Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

UNIVERSITY OF HAMBURG
BUSINESS SCHOOL

Causal Estimation and Predictive Uncertainty: Essays on Robust Statistical Learning

Cumulative Dissertation

to obtain the academic degree of a
Doctor rerum oeconomicarum (Dr. rer. oec.)
according to doctoral degree regulations 2024

at the University of Hamburg Business School
Moorweidenstr. 18
20148 Hamburg
Germany

submitted by: Jan Thilo Rabenseifner
born on September 08th 1994 in Stuttgart

Hamburg, February 19th 2026

Thesis Committee

Chair: Prof. Dr. Knut Haase
First Examiner: Prof. Dr. Michael Merz
Second Examiner: Prof. Dr. Martin Spindler
Third Examiner: Prof. Dr. Alexander Szimayer
Date of Disputation: 09.04.2026

Acknowledgments

To start my PhD under Michael Merz and Martin Spindler, I left my small study town Constance and moved almost all the way up north. I have never regretted this decision. Both Martin and Michael have been excellent advisors on both a personal and academic level. I want to thank Michael for teaching me to be meticulous about notation and details, for showing me how to remain creative and passionate about teaching – even if it is not always the most rewarding aspect of our field – and for providing me with the right tools and insights to complete my last project. I want to thank Martin for being a constant source of ideas, for sharing his excellent research network, and for providing me, at just the right moments, the necessary push to finish a project. Together, they made it possible for me to attend many conferences and summer schools, even in times when it was financially challenging.

I want to thank my wonderful co-authors Sven Klaaßen, Philipp Bach, and Jannis Kück. They have been a great team on many projects, and I am grateful that they welcomed me into their circle with open arms. I learned more at Sven's whiteboard than in many of the courses I have taken. All three were great mentors and did not hold back in sharing their knowledge on how to succeed in the academic world.

I want to express my special gratitude to Cathy Yi-Hsuan Chen, who took me under her wing and allowed me to join a challenging yet exciting project with Martin Spindler and Victor Chernozhukov. This experience enabled me to expand and deepen my methodological toolkit. Thank you also for inviting me to Edinburgh to present our work.

Through the advantage of shared supervision, I was also fortunate to have two excellent support groups that provided feedback whenever needed, offered well-timed breaks from work, and gave me the comforting feeling that I was never alone in my struggles. My great colleagues are Nha-Nghi de la Cruz, Valerian Fourel, Lucas Gomes, Julius Herzig, Marie Hielscher, Max Lüdecke, Oliver Schacht, Gangli Tan, Jan Teichert-Kluge, Erik Wendt, and Zihao Yuan. A big thanks also goes to Ira Widderich, who gave me my first city tour of Hamburg, always looked after everyone at our chair, and helped me with all organizational requirements.

I also want to thank my friends for providing all the necessary distractions and for asking just the right number of times when I would finally be finished – and I know most of you will probably never read this work.

My parents and my sister Katja believed in me in every possible way and supported me unconditionally. Thank you for always being there for me. I also want to thank my parents, in particular, for financing many of my adventures during my academic journey.

And most of all, I want to thank Anne Kümmerle. Anne, I dragged you with me into this roller coaster of extremes. You suffered with me during the most stressful moments and celebrated my greatest successes by my side. Even though I spent many hours in front of a screen, you were always first and foremost in my mind.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

1.1 Motivation

Modern statistical learning offers a rich toolkit for both causal inference and predictive modeling. Double machine learning (Chernozhukov et al. 2018) provides a principled framework for estimating treatment effects with high-dimensional nuisance parameters; conformal prediction (Gammerman et al. 1998, Vovk et al. 2022) delivers distribution-free prediction intervals with finite-sample coverage guarantees. These frameworks are powerful in their generality. Yet generality often comes at the cost of applicability: the assumptions that make a method elegant in theory can become restrictive, or even fail, when confronted with the structure of a specific empirical problem.

This dissertation is motivated by a recurring pattern. A researcher identifies a substantive question – the marginal effect of one treatment among many, the reliability of a propensity score estimator in small sample sizes, or the uncertainty of a claims frequency prediction – and turns to modern statistical learning for an answer. The relevant framework exists, but it does not fit the problem as stated. The treatment space is combinatorial rather than binary. The propensity scores are miscalibrated, and standard corrections destabilize under small samples. The prediction intervals achieve nominal coverage on average but systematically fail for the minority of observations that matter most. In each case, the gap between the framework and application demands methodological work: not a new paradigm, but a careful extension, adaptation, or diagnostic that makes the existing paradigm operational.

What unifies the three essays collected here is a concern with *estimation robustness*: whether a statistical procedure delivers reliable results not just under ideal conditions, but under the conditions actually encountered in applied work. The first two essays address robustness in the context of causal estimation with double machine learning; the third addresses robustness in the context of predictive uncertainty with conformal prediction.

1.2 Causal Inference with Machine Learning

Causal inference from observational data requires both identification and estimation. Identification establishes the conditions, typically unconfoundedness and overlap, under which a causal parameter is recoverable from the observed data distribution (Rosenbaum and Rubin 1983, Imbens and Rubin 2015). Estimation translates these conditions into a computable quantity, often involving nuisance parameters such as conditional expectations or

propensity scores that must be learned from data.

The double machine learning (DML) framework of Chernozhukov et al. (2018) provides a general approach to this estimation problem. By combining Neyman-orthogonal score functions with cross-fitting, DML yields $\sqrt{n}$-consistent and asymptotically normal estimators for low-dimensional causal parameters, even when nuisance functions are estimated with flexible machine learning methods. The key insight is that orthogonality renders the target estimator insensitive to first-order errors in nuisance estimation, so the regularization bias inherent in machine learning does not propagate into the causal estimate.

This framework has been remarkably successful and has been extended along multiple dimensions in recent years: to different treatment types including continuous (Colangelo and Lee 2023), difference-in-differences (Chang 2020), and dynamic settings (Lewis and Syrgkanis 2021); to different model classes such as regression discontinuity (Noack et al. 2024) and automatic debiased estimation (Chernozhukov et al. 2022b); and to practical considerations including sensitivity analysis (Chernozhukov et al. 2022a), clustered data (Chiang et al. 2022), hyperparameter tuning (Bach et al. 2024b), and software implementation (Bach et al. 2022, 2024a). Figure 1.1 provides a selective overview. Yet, the canonical DML setting – a single binary treatment, a well-specified propensity score, and sufficient overlap – remains the point of departure for most applications. Two gaps in this landscape motivate the first two essays of this dissertation.

From binary to combinatorial treatments. In many empirical settings, the outcome of interest is shaped not by a single intervention but by the joint configuration of several treatments. Examples range from the effect of lifestyle factors on blood pressure, where smoking, obesity, physical inactivity, and other conditions may interact in complex ways, to multi-item promotional bundles in retail marketing or multi-drug regimens in medicine. The causal attribution problem, which isolates the marginal contribution of a single treatment when the background configuration of other treatments matters, is fundamentally harder than in the binary case. The treatment space grows exponentially (2^d configurations for d binary treatments), overlap violations become structural rather than incidental, and the relevant estimand must be carefully defined to remain identifiable. While recent work has addressed multiple treatments through propensity-based approaches (Imbens 2000, Feng et al. 2012, Li and Li 2019), representation learning (Zhou et al. 2023), and interaction-specific DML (Xiang et al. 2025), a general framework for marginal attribution under combinatorial treatments within DML has been missing.

From well-specified to calibrated propensity scores. Even in the standard single-treatment setting, the practical performance of DML depends critically on the quality of nuisance function estimates. Propensity scores estimated by flexible machine learning methods such as gradient boosting or random forests may be poorly calibrated: the predicted probability of treatment may systematically deviate from the true conditional probability. This miscalibration is particularly consequential for inverse-propensity-weighted estimators, where extreme or distorted weights can inflate variance or introduce bias (Tan 2017, Deshpande and Kuleshov 2023, Gutman et al. 2024). The problem is exacerbated in small samples, under limited overlap, or with unbalanced treatment assignment – precisely

the settings where calibration would be most valuable, but where standard calibration procedures themselves become unstable (van der Laan et al. 2024a, Ballinari 2024). A systematic evaluation of how calibration methods interact with sample-splitting schemes and base learners within DML has been lacking.

1.3 Predictive Uncertainty for Imbalanced Count Data

Complementing the causal estimation perspective, the third essay addresses a different but related challenge: quantifying the uncertainty of point predictions. In many applied domains, a point forecast alone is insufficient. Insurance pricing requires prediction intervals to assess the range of plausible claims; risk management demands honest communication of forecast reliability; regulatory frameworks increasingly require uncertainty quantification alongside point estimates.

Conformal prediction (Vovk et al. 2022) offers a distribution-free framework for constructing prediction sets with finite-sample coverage guarantees, requiring only the assumption of exchangeability. Unlike parametric prediction intervals that may undercover when distributional assumptions fail, or bootstrap methods that provide only asymptotic guarantees, conformal prediction delivers mathematically rigorous coverage regardless of the underlying data distribution or the complexity of the predictive model. The framework has been extended to conformalized quantile regression (Romano et al. 2019), adaptive inference under distribution shift (Gibbs and Candès 2021), classification (Romano et al. 2020), and survival analysis (Candès et al. 2023).

Yet this generality can be deceptive when applied to imbalanced count data. In settings with severe zero inflation, the typical structure of insurance claims, where most policyholders file no claims, standard conformal methods achieve their marginal coverage guarantee by *over-covering the majority class* and *under-covering the minority* that matters most for actuarial applications (Tsoumas and Papadopoulos 2024). The prediction intervals are formally valid but practically misleading: they provide no useful uncertainty information for claimants, precisely in the cases of greatest interest.

1.4 Outline and Contributions

This dissertation consists of three self-contained essays. The first two operate within the double machine learning framework, addressing the gaps identified in Section 1.2. The third addresses predictive uncertainty through conformal prediction, motivated by the challenges described in Section 1.3. Figure 1.1 situates the causal estimation contributions within the broader DML landscape.

Chapter 2: Inference on Multiple Treatment Effects. Joint work with Cathy Yi-Hsuan Chen, Victor Chernozhukov, and Martin Spindler. This essay extends DML to combinatorial treatment regimes where d binary treatments interact in 2^d configurations.

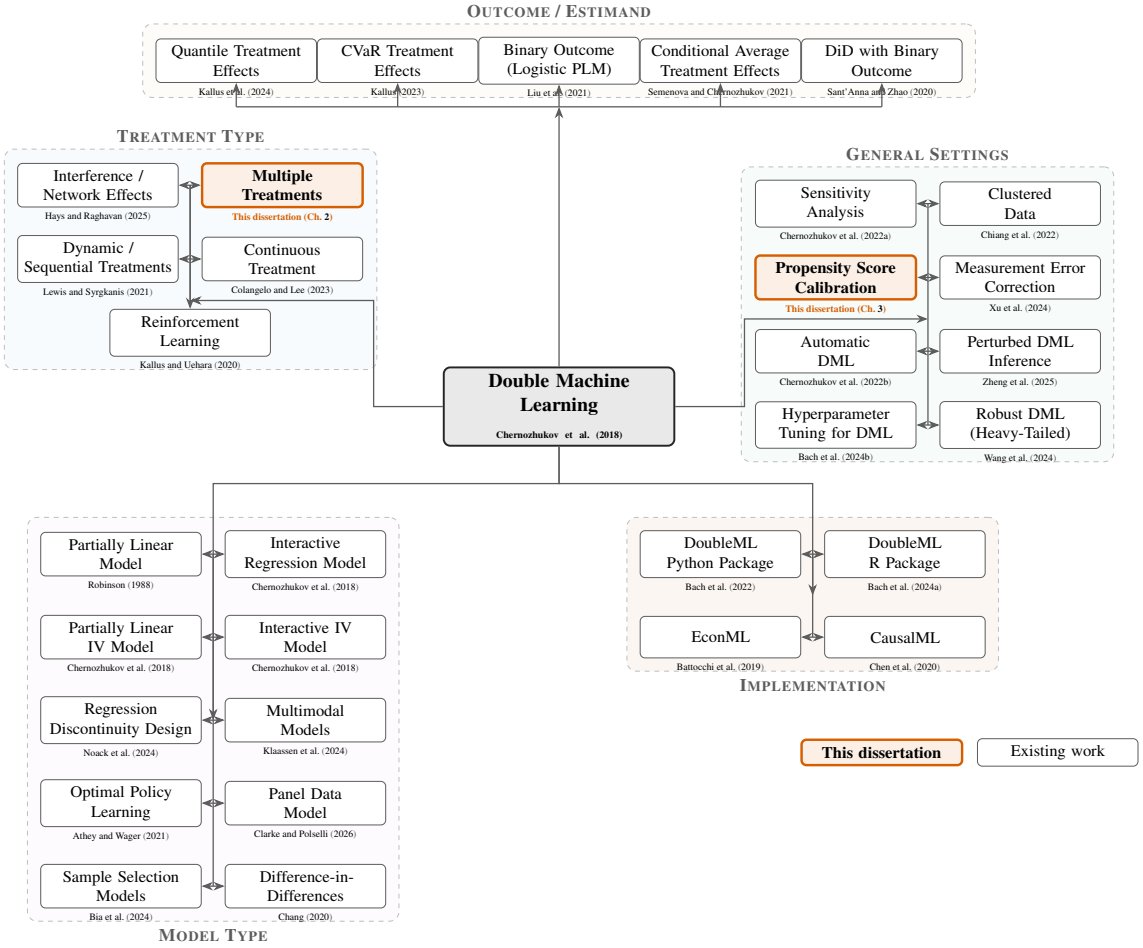

Figure 1.1: Extensions of the double machine learning framework (Chernozhukov et al. 2018). Nodes with red borders indicate contributions from this dissertation (Chapters 2 and 3). This overview is necessarily selective.

