baseline vocabulary, a special concern for Tier 2 literacy interventions, one would have to replace 37% (about 528) cases with cases for which the effect of the intervention did not depend on baseline vocabulary. We interpret these sensitivity analyses relative to the strengths and weaknesses of study design for estimating main and interaction effects.

Sensitivity Analysis

A Unified Approach for Assessing Sensitivity to Violations of Causal Assumptions Guilherme Jardim Duarte* Guilherme Duarte,

This paper introduces a general method for sensitivity analysis to assess the robustness of causal estimates when key assumptions are violated. Unlike prior sensitivity approaches that are developed on a case-by-case basis, it presents a unified framework that accommodates a wide variety of assumption types, including (1) functional-form violations, e.g. the presence of defiers in studies that assume monotonicity; (2) exclusion-restriction violations, e.g. an encouragement that directly affects the outcome in instrumental-variable studies; and (3) unconfoundedness violations, e.g. unobserved common causes in selection-on-observables studies. The key innovation is to allow for an assumption to be violated in a proportion of the data that is at most t; at t=0, the approach recovers the original estimates, and t=1 is equivalent to discarding the assumption entirely. By using recent developments in partial identification, the method derives sharp bounds representing the range of possible conclusions for a given t; if the best- and worst-case conclusions both have the same sign (e.g., both positive), then the original estimate is said to be t-robust. By varying t, researchers can determine the severity of the violation needed to reverse their causal estimates, expressed in a form that is easy to reason about with domain expertise. This method offers a widely applicable, easy-to-interpret tool for examining the implications of assumption violations on causal conclusions.

Sensitivity Analysis

Nonparametric Sensitivity Analysis for Unobserved Confounding with Survival Outcomes
Rui Hu* Rui Hu, Ted Westling,

In observational studies, the observed association between an exposure and outcome of interest may be distorted by unobserved confounding. Causal sensitivity analysis is often used to assess the robustness of observed associations to potential unobserved confounding. For time-to-event outcomes, existing sensitivity analysis methods rely on parametric assumptions on the structure of the unobserved confounders and Cox proportional hazards models for the outcome regression. If these assumptions fail to hold, it is unclear whether the conclusions of the sensitivity analysis remain valid. Additionally, causal interpretation of the hazard ratio is challenging. To address these limitations, in this paper we develop a nonparametric sensitivity analysis framework for time-to-event data. Specifically, we derive nonparametric bounds for the difference between the observed and counterfactual survival curves and propose estimators and inference for these bounds using semiparametric efficiency theory. We also provide nonparametric bounds and inference for the difference between the observed and counterfactual restricted mean survival times. We demonstrate the performance of our proposed methods using numerical studies and an analysis of the causal effect of physical activity on respiratory disease mortality among former smokers.

Sensitivity Analysis

Controlling the False Discovery Proportion in Observational Studies with Hidden Bias Mengqi Lin* Mengqi Lin, Colin Fogarty,

We propose an approach to exploratory data analysis in matched observational studies that investigates multiple outcomes while controlling for false discoveries and potential unmeasured confounding. For any candidate set of rejected hypotheses, our method provides sensitivity sets for the false discovery proportion, the proportion of rejected hypotheses that are actually true. For a set R containing |R| outcomes, the method describes how much unmeasured confounding would need to exist for us to believe that the proportion of true hypotheses is 0/|R|, 1/|R|,..., all the way to |R|/|R|. Moreover, the confidence statements are valid simultaneously over all possible choices of the set, allowing the researcher to look in an ad hoc manner for promising subsets of outcomes that maintain a large estimated fraction of true discoveries even if a large degree of unmeasured confounding is present. The approach is particularly well suited to sensitivity analysis, as conclusions that some fraction of outcomes were affected by the treatment exhibit larger robustness to unmeasured confounding than the conclusion that any particular outcome was affected. In principle, the method involves solving a series of quadratically constrained integer programs, yet we show that solutions can be found within reasonable time and that, in large samples, one can often bypass the integer program altogether. We illustrate the practical utility of the method through simulation studies and a data example.