We formalize marginal attribution through a regularized tempered policy that restricts estimation to empirically supported regions, yielding a probabilistic generalization of the Shapley value. A Method of Simulated Scores estimator addresses computational intractability while preserving $\sqrt{n}$-consistency and asymptotic normality. The empirical application estimates the marginal effects of lifestyle factors on blood pressure, drawing on U.S. data from the National Health and Nutrition Examination Survey.

Chapter 3: Calibration Strategies for Robust Causal Estimation. Joint work with Sven Klaassen, Jannis Kueck, and Philipp Bach. This essay investigates how post-hoc calibration of propensity scores affects causal estimators within DML. We establish that isotonic calibration preserves the convergence and complexity conditions required by DML theory, and we systematically evaluate the interaction between calibration methods, sample-splitting schemes, and base learners across four data-generating processes. The key finding is that full-sample calibration using cross-fitted propensity scores provides the most stable improvements.

Chapter 4: Uncertainty Estimation in Insurance Claims Modeling. Joint work with Michael Merz. This essay develops conformal prediction methods for zero-inflated count data. The recently proposed dynamically grouped conformal prediction (DGCP) framework (Papaioannou et al. 2026) offers a principled approach to outcome-conditional calibration, but its application to the specific challenges of zero-inflated count data, where the zero-generating mechanism differs qualitatively from the positive-count process, requires methodological adaptation. We propose a two-stage framework combining Mondrian conformal prediction for the claim/no-claim decision with dynamically grouped calibration for positive counts. Simulations across four data-generating processes and an application to German motor insurance data demonstrate balanced outcome-conditional coverage where standard methods systematically fail. A Burt distance-based clustering diagnostic identifies portfolio segments with elevated prediction uncertainty.

Across these three essays, a common methodological stance emerges: the value of a statistical framework is measured not by its elegance in the standard case, but by its reliability in the cases that practitioners actually encounter. Each chapter takes an established, general-purpose tool and asks what breaks when it meets a specific empirical reality, and what minimal adaptation restores its usefulness.

Chapter 2

Inference on Multiple Treatment Effects with an Application to Health Economics

Joint work with Cathy Yi-Hsuan Chen (Adam Smith Business School, University of Glasgow, UK), Victor Chernozhukov (Economics Department, MIT, USA), and Martin Spindler (Institute for Statistics, Hamburg Business School, University of Hamburg, Germany)

Abstract. With the rise of digitization more complex data sets are available for empirical research. In causal inference the focus has been traditionally on estimation of the causal effect of a selected treatment variable or very few. But in many relevant applications, there are very many treatment variables and their interactions can be very complex. An example are the causes for high blood pressure. In this paper we develop methods for estimation and inference of the marginal treatment effect of variables in such complex settings with modern machine learning methods. Based on the concept of Shapley values, we provide debiased estimators for the marginal contribution and derive their theoretical properties which allows for valid inference on the marginal effects. Because of the complex nature of the underlying estimation problem, also a simulation based debiased estimator is propoposed and analyzed. In a simulation study the small sample properties for the estimators are compared. Despite the challenging setting, our estimator overall gives good results. In an application we estimate the marginal effects of different lifestyle risks for high blood pressure using data from The National Health and Nutrition Examination Survey (NHANES).

Keywords: multiple treatments, double/debiased machine learning, causal machine learning, methods of simulated scores, marginal treatment effects.

2.1 Introduction

2.1.1 Motivation

Causal inference in high-dimensional, complex data has been an active field of research in the recent years because of its importance to analyze real world data. In the standard setting the causal effect of one or few treatment variables is considered. But in some situations, very many treatment variables are given, e.g. the causes of high blood pressure

in medicine. These treatment variables can interact in very complicated ways. In this paper, we extend the standard setting to deal with very many treatment variables and to conduct valid inference on causal quantities of interest, in particular average marginal treatment effects. Utilizing the concept of Shapely values, we provide debiased estimates for the target parameters of interest. To handle the computational complexity, we generalize the framework of Double Machine Learning (DML) as proposed by Chernozhukov et al. (2018) to accommodate multiple treatments and simulated estimators. If we consider a nonlinear setting with very many binary treatment variables, estimating the marginal ATE is a non-trivial task. In many situations, the treatments exhibit interactions, and the combination of treatments determines the outcome in a complex and nonlinear manner. Furthermore, the presence of potentially high-dimensional confounding variables exacerbates the complexity of this estimation process. Many important examples have this structure:

Remark 1 (Examples of Combinatorial Causal Structures)**.**

- ***Retail Marketing:*** *How does a multi-item promotional bundle affect a store's total sales volume?*
- ***Medicine & Health Economics:*** *Patient mortality is often the result of several interacting comorbidities. Causal attribution is required to understand the weight of a single disease within a complex cluster of health conditions.*
- ***Genomics:*** *Phenotypic traits are determined by the non-linear interaction of many genes. Evaluating the marginal effect of a single gene requires accounting for the latent genomic architecture.*
- ***Team Sports:*** *How can the value added of a single player be determined when their performance is intrinsically linked to the composition of the rest of the team?*
- ***Modern Central Banking:*** *How do multifaceted monetary policy implementations, combined with forward guidance and central bank communication, influence inflation? Isolating the effect of one instrument requires a multi-treatment framework.*
- ***Dynamic Treatment Regimes:*** *Varying sequences of treatments over time can be viewed as multiple treatment subsets. Estimating the effect of an intervention at a specific time step requires accounting for the history of prior interactions.*

We show consistency and normality of our proposed estimators for the marginal treatment effects in complex settings, where machine learning methods are used for the estimation of the nuisance parameters building on the Double Machine Learning (DML) framework of Chernozhukov et al. (2018) and extend it.

The complexity of treatment interactions across various subgroups and their interactions in a higher-order feature space poses a significant challenge to come up with an isolated estimator. Estimating the importance of treatment j can be related to the game-theoretic attribution technique based on the Shapley value, which is used to interpret black-box machine learning models. Shapley values have been widely used as a feature attribution method in machine learning in numerous prior works. Evaluating treatment importance in "black-box" causal models can be effectively characterized as a "coalitional game" in a set of treatments under a wide variety of treatment subset choices.

One of the novelties we introduce is to conduct simulation-based inference based on the Neyman-orthogonal score to circumvent the hard-to-compute score function that involve multiple integrations. We propose to undertake the method of simulated scores ($\tilde{\theta}_{j,N}$), introduced by Hajivassiliou and McFadden (1998), for the non-trivial moment conditions and establish the asymptotic properties for the $\tilde{\theta}_{j,N}$ estimators for marginal ATEs.

Moreover, in a simulation study we assess the small sample performance of our proposed estimators and compare it to benchmark estimators. Despite the challenging, complex setting our estimator performs well giving a good performance. Finally, we apply our method to analyze the effect of different causes on high blood pressure. High blood pressure is caused by a combination of many factors and understanding these pattern is key. We use the The National Health and Nutrition Examination Survey (NHANES) provided by CDC's National Center for Health Statistics (NCHS) for our empirical application.

2.1.2 Related Work

Estimating causal effects from observational data is a central challenge in empirical research. While the binary treatment setting is well-established, many applications involve comparing multiple treatments. As a result, various extensions of the propensity score have been introduced to support inference under multi-level and more complex treatment regimes.Joffe and Rosenbaum (1999) showed that a scalar balancing score is sufficient for ordinal dose groups under specific assumptions, while Imbens (2000) formalized the generalized propensity score (GPS) for unstructured categorical treatments. He demonstrated that conditioning on the treatment-specific conditional probability removes confounding bias. Imai and van Dyk (2004) further generalized this to the "propensity function" for arbitrary regimes, including continuous and multivariate treatments.

Building on these foundations, Feng et al. (2012) detailed practical GPS implementations using regression adjustment and inverse probability of treatment weighting (IPTW). In the regression adjustment approach, the outcome is modeled as a function of the GPS for each treatment group separately; in the weighting approach, observations are weighted by the inverse of their GPS to create a pseudo-population where confounders are balanced. To handle high-dimensional covariates, GPS estimation often employs multinomial logistic regression with regularization (Tibshirani 1996) or flexible machine learning techniques like Generalized Boosted Models (GBM), as detailed by McCaffrey et al. (2013), which better capture non-linear relationships. To address instability caused by extreme weights in IPTW, Li and Li (2019) proposed "balancing weights" and generalized overlap weights (GOW), targeting subpopulations with substantial probability of receiving any treatment.

Other approaches focus on outcome modeling or latent structures. Hu and Gu (2021) compared methods for multiple treatments with rare outcomes, finding that Bayesian Additive Regression Trees (BART) and Regression Adjustment on Multivariate Spline of GPS (RAMS) outperformed traditional weighting. Alternatively, Wang and Blei (2019) introduced the "deconfounder" algorithm, which fits a probabilistic factor model to infer latent variables substituting for unobserved multi-cause confounders. This approach has generated substantial debate regarding its theoretical foundations (Grimmer et al. 2023), particularly concerning the requirement for strong infinite confounding and its implications for overlap violations. Leveraging known causal graphs, Qian et al. (2021) proposed the Single-Cause Perturbation (SCP) method to decompose multi-cause problems into

sequential single-cause estimations. Through a data augmentation technique, where non-descendant causes are treated as additional covariates and descendant causes as additional outcomes, standard single-cause estimators can be applied sequentially. This approach avoids restrictive parametric assumptions on the outcome model but requires specifying a causal structure among the treatments.

Recent advances in deep learning include representation learning for counterfactual inference (Shalit et al. 2017) and meta-learning frameworks (Zhou et al. 2023). Zhou et al. (2023) addressed the challenge of data imbalance across treatment groups. Their method treats each treatment group as a separate learning task, training a meta-learner on treatment groups with sufficient data (source domains) and adapting it to groups with limited data (target domains) using gradient-based optimization. The framework incorporates a discrepancy loss based on Maximum Mean Discrepancy (MMD) to align latent representations across treatment groups, improving generalization in few-shot scenarios. This approach is particularly promising for observational studies where inherent selection biases lead to highly imbalanced treatment groups. Complex interaction effects have also been explored: Parbhoo et al. (2021) modeled cross-treatment interactions using neural networks, while in recommender systems, Schnabel et al. (2016) and Wang et al. (2019) applied propensity scoring and doubly robust estimators to treatment bundles.

Most recently, Xiang et al. (2025) proposed a Double Machine Learning framework specifically designed to estimate the effects of multiple treatments and their interactions. Their work represents a significant step in extending DML beyond binary or continuous single treatments, focusing on the identification of complex interaction terms in high-dimensional settings. For continuous treatments, Zhao et al. (2025) extended the DML framework to handle multi-dimensional interventions, incorporating monotonicity constraints for domain-specific applications.

Our work provides a general framework for attributing effects across arbitrary treatment combinations without relying on restrictive functional form assumptions or known causal treatment structures. Methodologically, we offer two key innovations. First, we formalize causal attribution through treatment subset averaging, a causal analogue of Shapley-value aggregation (Shapley 1953). Rather than imposing combinatorial weights based on random permutations, our estimator weights by the observational distribution $P(S|X)$, yielding marginal treatment effects that reflect real-world patterns. Second, to address the computational intractability of enumeration-based approaches in high-dimensional treatment spaces, where the power set 2^d explodes, we integrate simulation-based inference within the DML framework (Chernozhukov et al. 2018).

The paper is structured as follows: After the Introduction (Section 1) the model setting and some fundamental assumptions are presented in Section 2. In Section 3 the estimators, algorithms and their theoretical properties are given. Section 4 contains the results from a simulation study to evaluate the finite sample performance of our proposed estimator and compare it to selected benchmarks. An empirical application is presented in Section 5. Finally, we conclude in Section 6.

2.2 Regularized Policies for Combinatorial Treatments

We address the fundamental challenge of causal inference in combinatorial treatment regimes , where the outcome is driven by the non-linear interaction of d binary causes. In such high-dimensional settings, the cardinality of the treatment configuration space (2^d) grows exponentially, typically exceeding the available sample size. This "curse of dimensionality" results in structural violations of the positivity assumption (overlap), rendering traditional “atomic” (deterministic) interventions statistically unidentified.

While recent advances have addressed aspects of high-dimensional inference, a critical gap remains in the **attribution problem**. Chernozhukov et al. (2018) established the Double Machine Learning (DML) framework for controlling high-dimensional confounders, and Wang and Blei (2019) analyzed the aggregated impact of multiple causes. However, neither approach resolves the instability of isolating the marginal contribution of a specific treatment j when the background configuration a_{-j} has near-zero probability in the observational data. In these regimes, standard estimators effectively rely on dangerous extrapolation, leading to inflated variance and unreliable inference.

To resolve this, we propose a shift from atomic estimands to regularized tempered policies. We define the marginal effect with respect to a reference policy $\pi_{T,\delta}$ that is explicitly constructed to respect the support of the data. This regularization consists of two mechanisms:

- **Tempering (T):** A smoothing operation that relaxes the target distribution, allowing the estimator to borrow information from neighboring configurations for identification.
- **Truncation (δ):** A strict constraint that zeros out regions of the treatment space where empirical overlap is insufficient.

By integrating out the combinatorial complexity over this ”regularized support” leveraging by these two parameters, we replace a computationally intractable and statistically unstable summation with a robust, efficient estimation problem. This formulation provides a **probabilistic generalization of the Shapley value**, optimizing the trade-off between the bias of the identifiable target and the variance of the estimator.

2.2.1 Model Setup and Preliminaries

Let $Y \in \mathcal{Y} \subseteq \mathbb{R}$ denote the outcome variable, and $X \in \mathcal{X} \subseteq \mathbb{R}^p$ denote a high-dimensional vector of confounders. Let $\mathcal{D} = \{1, \dots, d\}$ be the index set of treatment variables. We define the treatment assignment as a random vector $A = (A_1, \dots, A_d)^\top \in \{0,1\}^d$. The space of all possible treatment vectors is denoted by $\mathcal{A} = \{0,1\}^d$, which has cardinality $|\mathcal{A}| = 2^d$. For example, if $d = 3$, the space of possible treatment vectors is:

$$\mathcal{A} = \big\{(0,0,0)^t, (1,0,0)^t, \dots, (1,1,1)^t\big\}.$$

While each element $a \in \mathcal{A}$, as a d-dimensional binary vector,corresponds uniquely to a subset of active treatments, we rely on vector notation for algebraic clarity in the regression framework.

We adopt the potential outcomes framework. For any treatment vector $a \in \mathcal{A}$, let $Y(a)$ denote the potential outcome under treatment vector a. The observed outcome is $Y = \sum_{a \in \mathcal{A}} \mathbb{1}(A = a) Y(a)$. We assume the data generating process follows a generalized interactive model:

$$Y = \mu(A, X) + U, \quad E[U \mid X, A] = 0 \tag{2.1}$$
$$A \sim P(A \mid X) \tag{2.2}$$

where $\mu(a, x) = E[Y(a) \mid X = x]$ is the conditional outcome expectation and and $P(a \mid x) = \mathbb{P}(A = a \mid X = x)$ denotes the **joint propensity score**, representing the probability of observing the specific treatment configuration a given covariates.

We are interested in the marginal effect of a specific treatment variable $j \in \mathcal{D}$. Let e_j be the unit vector of dimension d with the j-th entry equal to 1. We define the set of "background" treatment vectors where treatment j is inactive (zero) as:

$$\mathcal{A}_{-j} = \{a \in \mathcal{A} : a_j = 0\}.$$

To prevent ambiguity, we distinguish the usage of treatment vector symbols. We use a as an iterator or dummy variable when defining sets (e.g., $a \in \mathcal{A}_{-j}$) or computing aggregations (e.g., sums $\sum_a$). We use s to denote a specific fixed point or background configuration. Consequently, an expression like $E[Y|A = s]$ refers to the expected outcome conditional on the observed treatment vector being fixed to the specific value s, whereas the summation $\sum_a \pi(a|X)$ implies iterating over the entire support of the distribution.

Consider a specific background configuration $s \in \mathcal{A}_{-j}$. We define the *conditional marginal effect* $\theta_j(s)$ as the incremental benefit of adding treatment j, conditional on the sub-population that naturally selected the background profile s:

$$\theta_j(s) = \mathsf{E}_{X|A=s}[\mu(s + e_j, X) - \mu(s, X) \mid A = s]. \tag{2.3}$$

The expectation $\mathsf{E}_{X|A=s}$ is taken with respect to the covariate distribution of the treated group ($X \sim P(X \mid A = s)$). This quantity measures the effect specifically for the patients who experienced history s, analogous to the Average Treatment Effect on the Treated (ATT).

Let denote the observed data as $W = (Y, A, X)$. Throughout our theoretical exposition, we denote $\mathbb{E}_P[\cdot]$ as the expectation with respect to the true joint distribution of the observed data $W = (Y, A, X)$. This unifies our notation for both the moment conditions and the target parameter.

Assumption 1 (Causal Identification)**.** *For all treatment vectors $a \in \mathcal{A}$ and all covariates $x \in \mathcal{X}$, the following conditions hold: (i)* ***unconfoundedness:*** *the potential outcomes are independent of the observed treatment assignment conditional on covariates $Y(a) \perp A \mid X = x$. (ii)* ***strict overlap:*** *the joint probability of observing any treatment configuration is strictly bounded away from zero: $P(A = a \mid X = x) \geq \epsilon > 0$, for some constant ϵ. (iii)* ***identification:*** *under the conditions above, the moment conditions for the subset-specific scores are satisfied at the true parameters:*

$$\mathsf{E}_P[\psi_j(W; \theta_j(\pi_{T,\delta})(s), \eta_0)] = 0.$$

Proposition 1 (Atomic Identification)**.** *Under Assumptions 1, the conditional marginal effect $\theta_j(s)$ is identified by the efficient score:*

$$\psi_{j,s}(W) = \psi_{s+e_j}(W) - \psi_s(W), \tag{2.4}$$

where $\psi_s(W) = \frac{\mathbb{I}(A=s)}{P(s|X)}(Y - \mu(s,X)) + \mu(s,X)$ is the efficient influence function for the mean potential outcome of subset s.

While Proposition 1 establishes theoretical identification, the estimator for $\theta_j(s)$ relies on the inverse probability weight $1/P(s|X)$. In high-dimensional settings, the probability of observing any specific configuration s may be negligible ($P(s|X) \approx 0$), leading to violations of the overlap assumption and unbounded variance.

The Aggregation Problem: Shapley Values While $\theta_j(s)$ provides granular insight, policy analysis requires a global measure of variable importance. This necessitates aggregating these local effects over all possible background configurations $\mathcal{A}_{-j}$. A standard approach in interpretable machine learning is the Shapley value, which aggregates these marginal effects using a uniform (or combinatorial) probability measure. To obtain a population-level metric, the Shapley value averages the marginal effects over the global covariate distribution $P(X)$:

$$\theta_j^{\text{Shapley}} = \frac{1}{|\mathcal{A}_{-j}|} \sum_{s \in \mathcal{A}_{-j}} E_{P(X)}[\mu(s + e_j, X) - \mu(s, X)]. \tag{2.5}$$

In (2.5), every possible treatment combination s contributes equally to the final estimate (weight $= 1/|\mathcal{A}_{-j}|$). While this satisfies game-theoretic axioms for attributing model variance, it is often misleading for causal policy analysis. In health economics, treatment combinations are rarely uniformly distributed; many subsets s (e.g., contraindicated drug pairs) may correspond to null populations where the treatment is never administered. Assigning equal importance to these irrelevant regions distorts the estimated effect.

2.2.2 Tempered Policy and Overlap Regularization

Regularisation for limited overlap

While a traditional causal estimand typically targets a fixed, deterministic configuration s, such an approach is highly susceptible to violations of the *strict overlap assumption* (positivity) when the number of potential treatment combinations is large. Kennedy (2019) identifies the failure of the positivity assumption (overlap) as the primary barrier to causal inference in high-dimensional or continuous settings. When the observational density approaches zero at the target value of interest, the standard atomic causal effect is statistically unidentified, leading to infinite bounds on the estimator variance. We redefine our target as a weighted average of potential outcomes over a tempered distribution $\pi(s|X)$ rather than a single mass point where we assign or smooth non-zero density to a range of configuration states.

By smoothing the target distribution via temperature scaling, we prevent the Riesz representer, e.g. $\frac{\mathbb{I}(A=s)}{P(s|X)}$ in (2.3), from diverging in regions of sparse data. In a high-dimensional treatment regime, the resulting RR in the framework of Rosenbaum and Rubin

(1983) is highly unstable. Thus, the tempered policy acts as a regularization device for the supports of treatment combinations, allowing us to compute a policy-relevant marginal effect that remains robust to the curse of dimensionality and the empirical sparsity of complex treatment regimes.

We propose a tempered policy based on energy-based models (LeCun et al. 2006):

$$\pi_T(a \mid X) \propto P(A = a \mid X)^{1/T}. \tag{2.6}$$

where $T \geq 1$ is a temperature scalar. This formulation governs a bias-variance trade-off: higher T improves overlap (reducing estimator variance) by redistributing mass from high-probability to low-probability configurations, at the cost of shifting the estimand away from the natural observational distribution (Levine et al. 2020) which is bias. To see it clearly, we can frame **tempering as entropy regularization**. The proposed tempered policy can be theoretically justified as the solution to a constrained optimization problem. Seeking a reference distribution π that balances fidelity to the observed data structure with maximum overlap (uniformity) corresponds to maximizing the entropy-regularized likelihood:

$$\pi^* = \arg\max_{\pi} \left(\mathbb{E}_\pi[\log P(A|X)] + T \cdot H(\pi) \right). \tag{2.7}$$

where $H(\cdot)$ is an entropy operator. The solution to this objective is exactly $\pi^*(a) \propto P(a|X)^{1/T}$ that preserves strictly positivity and rank monotonicity. Thus, the temperature T serves as the regularization coefficient, analogous to λ in Ridge regression, preventing the distribution from collapsing onto sparse atomic configurations.

We discuss temperature regimes and the regularization path that leads to different tempered policy π_T: (i) $T \to \infty$: maximal exploration where π_∞ converges to the discrete uniform distribution $1/|\mathcal{A}|$. This aligns with the Shapley value framework, treating all counterfactual configurations as equally relevant. (ii) $T \downarrow 0$: π_0 converges to a Dirac delta δ_{a^*} centered at the mode $a^* = \arg\max P(a|X)$. This recovers the deterministic "atomic" treatment effect. (iii) $T = 1$, π_1 coincides with the observational propensity $P(A|X)$. Our proposed estimator utilizes $T > 1$ as a structural regularizer, softening the sharp peaks of the observational distribution to ensure the Riesz weights remain well-behaved while maintaining the characteristic geometric structure of the treatment space. In later section, we discuss rate of convergence of the estimand that requires the regularity conditions on T.

Comparison with exponential tilting and shifting

Consider any two treatment vectors $a, a' \in \mathcal{A}$ with $P(a|X) > P(a'|X)$. The relative weight assigned to the secondary configuration a' versus the primary configuration a increases with T:

$$\frac{\pi_T(a'|X)}{\pi_T(a|X)} = \left(\frac{P(a'|X)}{P(a|X)} \right)^{1/T}. \tag{2.8}$$

For $T > 1$, this ratio is strictly larger than the observational ratio. This implies a systematic "lifting" of the probability mass from the tails (low propensity) towards the peaks (high propensity), effectively flattening the distribution. It is instructive to contrast our tempering approach with the exponential tilting framework common in the stochastic intervention literature (e.g., Kennedy (2019), Muñoz and Van Der Laan (2012)). The stabilizing mechanism of the tempered policy can be understood through the lens of *logit shrinkage*.

Let $\log(P(a'|X)/P(a|X))$ be the observational log-odds ratio between two configurations. For an exponential tilting ($\pi_T \propto e^{Ta}P$), it imposes an *additive* shift on the log-odds:

$$\log \frac{\pi_T(a')}{\pi_T(a)} = \log(P(a'|X)/P(a|X)) + T(a' - a).$$

Clearly, it shifts the mean of the distribution. While useful for estimating "incremental effects" (e.g., what if everyone smoked slightly less?), it implies a directional preference and can inadvertently shift probability mass into regions of poor overlap if T is misspecified.

For tempering ($\pi_T \propto P^{1/T}$), it imposes a *multiplicative* scaling on the log-odds:

$$\log \frac{\pi_T(a')}{\pi_T(a)} = \frac{1}{T} \cdot \log(P(a'|X)/P(a|X)).$$

, increasing the variance (shrinking the logits) without altering the mode or the rank-ordering of the configurations. Unlike shifting, tempering is structurally conservative: it forces the intervention distribution to relax towards uniformity within the existing data support, strictly improving overlap rather than risking its violation. We further can link it with Ridge regression. Tempering is mathematically equivalent to *coefficient shrinkage*. This mirrors the mechanism of L_2-regularized logistic regression, where the estimated coefficients $\hat{\beta}_{Ridge}$ are shrunk relative to the maximum likelihood estimates by a factor of roughly $(1 + \lambda)^{-1}$. Identifying $T \approx 1 + \lambda$, we observe that our method applies a "post-estimation Ridge penalty." Whereas standard Ridge regression (in DML procedure) applies shrinkage during the *learning* phase to minimize prediction variance, tempering applies shrinkage during the *inference* phase to minimize the variance of the inverse probability weights.

2.2.3 Estimand under Tempered Policy and its Riesz Representers

How the tempered policy stabilise the DML framework when the score is constructed doubly robust ? To answer it, we may consider the Riesz representers in the moment condition.

By the Riesz representation theorem, there exists a unique random variable $\alpha_0(A, X)$, known as the Riesz representer (RR), such that for any square-integrable function g,

$$\mathsf{E}_P[\pi(A \mid X)g(A, X)] = \mathsf{E}_P[\alpha_0(A, X)g(A, X)]. \tag{2.9}$$

By the Radon-Nikodym theorem the Riesz representer is explicitly the density ratio between the target policy $\pi(A \mid X)$ and the observational propensity $P(A \mid X)$:

$$\alpha_0(A, X) = \frac{\pi(A \mid X)}{P(A \mid X)}. \tag{2.10}$$

The Riesz Representation Theorem states that for any continuous linear functional $L : \mathcal{H} \to \mathbb{R}$ on a Hilbert space $\mathcal{H}$, there exists a unique element $\alpha_0 \in \mathcal{H}$ such that $L(g) = \langle \alpha_0, g \rangle_{\mathcal{H}}$ for all $g \in \mathcal{H}$. In the context of causal inference, we define the inner

product using the observational measure P:

$$\langle \alpha_0, g \rangle_P = \int \alpha_0(a,x) g(a,x) dP(a,x) = \mathsf{E}_P[\alpha_0(A,X) g(A,X)]. \tag{2.11}$$

When the target functional is π, we have $L(g) = \mathsf{E}_\pi[g(A,X)]$. By the Radon-Nikodym theorem, if π is absolutely continuous with respect to P, the representer is simply the density ratio $\alpha_0 = d\pi/dP$. This equality allows us to evaluate counterfactual policies using the observational distribution by re-weighting the observed realizations $g(A,X)$ by the factor $\alpha_0(A,X)$.

The semi-parametric efficiency bound (and thus the asymptotic variance of our DML estimator) scales with the second moment of this representer, $E[\alpha_0^2(A,X)]$ (We will discuss it in the later section). When the overlap assumption is violated—i.e., when $P(A = a \mid X)$ approaches zero for treatment vectors where $\pi(a \mid X) > 0$—the RR diverges ($\alpha_0 \to \infty$), causing the variance to explode. This leads to the "relevance (bias)-stability (variance)" trade-off, and can be understood in the following scaling strategy.

While tempering smooths the distribution, it does not strictly bound the variance of RR as the overlap vanishes ($P(a|X) \to 0$) in (2.10). Thus, tempering alone changes the *rate* of explosion but does not prevent it. To ensure finite variance of the estimator, strictly bounded Riesz weights are required. The threshold $\pi(a \mid X) > \delta$ where $\delta > 0$ provides this guarantee by imposing a hard constraint on the support. The resulting weights are strictly bounded to sustain the asymptotic theorem. For this reason, we involve dual regularisation parameters in the policy denoted by by (T, δ) serves as a structural regularization device to address the identification-variance trade-off in high-dimensional treatment interaction models. T governs the *shape* of the integration (exploration vs. specificity), while δ governs the *safety* of the integration (excluding non-identified regions). This formulation allows us to target the most informative causal parameter that is statistically identified by the data.

Let $T \geq 1$ be the temperature parameter and $\delta > 0$ be the overlap threshold. We define the **Truncated Tempered Policy** $\pi_{T,\delta}(a \mid X)$ restricted to the feasible support:

$$\pi_{T,\delta}(a \mid X) = \frac{1}{\mathcal{Z}_{T,\delta}(X)} \left(P(a \mid X)^{1/T} \cdot \mathbb{I}\left(P(a \mid X) \geq \delta\right) \right), \tag{2.12}$$

where $\mathcal{Z}_{T,\delta}(X) = \sum_{a' \in \mathcal{A}} P(a' \mid X)^{1/T} \cdot \mathbb{I}(P(a' \mid X) \geq \delta)$ is the normalizing constant. The corresponding **Regularized Riesz Representer** $\alpha_{T,\delta}(a,X)$ is given by the density ratio:

$$\alpha_{T,\delta}(a,X) \equiv \frac{\pi_{T,\delta}(a \mid X)}{P(a \mid X)} = \frac{1}{\mathcal{Z}_{T,\delta}(X)} \left(P(a \mid X)^{\frac{1}{T}-1} \cdot \mathbb{I}\left(P(a \mid X) \geq \delta\right) \right). \tag{2.13}$$

Proposition 2 (Boundedness of the Regularized Riesz Representer)**.** *Let $T \geq 1$ and $\delta > 0$. The Regularized Riesz Representer $\alpha_{T,\delta}(a,X)$ defined in (2.13) satisfies the uniform bound:*

$$\sup_{a \in \mathcal{A}, x \in \mathcal{X}} |\alpha_{T,\delta}(a,x)| \leq \frac{1}{\delta} < \infty. \tag{2.14}$$

Furthermore, considering the specific structure of the tempered policy, the bound is tighter

relative to the partition function $\mathcal{Z}_{T,\delta}(X)$:

$$\alpha_{T,\delta}(a,x) \leq \frac{\delta^{\frac{1}{T}-1}}{\mathcal{Z}_{T,\delta}(x)}. \tag{2.15}$$

We define the target parameter under truncated tempered policy $\pi_{T,\delta}(a|X)$. Unlike the atomic effects $\theta_j(s)$ which are conditional on the specific sub-population $A = s$, the target parameter seeks a population-level average. We therefore aggregate the effects over the marginal distribution of covariates $P(X)$. Note that while $\pi_{T,\delta}$ is defined over all $\mathcal{A}$, the marginal effect for treatment j sums over the background set $\mathcal{A}_{-j}$:

$$\theta_j(\pi_{T,\delta}) = \mathsf{E}_P \left[\sum_{s \in \mathcal{A}_{-j}} \pi_{T,\delta}(s \mid X) \left(\mu(s + e_j, X) - \mu(s, X) \right) \right], \tag{2.16}$$

where the expectation is taken with respect to the marginal $P(X)$ implied by the joint distribution $P(W)$. This estimand exploiting tempered policy $\pi_{T,\delta}$, which is fully contained within the support of the observational data, resolves the identification issue if positivity is potentially violated in (2.3). From a regulatory perspective, the tempered estimand provides a *conservative lower bound* on the potential benefit. Because we do not extrapolate to regions of zero exposure (which are truncated by δ), we avoid the risk of over-promising benefits based on unstable statistical models. We provide a rigorous estimate of the gains achievable *within the reliable support* of environmental conditions.

2.3 Main Theoretical Results

Table 2.1: Summary of notations for causal estimands and estimators

Symbol	Definition
$\theta_j(a^*)$	**Ideal Atomic Target**. The causal effect of the deterministic intervention at the local mode a^*. This is the scientific quantity of interest (often unidentified without smoothing).
$\theta_j(\pi_{T,\delta})$	**Tempered Target**. The causal effect under the tempered policy $\pi_{T,\delta}$. This is the feasible population parameter we actually target for fixed T, δ.
$\hat{\theta}_{j,N}$	**DML Estimator**. The numerical estimate derived from sample size N using the tempered policy.
$\alpha_{T,\delta}(a,x)$	**Regularized Riesz Representer**. The density ratio (importance weight) for the tempered policy, bounded by $1/\delta$.

We introduce a tempered policy $\theta(\pi_{T,\delta})$ which satisfies identifiability but is biased. Our main results characterize the trade-off between these quantities in Table 2.1 as follows:

1. **Inference on the constraint support (Theorem 1).** We address the practical question of inference for a fixed regularization parameters (T, δ), we derive the

asymptotic properties of the estimator relative to the *identifiable target*. We establish that the DML estimator $\hat{\theta}_{j,N}$ is $\sqrt{N}$-consistent and asymptotically normal around the tempered estimand $\theta_j(\pi_{T,\delta})$.

2. **Approximation risk (Theorem 2).** We formalise the deterministic bias incurred by replacing the unidentified atomic target $\theta_j(a^*)$ with the regularized proxy $\theta_j(\pi_{T,\delta})$. We show that this bias is controlled by the outcome smoothness and the propensity tail.

3. **Rate of convergence (Theorem 3).** We consider regularization parameters (T_N, δ_N) to demonstrate structural consistency. We ask if we can recover the true atomic effect as the sample size grows. By allowing the regularization parameters (T_N, δ_N) to decay towards the atomic limit at optimal rates, we show that the total error (Bias + Variance) vanishes, achieving the **minimax** optimal convergence rate.

4. **Asymptotic normality(Theorem 4).** Finally, we characterize the limiting distribution of the estimator relative to the *ideal atomic target*. We establish the scaling factor $\sqrt{N\delta_N}$ required to stabilize the variance as we approach the identification boundary.

Together, these results provide a complete picture: Theorem 2 justifies the *design*, Theorem 1 validates the *practice*, and Theorems 3–4 prove the *fundamental validity* of the method.

2.3.1 Fixed T, δ Case

To establish valid inference for the marginal treatment effect $\theta_j(\pi_{T,\delta})$ or shorthand $(\theta_j(\pi))$ using Double Machine Learning, we must verify a set of regularity conditions adapted to our multiple treatment setting. Let denote the observed data as $W = (Y, A, X)$, and define the nuisance parameters as $\eta = (\mu, P)$, where $\mu(a, x) = \mathsf{E}[Y \mid A = a, X = x]$ is the outcome regression and $\alpha(a, x)$ is the Riesz representer associated with the target functional $\theta_j(\pi)$. Denote the expectation and variance operators with respect to the observational distribution P as $\mathsf{E}_P[\cdot]$ and $\mathrm{Var}_P[\cdot]$, respectively. The score function for treatment j is given by: Let $\mathcal{T}$ be the function space containing the Riesz representer α_0. We assume $\mathcal{T} \subset L_2(P)$ is a convex set of square-integrable functions. We adapt the assumptions from Chernozhukov et al. (2025) to our multiple treatment setting.

Assumption 2 (Mean Square Continuity of the Functional)**.** *The target functional $\theta(\mu)$ is mean square continuous with respect to the L_2 norm and the inner product by $\|\mu\|_2 := \sqrt{\mathsf{E}[\mu(X)^2]}$ and $\langle \mu, \cdot \mu \rangle$, respectively. That is, there exists a constant $M < \infty$ such that for any μ in the parameter space:*

$$\mathsf{E}[m(W; \mu)^2] \leq M \|\mu\|_2^2. \tag{2.17}$$

As established in Chernozhukov et al. (2025) (Section 2), this condition is sufficient for the existence of the Riesz representer $\alpha_0 \in L_2(P)$. In our context of stochastic policies, this condition requires that the importance weights are square-integrable, i.e., $\mathsf{E}[(\pi(A|X)/P(A|X))^2] < \infty$. The use of temperature scaling ($T > 1$) ensures this condition holds even when strict overlap is violated for the natural distribution.

Assumption 3 (Mixed Bias Condition)**.** *Let $\hat{\mu}$ and $\hat{\alpha}$ denote the estimators for the outcome regression and Riesz representer, respectively. We assume the product of their estimation errors vanishes at a rate faster than $N^{-1/2}$:*

$$\sqrt{N}\mathsf{E}_P\big[\big(\hat{\alpha}(A,X)-\alpha_0(A,X)\big)\big(\hat{\mu}(A,X)-\mu_0(A,X)\big)\big] \xrightarrow{p} 0. \tag{2.18}$$

This is satisfied if $\|\hat{\alpha}-\alpha_0\|_2 \cdot \|\hat{\mu}-\mu_0\|_2 = o_P(N^{-1/2})$. This condition allows for a trade-off in complexity: if $\hat{\mu}$ is estimated at a slower non-parametric rate, $\hat{\alpha}$ must be estimated with sufficient precision to compensate (Chernozhukov et al. 2025, Assumption 3).

Assumption 4 (Critical Radius and Complexity)**.** *Let m_N be the critical radius of the function space $\mathcal{T}$ used to estimate α. We assume the complexity of the space, captured by Rademacher complexity and denoted by m_N, and the estimation error r_N satisfy the condition $\sqrt{N}(m_N r_N + m_N^2) \to 0$. This "complexity-rate robustness" ensures that the empirical process term in the score decomposition is asymptotically negligible without requiring strict Donsker conditions (Chernozhukov et al. 2025, Theorem 3).*

Assumption 5 (Square-Integrability of the Riesz Representer)**.** *The Riesz representer $\alpha_0(A,X)$ satisfies the square-integrability condition:*

$$\|\alpha_0\|_2^2 = \mathsf{E}_P[\alpha_0(A,X)^2] \le C < \infty, \tag{2.19}$$

for some finite constant C. This condition ensures that the target functional $\theta(\mu)$ is mean square continuous (Chernozhukov et al. 2025, Assumption 1). In practice, choosing a tempered reference policy π_T with $T > 1$ serves as a regularization to satisfy this condition even when strict overlap $P(A|X) > \epsilon$ is empirically fragile.

Assumption 6 (Linearity and Moment Conditions)**.** *The score function $\psi_j(W;\theta,\eta)$ satisfies the following properties: (i)**linearity:** the score is affine in the parameter of interest θ:*

$$\psi_j(W;\theta,\eta) = \psi_j^a(W;\eta)\cdot\theta + \psi_j^b(W;\eta),$$

*where in our formulation, $\psi_j^a = -1$ and $\psi_j^b = \Gamma(X;\mu) + \alpha(A,X)(Y-\mu(A,X))$. (ii)**moment condition:** At the true parameter θ_0 and true nuisance η_0, the score is centered:*

$$\mathsf{E}_P[\psi_j(W;\theta_0,\eta_0)] = 0.$$

It holds over the joint law of W. This implies that the score has mean zero when averaged over the population of units, treatments, and outcomes. This linearity ensures that the DML estimator $\hat{\theta}_j = -\mathsf{E}_N[\psi_j^b]/\mathsf{E}_N[\psi_j^a]$ is well-defined and has a unique solution (Chernozhukov et al. 2025, Section 4).

We now formally define the score function for the tempered marginal treatment effect $\theta_j(\pi)$. This score generalizes the subset-specific scores derived in Section 2.1 by aggregating them according to the stochastic reference policy $\pi(a|X)$.

Let the nuisance parameters be $\eta = (\mu, P)$, where $\mu(A,X) = E[Y|A,X]$ is the outcome regression and $P(A|X)$ is the propensity score. The *Tempered Neyman Orthogonal Score* $\psi_j(W;\theta,\eta)$ for the marginal effect of treatment j is defined as:

$$\psi_j(W;\theta,\eta) = \Gamma(X;\mu) + \alpha(A,X)(Y-\mu(A,X)) - \theta, \tag{2.20}$$

where $\Gamma(X;\mu)$ is the plug-in estimator for the conditional marginal effect, averaged over the reference policy:

$$\Gamma(X;\mu) = \sum_{a \in \mathcal{A}_{-j}} \pi(a \mid X) \left(\mu(a + e_j, X) - \mu(a, X)\right). \tag{2.21}$$

Without ambiguity, we use $\pi := \pi_{T,\delta}$ for notational simplicity. Indeed, the tempered Neyman orthogonal score in (2.20) corresponds to an efficient influence curve under $P(W)$ that can be seen in result 1 of Muñoz and Van Der Laan (2012). Our theoretical framework builds on the stochastic intervention calculus of Muñoz and Van Der Laan (2012). However, while they consider general stochastic policies, we introduce a specific family of tempered policies to ensure the Riesz representer remains bounded in high-dimensional combinatorial spaces.

Since our functional $\theta_j(\pi)$ is a contrast (difference) between the treated state ($A_j = 1$) and untreated state ($A_j = 0$), the representer is the *Signed Density Ratio*:

$$\alpha(A, X) = \underbrace{\frac{\pi(A_{-j} \mid X)}{P(A \mid X)}}_{\text{Magnitude}} \cdot \underbrace{(2A_j - 1)}_{\text{Sign Flip}}. \tag{2.22}$$

where π is defined in (2.12). $\alpha(A, X)$ automatically handles the aggregation in (2.20). The term $(2A_j - 1)$ flips the sign of the residual based on whether the observed unit is in the treated ($A_j = 1$) or control ($A_j = 0$) group relative to the contrast j. Observed treated units ($A_j = 1$) contribute positively to the estimator, while observed control units ($A_j = 0$) contribute negatively.

A crucial property of this score is its insensitivity to local perturbations in the nuisance parameters. We formalize this property in the following proposition.

Proposition 3 (Neyman Orthogonality of the Tempered Score)**.** *The score $\psi_j(W;\theta,\eta)$ in (2.20) satisfies the Neyman orthogonality condition with respect to the nuisance parameter $\eta = (\mu, P)$ at the true value η_0. Specifically, the Gateaux derivative of the expected score with respect to η vanishes:*

$$\partial_r E_P[\psi_j(W;\theta_0, \eta_0 + r(\eta - \eta_0))]\Big|_{r=0} = 0. \tag{2.23}$$

While our framework allows for arbitrary correlation structures among treatments, it is instructive to consider the case of treatment independence. Under the assumption that the components of A are conditionally independent given X, the joint propensity score $P(a|X)$ factorizes into the product of marginal propensity scores, $\ell_j(X) = P(A_j = 1|X)$, which is widely adopted by the existing research (Wang and Blei 2019):

$$P(a \mid X) = \prod_{k=1}^{d} \ell_k(X)^{a_k} (1 - \ell_k(X))^{1-a_k}. \tag{2.24}$$

In this specific regime, the magnitude of the Riesz representer $\alpha(A, X)$ simplifies to:

$$|\alpha(A, X)| = \frac{\pi(A_{-j} \mid X)}{\prod_{k \neq j} \ell_k(X)^{A_k}(1 - \ell_k(X))^{1-A_k} \cdot P(A_j \mid X)}. \tag{2.25}$$

If the reference policy π is similarly chosen to be independent, the terms corresponding to $k \neq j$ in the numerator and denominator cancel out, reducing the Riesz representer to the standard inverse probability weight for the j-th treatment:

$$\alpha(A, X) = \frac{2A_j - 1}{P(A_j \mid X)}. \tag{2.26}$$

This reduction confirms that *in the absence of treatment correlations*, our estimator recovers the standard weighting schemes used in single-treatment DML (Chernozhukov et al. 2018), while the general form in(2.20) remains necessary for settings with complex interactions and bundled interventions.

Remark 2 (Computational Efficiency and Treatment-Specific Attribution)**.** *The primitive nuisance parameters, $\eta = (\mu, P)$, represent fundamental properties of the data-generating process, in terms of the global outcome response and the joint treatment assignment mechanism. Consequently, $\hat{\mu}$ and $\hat{P}$ are estimated only once for the entire vector A. While the primitives are shared, the causal contrast of interest varies for each treatment $j \in \{1, \ldots, d\}$. This is handled efficiently through the treatment-specific Riesz representer, $\alpha_j(A, X)$, which is derived algebraically from the shared joint propensity score $\hat{P}$ and the reference policy π. This structure allows the researcher to conduct inference on all d marginal effects without retraining the underlying machine learning models for each specific treatment, providing a scalable solution for high-dimensional combinatorial causal inference.*

We now state the main theorem governing the inference of the tempered estimator. This theorem incorporates the stability conditions discussed in Chernozhukov et al. (2025) to ensure the variance remains bounded.

Theorem 1 (Inference on the Tempered Marginal Estimand at fixed T, δ)**.** *Consider the estimation of the j-th marginal effect for a fixed component $j \in \{1, \ldots, d\}$ and fixed regularization parameters (T, δ). Let $\hat{\theta}_{j,N}$ be the DML estimator solving the efficient score equation $\sum_{i=1}^{N} \psi_j(W_i; \hat{\theta}_{j,N}, \hat{\eta}) = 0$.*

Under Assumptions 1–6, the estimator is asymptotically normal around the tempered proxy target:

$$\sqrt{N}\left(\hat{\theta}_{j,N} - \theta_j(\pi_{T,\delta})\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{j,T,\delta}), \tag{2.27}$$

where the asymptotic variance $\Sigma_{j,T,\delta}$ is determined by the variance of the efficient score for component j:

$$\Sigma_{j,T,\delta} = \mathbb{E}\left[\alpha_{j,T,\delta}^2 \sigma^2(A, X) + (\mu_j(X) - \theta_j(\pi_{T,\delta}))^2\right]. \tag{2.28}$$

Here, $\alpha_{j,T,\delta}$ denotes the Regularized Riesz Representer specific to the j-th marginal shift, and $\sigma^2(A, X) = Var(Y|A, X)$.

The variance decomposition of $\Sigma_{j,T,\delta}$ highlights the fundamental challenge of causal inference for violation of positivity assumption. This first term represents the inflation of variance caused by the fact that we do not observe the complete vector of potential outcomes for any unit (the fundamental problem of causal inference). We only observe $Y(A)$, but we need to estimate $Y(s)$ for unobserved configurations s. This increment of uncertainty reflects (i) reweighting penalty: to recover these unobserved counterfactuals, we re-weight the observed data using the Riesz representer α_0. In the sequel, it scales with $\mathsf{E}[\alpha_0^2]$; (2) relation to rare support:"missing counterfactuals" is distinct from "rare support," but they interact dangerously here. We always miss counterfactuals (even with good overlap). However, when support is rare (i.e., $P(s|X) \approx 0$), we have almost no observational proxies for the counterfactual state s. To compensate, the Riesz representer $\alpha_0 \propto 1/P(A|X)$ becomes unstable, forcing us to place massive weight on the few units we do observe. Using tempering the policy (choosing $T > 1, \delta > 0$), we ensure that the target distribution π does not demand inference on subsets where $P(s|X)$ is negligible. This bounds α_0 in Proposition 2, keeping Σ finite and ensuring statistical stability.

2.3.2 Asymptotic Theorems under T_N, δ_N

Our goal is to maximize the causal signal while minimizing the approximation bias arising from overlap regularisation. To establish the approximation error bound of $\hat{\theta}_{j,N}$ that quantifies bias of estimand by adopting truncated and tempered policy, we require additional regularity conditions.

Assumption 7 (Lipschitz Continuity of Outcome)**.** *Let $\mathcal{A} = \{0,1\}^d$ be the space of treatment configurations endowed with the Hamming distance metric $d_H(a,a') = \sum_{k=1}^{d} \mathbb{I}(a_k \neq a'_k)$. We assume the outcome regression function $\mu(a,x)$ is L-Lipschitz continuous with respect to d_H for all $x \in \mathcal{X}$. That is, for any pair of treatment vectors $a, a' \in \mathcal{A}$:*

$$|\mu(a,x) - \mu(a',x)| \leq L(x) \cdot d_H(a,a'), \tag{2.29}$$

where $L(x)$ is a bounded Lipschitz constant such that $\sup_x L(x) \leq L_{max} < \infty$.

If the outcome surface $\mu(a,x)$ is smooth, the error introduced by averaging over the tempered support is bounded by $O(T-1)$. Thus, the "cost" paid in bias is mathematically controlled by the smoothness of the causal mechanism, allowing for valid inference in regimes where unregularized (atomic) estimators have undefined variance.

Assumption 8 (Propensity tail bound)**.** *The probability mass of the propensity scores near zero decays polynomially. Specifically, there exist constants $c > 0$ and $\gamma > 0$ such that for small $t > 0$:*

$$\mathbb{P}_X(P(A|X) \leq t) \leq c \cdot t^{\gamma}. \tag{2.30}$$

The parameter γ characterizes the difficulty of the overlap problem: larger γ implies fewer units have near-zero propensity.

Theorem 2 (Approximation Error Bound)**.** *Let $\theta(a^*)$ be the target causal effect at the local mode a^*, and let $\theta(\pi_{T,\delta})$ be the feasible estimand under the truncated tempered policy defined in (2.12). Under Assumptions 7 and 8, and assuming bounded outcomes $|Y| \leq B$,*

the approximation bias is bounded by:

$$\textit{Bias}(T,\delta) = |\theta(\pi_{T,\delta}) - \theta(a^*)| \leq \underbrace{L_{max} \cdot d \cdot \exp\left(-\frac{\Delta}{T}\right)}_{\textit{Tempering Cost (Smoothing)}} + \underbrace{2B \cdot c \cdot \delta^{\gamma}}_{\textit{Truncation Cost (Coverage)}}, \quad (2.31)$$

where $\Delta = -\log(\rho) > 0$ *represents the energy gap, with* $\rho = \frac{\max_{a \neq a^*} P(a)}{P(a^*)} < 1$.

The conditional mode $a^*(X)$ corresponds to the limit of the tempered policy as $T \to 0$. Thus, comparing $\pi_{T,\delta}$ to a^* precisely quantifies the "smoothing cost" incurred by inflating the temperature to $T > 1$ for variance control, along with truncation cost by δ. Theorem 2 proves that this smoothing bias vanishes exponentially fast ($\exp(-\Delta/T)$), ensuring that the stability gains of $T > 1$ come at a negligible asymptotic cost compared to the primary truncation error, that will be discussed in the following theorem.

Lemma 1 (Optimal Truncation Rate)**.** *Let the estimation error be decomposed into a variance component scaling with* $(N\delta_N)^{-1}$ *and a squared bias component scaling with* $\delta_N^{2\gamma}$ *(under Assumption 8). The truncation sequence* δ_N *that minimizes the Mean Squared Error (MSE) rate is given by:*

$$\delta_N \asymp N^{-\frac{1}{2\gamma+1}}. \quad (2.32)$$

The optimal truncation rate $\delta_N \asymp N^{-\frac{1}{2\gamma+1}}$ in Lemma 1 quantifies the exact speed at which we must approach the "danger zone" of identification.

Theorem 3 (Rate of convergence)**.** *Let* $\hat{\theta}_{j,N}$ *be the estimator satisfying the asymptotic normality condition of Theorem 4. By Lemma 1, consider a sequence of regularization parameters* (T_N, δ_N) *satisfying the following rates:*

$$\delta_N \asymp N^{-\frac{1}{2\gamma+1}}, \quad \textit{and} \quad T_N \to 0 \textit{ such that } T_N \cdot \log N \to \infty.$$

Then, the total estimation error relative to the ideal atomic target satisfies:

$$|\hat{\theta}_{j,N} - \theta_j(a^*)| = \underbrace{O_p\left(\frac{1}{\sqrt{N\delta_N}}\right)}_{\textit{Stochastic Variance}} + \underbrace{O(\delta_N^{\gamma} + e^{-1/T_N})}_{\textit{Approximation Bias}}. \quad (2.33)$$

Under this regularised growth rate for (T_N, δ_N)*, the estimator achieves the* ***minimax*** *convergence rate:*

$$|\hat{\theta}_{j,N} - \theta_j(a^*)| = O_p\left(N^{-\frac{\gamma}{2\gamma+1}}\right). \quad (2.34)$$

Our theoretical analysis establishes that the proposed Tempering framework is not merely a heuristic for variance reduction, but a statistically optimal strategy for recovering causal effects under structural positivity violations. The derivation of Theorem 3 yields three fundamental insights: First, we show *Minimax Optimality under Limited Overlap.* No estimator can achieve the standard parametric rate $N^{-1/2}$ without stronger (and likely unrealistic) assumptions about the overlap geometry.

Secondly, we show how crucial the optimal truncation rate in Lemma 1 is in the derived rate. If γ is large (easy overlap), δ_N decays rapidly, and we recover near-parametric speeds.

If $\gamma \to 0$ (severe violation), δ_N decays slowly, reflecting the inherent difficulty of the scientific question. Our framework automatically adapts to this hardness parameter. While γ is a theoretical quantity governing the minimax rate, in practice we do not assume a specific value. Instead, we select the truncation threshold δ adaptively via cross-validation (or by minimizing the estimated efficient influence function variance), which allows the estimator to automatically calibrate to the unknown overlap geometry of the dataset. Unlike standard truncation which incurs purely "negative" bias (loss of coverage), our dual regularization (T, δ) incurs "constructive" bias that can vanish if the sequence of (T_N, δ_N) can behave well.

Theorem 4 (Asymptotic Normality of the Tempered DML (T-DML) Estimator)**.** *Consider the estimation of the j-th marginal effect using a sequence of regularization parameters (T_N, δ_N) satisfying the conditions of Theorem 3. Let $\hat{\theta}_{j,N}$ be the DML estimator defined by the sample average of the efficient score constructed with these parameters:*

$$\hat{\theta}_{j,N} = \frac{1}{N} \sum_{i=1}^{N} \psi_j(W_i; \hat{\eta}_N, T_N, \delta_N), \tag{2.35}$$

where $\hat{\eta}_N$ denotes the nuisance parameters estimated via cross-fitting. Under the Identification Assumptions 1–6 and the Regularity Assumptions 7–8, as $N \to \infty$, the centered and scaled estimator satisfies:

$$\sqrt{N\delta_N} \left(\hat{\theta}_{j,N} - \theta_j(a^*) - \textit{Bias}_j(T_N, \delta_N) \right) \xrightarrow{d} \mathcal{N}(0, \sigma_\infty^2), \tag{2.36}$$

where $\textit{Bias}_j(T_N, \delta_N) = \theta_j(\pi_{T_N,\delta_N}) - \theta_j(a^*)$ is the deterministic approximation error characterized in Theorem 2. The limiting variance is given by $\sigma_\infty^2 = \lim_{N\to\infty} \delta_N \cdot \textit{Var}(\psi_j)$.

Remark 3 (Connection to Algorithm 1)**.** *The estimator $\hat{\theta}_{j,N}$ analyzed in Theorem 4 corresponds empirically to the output of Algorithm 1 when the input parameters are set to the optimal tuning values $(T, \delta) = (T_N, \delta_N)$. While Algorithm 1 computes the finite-sample estimate $\hat{\theta}_{j,N}$ for a fixed configuration, Theorem 4 guarantees that this output approximates the ideal atomic target $\theta_j(a^*)$ with a stable Gaussian limit if the parameters are tuned according to the sample size.*

2.3.3 Inference with Monte Carlo Approximation

Theorem 1 characterizes the asymptotic distribution of the tempered estimator assuming the target parameter $\theta_j(\pi_{T,\delta})$ can be evaluated exactly. However, in high-dimensional settings involving continuous or multi-valued treatments, computing the integral $\int_{\mathcal{A}} \pi_{T,\delta}(a|x)\mu(a,x)da$ is analytically intractable. Algorithm 1 overcomes this by approximating the integral via Monte Carlo simulation (Step 2a). We replace the exhaustive sum with a Monte Carlo approximation by drawing M trajectories from the tempered policy. This introduces an additional source of stochasticity: the simulation noise. The following theorem extends Theorem 1 to explicitly account for this computational uncertainty, treating the regularization parameters (T, δ) as fixed for the purpose of finite-sample inference.

Unbiased Simulator of the Score For each observation i in the sample, we generate M independent draws of background configurations $\{s_{im}\}_{m=1}^{M}$ such that each $s_{im} \sim \pi(\cdot|X_i)$. The simulated version of the plug-in term is:

$$\tilde{\Gamma}(X_i;\mu) = \frac{1}{M}\sum_{m=1}^{M}\left[\mu(s_{im}+e_j, X_i) - \mu(s_{im}, X_i)\right]. \tag{2.37}$$

By the Law of Large Numbers, $\tilde{\Gamma}(X_i;\mu) \xrightarrow{p} \Gamma(X_i;\mu)$ as $M \to \infty$. Because π is a known (tempered) distribution given (T,δ), we can sample from it efficiently using (2.12). The final simulated T-DML estimator $\tilde{\theta}_{j,N}$ is the solution to the simulated empirical score:

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{\psi}_{j,i}(\tilde{\theta}_{j,N}) = 0. \tag{2.38}$$

where $\tilde{\psi}_{j,i}(\theta) = \left(\tilde{\Gamma}(X_i;\hat{\mu}) + \hat{\alpha}(A_i,X_i)(Y_i - \hat{\mu}(A_i,X_i)) - \theta\right)$ is simulated tempered score. Because the simulation noise is independent across observations i, the simulated T-DML estimator remains consistent even for a small fixed M as $N \to \infty$, though increasing M reduces the simulation-induced variance.

Theorem 5 (Asymptotic normality of the simulated tempered estimator)**.** *Let $\tilde{\theta}_{j,N}$ be the estimator of (2.38) and computed via Algorithm 1 using M Monte Carlo draws per unit and fixed (T,δ). Suppose the conditions of Theorem 1 hold. Additionally, assume the simulation process satisfies: (1) unbiasedness: conditional on the data W_i and nuisance estimates $\hat{\eta}$, the simulated score $\tilde{\psi}_i$ is an unbiased estimator of the exact efficient score ψ_i:*

$$\mathbb{E}_{sim}[\tilde{\psi}_i \mid W_i, \hat{\eta}] = \psi_i(W_i;\hat{\eta}).$$

(ii) finite simulation variance:

$$\Sigma_{sim} = \mathbb{E}_P\left[\mathit{Var}_{sim}(\tilde{\psi}_i \mid W_i)\right] < \infty.$$

Then, as $N \to \infty$, the stimulated estimator satisfies:

$$\sqrt{N}\left(\tilde{\theta}_{j,N} - \theta_j(\pi_{T,\delta})\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{j,T,\delta} + \frac{1}{M}\Sigma_{sim}\right), \tag{2.39}$$

where $\Sigma_{j,T,\delta}$ is the efficient variance of the exact estimator (defined in Theorem 1), and $\frac{1}{M}\Sigma_{sim}$ represents the variance inflation due to finite computational budget.

The parameter Σ_{sim} quantifies the variability of the stochastic policy π_T conditional on the covariates. Specifically, it is the expected variance of the plug-in contrast when drawn from the truncated policy (to be discussed later):

$$\Sigma_{sim} = \mathbb{E}\left[\mathrm{Var}_{s\sim\pi(\cdot|X)}\left(\hat{\mu}^{(-k)}(s+e_j, X) - \hat{\mu}^{(-k)}(s,X)\right)\right].$$

Intuitively, if the outcome surface $\hat{\mu}$ is flat or the policy π is very concentrated, Σ_{sim} is small. If the policy explores a rugged landscape of treatment effects, Σ_{sim} is large.

Theorem 5 formalizes the behavior of our computational strategy, offering two key practical insights. The first one is additive variance cost. The theorem establishes that the cost of using the Method of Simulated Scores is strictly additive. This decomposition is powerful because it allows the researcher to control the precision budget. Σ is fixed by the dataset size N and the population structure (as established in Theorem 1). However, the second term is purely computational. By increasing the number of internal simulation draws M, we can drive the simulation noise to zero, recovering the semiparametric efficiency bound Σ asymptotically.

2.3.4 Algorithmic Implementation for Tempered-DML Estimation Using Simulation

The proposed tempered-DML (T-DML) algorithm in Algorithm 1 is designed to satisfy two competing theoretical requirements: statistical consistency via Neyman Orthogonality and finite-sample stability via truncated policies.

Orthogonality requires that the noise in the nuisance estimators is uncorrelated with the noise in the outcome Y. By training $\hat{\mu}^{(-k)}$ and $\hat{P}^{(-k)}$ on the auxiliary folds (Step 1) and evaluating the score on the held-out fold $\mathcal{I}_k$ (Step 2), we ensure statistical independence. Consequently, the cross-product of estimation errors—the dominant source of bias—vanishes:

$$\mathbb{E}_P\left[(\hat{\alpha}^{(-k)} - \alpha_0)(\hat{\mu}^{(-k)} - \mu_0)\right] \approx 0.$$

The algorithm thus produces a valid orthogonalized signal $\tilde{\psi}_i$ for every data point, allowing for valid inference despite the high-dimensional complexity of the nuisance models.

Standard inverse probability weighting approaches often employ "clipping" (enforcing $\hat{P} \geq \epsilon$) to prevent variance explosion. However, this introduces undefined bias. Instead, we adopt a tempered policy approach that regularises overlap to facilitate identification. This is implemented in the algorithm via two mechanisms: (i) Rejection Sampling (Step 2a): we only simulate background counterfactuals s that are empirically supported by the data ($\hat{P} \geq \delta$) given T, the overlap regularisation parameter. This ensures the plug-in term $\tilde{\Gamma}_i$ represents a realistic intervention. (ii) bounded Riesz weight (Step 2b): for the correction term, if a unit's observed treatment A_i has negligible probability ($\hat{P} < \delta$), we set $\hat{\alpha}_i = 0$. This ensures the Riesz representer is strictly bounded by definition, guaranteeing finite variance and effectively excluding impossible units from the de-biasing step.

Finally, the score $\tilde{\psi}_i$ is composed of two parts: the plug-in estimator $\tilde{\Gamma}_i$ and the bias-correction term $\mathcal{C}_i$. We give some insights on these two terms. (i) $\tilde{\Gamma}_i$ captures the direct effect estimated by the outcome model. (ii) $\hat{\alpha}_i(Y_i - \hat{\mu}(A_i, X_i))$ estimates the residual bias. If $\hat{\mu}$ under-predicts the outcome for a treated group, the residual $(Y - \hat{\mu})$ will be positive. The Riesz weight $\hat{\alpha}$ scales this residual to exactly offset the error in the plug-in term.

The proposed tempered-DML (T-DML) algorithm offers distinct advantages over naive aggregation. We discuss the computational complexity and theoretical consistency. The primary bottleneck in estimating marginal effects for high-dimensional interactive models is the summation over the background space $\mathcal{A}_{-j}$. An exact calculation requires evaluating the outcome model 2^{d-1} times for each unit. For $d = 30$, this implies over 500 million evaluations per data point, which is computationally infeasible. By contrast, Monte Carlo integration reduces the complexity from exponential $O(N \cdot 2^d)$ to linear $O(N \cdot M)$. Since

Algorithm 1 Tempered Double Machine Learning (T-DML) for Marginal Effects

Require: Data $Z = \{Y_i, A_i, X_i\}_{i=1}^N$, Folds K, Monte Carlo Draws M, Fixed Parameters (T, δ)
Ensure: Marginal Effect Estimator $\tilde{\theta}_{j,N}$

Partition: Randomly split indices $\{1, \dots, N\}$ into K disjoint folds $\mathcal{I}_1, \dots, \mathcal{I}_K$.
for $k \leftarrow 1$ **to** K **do**
  Step 1: Nuisance Estimation (on Auxiliary Sample $\mathcal{I}_k^c$)
  Using training data $i \notin \mathcal{I}_k$, estimate:
    1. Outcome Regression: $\hat{\mu}^{(-k)}(a, x) \approx \mathbb{E}[Y|A = a, X = x]$.
    2. Propensity Score: $\hat{P}^{(-k)}(a \mid x) \approx P(A = a|X = x)$.
    3. Tempered Policy: $\pi_T^{(-k)}(a|x) \propto \hat{P}^{(-k)}(a|x)^{1/T}$.
  Step 2: Score Evaluation (on Evaluation Sample $\mathcal{I}_k$)
  for each unit $i \in \mathcal{I}_k$ **do**
    (a) Target Parameter Prediction (Reg. Adjustment):
    ▷ Estimate $\Gamma(X_i) = \sum \pi(a_{-j})[\mu(1, a_{-j}) - \mu(0, a_{-j})]$
    Initialize $\tilde{\Gamma}_i = 0$.
    for $m \leftarrow 1$ **to** M **do**
      Draw proposal $s \sim \pi_T^{(-k)}(\cdot \mid X_i)$.
      if $\hat{P}^{(-k)}(s \mid X_i) \geq \delta$ **then**
        Compute marginal contrast: $\Delta_m = \hat{\mu}^{(-k)}(s + e_j, X_i) - \hat{\mu}^{(-k)}(s, X_i)$.
        $\tilde{\Gamma}_i \leftarrow \tilde{\Gamma}_i + \Delta_m$.
      end if
    end for
    Set $\hat{\theta}_{target}(X_i) = \tilde{\Gamma}_i / M$.
    (b) Riesz Correction (Bias Correction):
    ▷ Construct Marginal Riesz Weight $\hat{\alpha}_j$
    Evaluate observational propensity $p_i = \hat{P}^{(-k)}(A_i \mid X_i)$.
    if $p_i < \delta$ **then**
      Set weight $\hat{\alpha}_i = 0$.
    else
      $\hat{\alpha}_i = \frac{\pi_T^{(-k)}((A_i)_{-j}|X_i)}{p_i} \times (2A_{ij} - 1)$.
    end if
    Compute correction term: $\mathcal{C}_i = \hat{\alpha}_i \times (Y_i - \hat{\mu}^{(-k)}(A_i, X_i))$.
    (c) Compute Efficient Score:
    $\hat{\psi}_i = \hat{\theta}_{target}(X_i) + \mathcal{C}_i$.
  end for
end for
Step 3: Aggregation
Compute final global estimator: $\tilde{\theta}_{j,N} = \frac{1}{N} \sum_{i=1}^N \hat{\psi}_i$.

the simulation error decreases at a rate of $1/\sqrt{M}$ independent of the dimension d, we can approximate the integral efficiently even in very high-dimensional spaces. This allows the estimator to scale to settings with hundreds of binary treatments, provided the outcome model $\hat{\mu}$ allows for efficient batch inference.

A crucial theoretical feature of this algorithm is that consistency does not require

$M \to \infty$. Because the simulated score $\tilde{\psi}_{ij}$ is an unbiased estimator of the true efficient score ψ_{ij} (conditional on the data), the Law of Large Numbers applied over the sample size N ensures convergence to the true parameter:

$$\tilde{\theta}_{j,N} \xrightarrow{p} \mathbb{E}_P[\tilde{\psi}(W)] = \mathbb{E}_P[\mathbb{E}_{sim}[\tilde{\psi}(W)|W]] = \mathbb{E}_P[\psi(W)] = \theta_j(\pi_{T,\delta}).$$

2.4 Simulation Study

In this section, we investigate the finite-sample performance of our proposed estimators for unconditional marginal treatment effects in multiple treatment settings through an extensive simulation study[1]. We evaluate the performance of our T-DML Algorithm 1 against established benchmarks, including Double Machine Learning (DML) approaches and Targeted Maximum Likelihood Estimation (TMLE). Performance is assessed using standard causal estimation metrics (RMSE, standard deviation, confidence interval length, and coverage) to evaluate the accuracy and inferential validity of the marginal treatment effect estimates.

2.4.1 Simulation Setting

One obstacle in evaluating causal estimators for marginal treatment effects is that the true value of the causal parameter θ_0 is not observed in observational studies. To enable a fair evaluation, we design a flexible data generating process (DGP) that generates d binary treatment variables $A = (A_1, \ldots, A_d)$ with known ground truth effects. The DGPs are adapted from Belloni et al. (2017)[2]. Covariates $X \in \mathbb{R}^p$ are drawn from a multivariate normal distribution with Toeplitz covariance structure $\Sigma_{ij} = 0.5^{|i-j|}$. Each treatment A_j is assigned via logistic propensity scores with treatment-specific coefficients: $P(A_j = 1|X) = \text{logit}^{-1}(X^\top \beta_j \cdot c_A)$, where β_j follows a decaying coefficient structure with $\beta_{j,k} \propto 1/k^2$, and c_A is a scaling factor calibrated to achieve a target R^2_A for treatment predictability. For the *Simple* and *3-Way* DGPs, treatments are assigned independently. For the *2-Way* and *Complex* DGPs, pairwise correlation ($\rho = 0.5$) is induced between adjacent treatment pairs $(A_1, A_2), (A_3, A_4), \ldots$ using a Gaussian copula structure.

The outcome is generated as

$$Y = \sum_{j=1}^{d} \theta_j A_j + f_{\text{int}}(A) + X^\top \beta_y \cdot c_y \cdot \sum_{j=1}^{d} A_j + \epsilon, \tag{2.40}$$

where $\epsilon \sim N(0, 1)$, θ_a are randomly drawn baseline coefficients fixed across simulations, and c_y is calibrated to achieve $R^2_y = 0.9$. The interaction function $f_{\text{int}}(D)$ varies across DGP variants. Crucially, the true unconditional marginal treatment effect for treatment j,

$$\theta^*_j = \mathbb{E}[Y(A_j = 1, A_{-j}) - Y(A_j = 0, A_{-j})], \tag{2.41}$$

depends not only on the baseline coefficient θ_j but also on the realized treatment configurations of the other treatments A_{-j} through the interaction terms. Consequently, the

[1]The simulation is executed on an HPC cluster in parallel, using different seeds for the DGPs.

[2]The DGP is available for single binary treatments at DoubleMLIRMData.

true marginal effects vary across simulated datasets and are computed exactly for each dataset by enumerating counterfactual outcomes for each observation and averaging over the sample.

We consider four DGP variants that vary along two dimensions, treatment dependence structure and outcome interaction complexity: The *Simple* DGP features independent treatments with no outcome interactions. The *2-Way* DGP introduces pairwise correlated treatments with two-way outcome interactions $\sum_j A_{2j} \cdot A_{2j+1}$. The *3-Way* DGP maintains independent treatments but includes three-way outcome interactions $A_1A_2A_3 + A_1A_4A_6$. Finally, the *Complex* DGP combines pairwise correlated treatments with both two-way and three-way outcome interactions. Notably, the *Simple* and *3-Way* DGPs share identical treatment assignment mechanisms, as do the *2-Way* and *Complex* DGPs. This design allows us to separately assess the impact of treatment dependence on propensity score estimation difficulty versus outcome interaction complexity on treatment effect estimation accuracy.

Beyond the baseline configuration, we conduct an extensive sensitivity analysis to assess the robustness of our findings across varying estimation challenges. Table 2.2 summarizes the parameter grid, with baseline values indicated in bold. We systematically vary each parameter while holding the others fixed at their baseline levels.

Table 2.2: Simulation parameter grid for sensitivity analysis. Baseline values are indicated in bold.

Parameter	Values	Baseline
Sample size (n)	1000; 2,000; 4,000; 6,000; 10,000	**2,000**
Treatment dimensionality (d)	4; 6; 8; 10	**6**
Covariate dimensionality (p)	5; 10; 20; 50	**10**
Treatment predictability (R^2_A)	0.1; 0.3; 0.5; 0.7; 0.9	**0.5**
Outcome predictability (R^2_y)	0.1; 0.3; 0.5; 0.7; 0.9	**0.9**

Varying the sample size n allows us to evaluate finite-sample performance and convergence behavior. The treatment dimensionality d assesses scalability, as the number of potential treatment combinations grows exponentially as 2^d, creating increasing sparsity in the treatment space. Covariate dimensionality p examines how the complexity of the confounding structure affects estimation performance. Finally, treatment predictability R^2_A captures varying degrees of overlap, ranging from near-random treatment assignment ($R^2_A = 0.1$) to highly predictable treatments ($R^2_A = 0.9$) that challenge the positivity assumption. This sensitivity analysis enables us to characterize the conditions under which each estimator performs well and to identify potential failure modes in more challenging scenarios.

2.4.2 Estimators and Implementation

Proposed Method

We implement our proposed simulated score estimator (Algorithm 1) with the following implementation choices. The propensity score model estimates the joint treatment distribution $\hat{p}(S|X)$ over all 2^d treatment combinations using multinomial logistic regression with

L_1 regularization. For subset sampling, we draw $R = 100$ treatment configurations from the estimated propensity model using a Gibbs sampler with adaptive temperature scaling. The Gibbs sampler iterates through each treatment dimension $k \neq j$ (excluding the primary treatment of interest), sampling from the conditional distribution while maintaining $A_j = 0$ to generate valid counterfactual configurations. We use a burn-in period of 500 iterations before collecting R samples. We use $K = 5$ folds for cross-fitting to avoid overfitting and ensure Neyman orthogonality.

Outcome Models

To assess robustness to model specification, we evaluate each method with different outcome model specifications $\hat{\mu}(A, X)$, where $A = (A_1, \ldots, A_d)$ denotes the treatment vector and X the covariates. The *Linear* model uses linear regression augmented with covariate-treatment interactions and all pairwise treatment interactions, with standardized features. *Linear3W* further extends this specification to include all three-way treatment interactions $\sum_{j<l<m} A_j A_l A_m$, capturing higher-order treatment synergies. The *NeuralNet* model employs a feedforward neural network with separate pathways for covariates, treatments, and explicit pairwise treatment interactions; the architecture consists of parallel embedding layers (covariate pathway: $p \to 32$; treatment pathway: $d \to 32$; interaction pathway: $\binom{d}{2} \to 16$), concatenated and passed through a hidden layer ($80 \to 32$ units, ReLU activation, 20% dropout) before the output layer, with training via Adam optimization for 100 epochs at batch size 256. Finally, *LGBM* uses LightGBM gradient boosting (Ke et al. 2017) with default hyperparameters, allowing the tree ensemble to automatically capture nonlinearities and interactions.

Comparison Methods

We compare our proposed estimator against several established approaches that represent current methodological standards for causal effect estimation. The T-Bench method represents a naive version of the proposed T-DML estimator that samples only subsets with single active treatments and therefore ignores treatment interactions, providing a baseline for evaluating the importance of accounting for treatment interactions. DML-IRM applies Double Machine Learning with the Interactive Regression Model, extended to handle multiple treatments by running the procedure separately for each treatment. DML-PLR uses the Partially Linear Regression model (Chernozhukov et al. 2018, Bach et al. 2022), which assumes an additive structure and may therefore be misspecified when treatment interactions are present. Like DML, the Targeted Maximum Likelihood estimator (TMLE) (Gruber and Laan 2010) is based on influence functions, but it differs in being an iterative procedure in which the outcome regression is first estimated and the ATE estimate is then updated using a propensity-score-based targeting step. All comparison methods rely on the same outcome model specifications (Linear, Linear3W, NeuralNet, LGBM) to ensure a fair comparison across estimators.

Performance Metrics

We evaluate estimator performance using several metrics computed over 100 simulation replications and aggregated across all d treatments. Let $\hat{\theta}_j^{(r)}$ denote the estimate for

treatment j in replication r, and $\theta_j^{(r)}$ the corresponding true marginal effect, which varies across replications due to the interaction structure. Root Mean Squared Error (RMSE) is computed as

$$\sqrt{\frac{1}{100 \cdot d} \sum_{r=1}^{100} \sum_{j=1}^{d} (\hat{\theta}_j^{(r)} - \theta_j^{(r)})^2}$$

, measuring overall estimation accuracy. We report the empirical standard deviation of estimates across replications and treatments, $\text{Std}(\{\hat{\theta}_j^{(r)}\}_{r,j})$. Confidence interval length represents the average length of 95% confidence intervals across all replications and treatments. We additionally apply a Bonferroni correction to account for multiple testing. Coverage probability measures the proportion of replication-treatment pairs where the 95% confidence interval contains the true parameter $\theta_j^{(r)}$, with nominal coverage of 0.95. Confidence intervals for our proposed estimator are constructed using the asymptotic variance derived in Section 2.3, while comparison methods use their respective standard inferential procedures.

2.4.3 Simulation Results

Figure 2.1 summarizes the RMSEs across the different DGPs and outcome models for our baseline setting of n = 2,000 observations, A = 6 treatments, R2_y = 0.9, and R2_A = 0.5. A more detailed overview of the results[3] together with a sensitivity analysis for

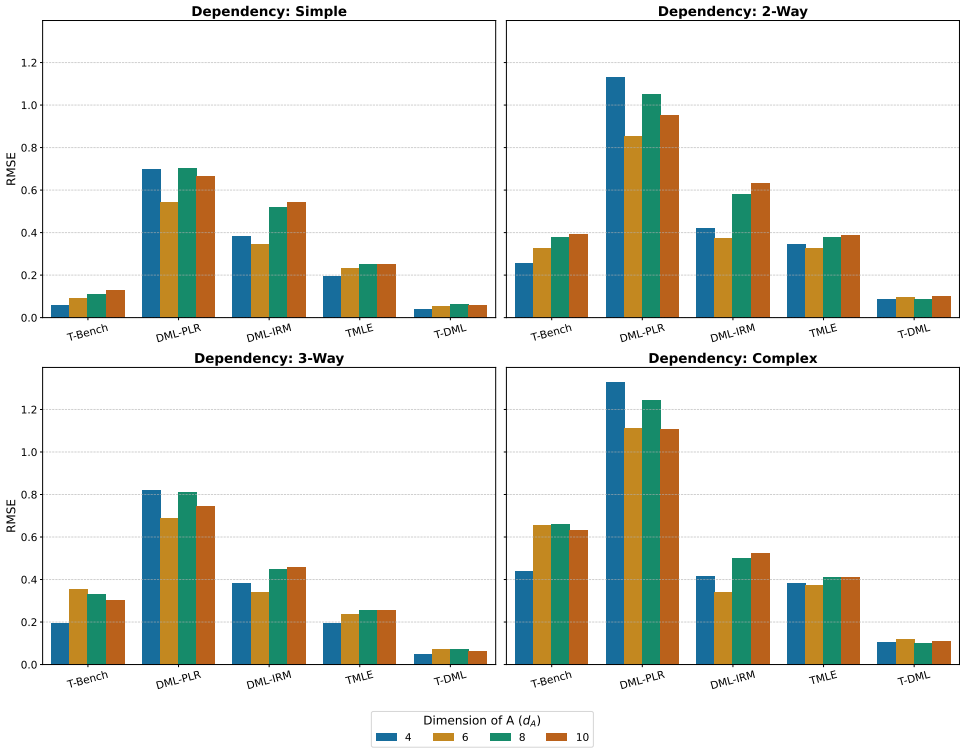

Figure 2.1: RMSE for varying treatments dimension and for all Models and DGPs. The outcome regression is a simple linear regression with two-way treatment interactions.

[3]Table 2.3 provides a summary for the baseline parameter combination.

alternative DGP specifications is provided in Appendix 2.7.2. Our proposed T-DML estimator performs best in the Simple DGP without treatment interactions, achieving the lowest RMSE (0.02–0.03) while maintaining proper coverage (0.90–1.00) across all outcome models. This confirms that the simulated-score approach is highly effective when the DGP structure is favorable. As DGP complexity increases, performance deteriorates for all estimators. Nevertheless, our estimator remains competitive in the 2-Way and 3-Way DGPs: its RMSE stays below 0.08 even in the most complex setting, consistently outperforming the benchmark models. Coverage patterns are more sensitive to model choice. With correctly specified linear outcome models, our estimator achieves near-perfect coverage ($\approx$ 1.00). The coverage declines mildly (0.88 - 0.93) when more flexible models, such as LGBM or neural networks, are used. For the linear outcome regression in figure 2.2, DML-IRM and DML-PLR estimators show solid performance in the simple DGPs, with coverage around 0.97. However, their coverage drops sharply (0.33–0.45) in the correlated treatments DGPs *2-Way* and *Complex*, as standard DML methods cannot capture higher-order treatment interactions when estimating marginal effects. The naive benchmark performs poorly whenever treatment interactions are present, with coverage falling to zero in many scenarios. This highlights the need for specialized estimators in multi-treatment settings. Finally, increased flexibility in the outcome model does not

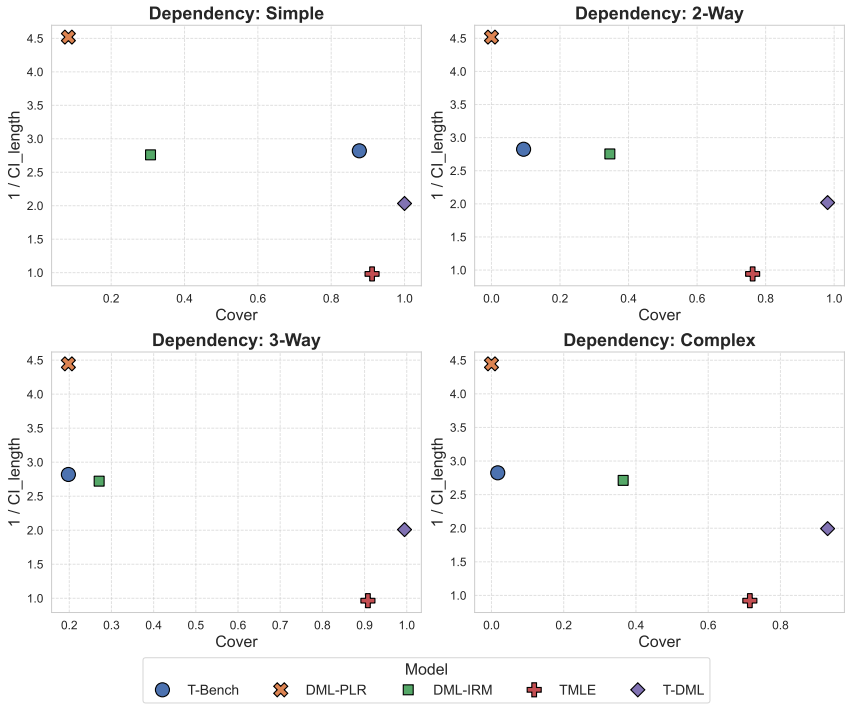

Figure 2.2: Inverse Confidence Interval Width and Coverage over all four DGPs.

automatically translate into better performance. In many cases, simple linear models yield better coverage and similar RMSE when paired with our proposed estimator.

2.4.4 Impact of Propensity Model Misspecification on Subset Sampling

We also investigate why performance varies across DGPs. The logistic-L1 propensity model is correctly specified for independent treatment assignment (None and Intermediate DGPs), but misspecified for correlated treatment assignment (Simple and Complicated DGPs). This misspecification directly affects subset sampling within our T-DML. The degree of confounding in the treatment assignment is influenced by the underlying correlation structure among treatments. This relationship is reflected in the mean effective support metric, which measures the average number of treatment subsets receiving more than 1% probability mass under the estimated propensity model. As shown in Figure 2.3, this metric separates the two classes of data-generating processes: independent treatments yield a mean effective support of 46.40, whereas correlated treatments yield only 36.44. This reduction indicates that the misspecified propensity model concentrates probability mass on a smaller subset of treatment combinations, failing to represent portions of the treatment space during subset sampling. As treatment confounding increases, this issue becomes more pronounced across all DGPs, and the effective subset sample size deteriorates accordingly. Such concentration is problematic for the simulated-score estimator, which relies on broad exploration of treatment combinations. When the propensity model mischaracterizes the joint treatment distribution, the Gibbs sampler may under-sample important regions, resulting in bias and lower coverage. The impact is especially severe if the outcome model is flexible, allowing overfitting to a few subsets, as can be seen in figure 2.4 for LightGBM as the outcome learner. Our

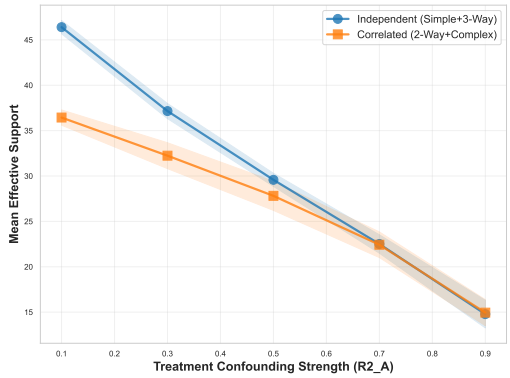

Figure 2.3: Mean effective support for varying Treatment Confounding Strength. DGPs are grouped, as the treatment assignment is the same for the DGPs *Simple & 3-Way* as well as *2-Way & Complex*.

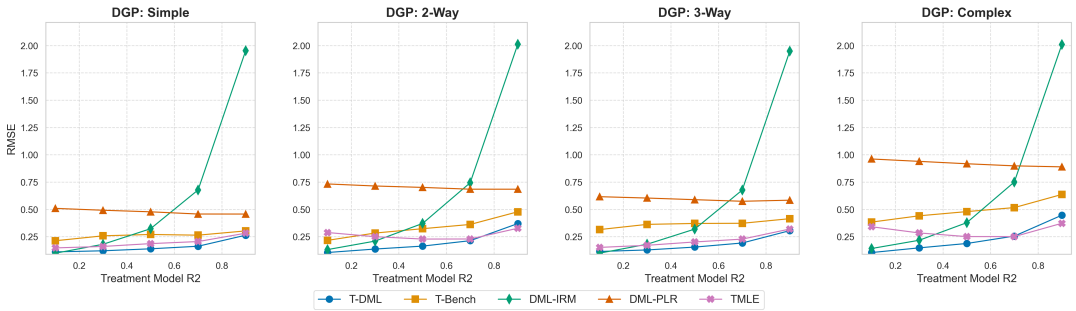

Figure 2.4: RMSE for different levels of treatment confounding strength with LightGBM as outcome regression.

adaptive temperature scaling partially mitigates this issue by encouraging wider exploration when the propensity distribution becomes overly concentrated. However, the simulation results show that this mechanism cannot fully offset misspecification, especially when combined with flexible outcome models that contribute additional estimation noise.

2.5 Empirical Application: Lifestyle Choices and Circumstances and Their Effect on Blood Pressure Risk

Data and Research Question

The National Health and Nutrition Examination Survey (NHANES) provides comprehensive health data collected by the Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) (2024) through interviews, physical examinations, and laboratory tests. This rich dataset offers an ideal setting to study the complex interplay between multiple lifestyle factors and health outcomes, particularly high blood pressure. Hypertension affects approximately 45% of adults in the United States and is a major risk factor for cardiovascular disease, stroke, and kidney failure. Our research investigates the marginal effects of various lifestyle factors on systolic blood pressure in a population not currently taking antihypertensive medication. This focus addresses a crucial public health question: which modifiable risk factors most significantly influence blood pressure in individuals who have not yet initiated pharmacological treatment? Understanding these relationships can inform targeted prevention strategies for at-risk populations before they require medication.

Data Structure and Covariates

The analysis utilizes data from NHANES cycles spanning 2015-2023. We focus on adult participants aged 20 years and older with complete data on blood pressure measurements, lifestyle factors, and relevant covariates. After applying exclusion criteria for missing treatment data and current use of blood pressure medication, our analytical sample comprises 9,737 participants. The dataset includes comprehensive information across several domains that serve as covariates in our analysis. Demographic factors include age, gender, race/ethnicity, Hispanic status, marital status, and household size. Socioeconomic indicators encompass education level, occupation category, and health insurance status. Dietary factors are captured through detailed 24-hour dietary recall data, including sodium, potassium, fiber, and caffeine intake, as well as omega-3 fatty acids and sodium-potassium ratio. Dietary habits are further characterized by salt usage type and frequency at the table.

Treatment Variables

The treatment variables of interest include eight modifiable lifestyle factors and health conditions that previous literature has associated with blood pressure regulation. Each treatment is coded as binary, indicating the presence or absence of the risk factor, with binarization based on established clinical thresholds and survey instruments. Smoking status is defined as current smoking of 100 or more cigarettes in one's lifetime combined with current smoking frequency. Heavy drinking uses gender-specific thresholds: more than one drink per day for men and more than half a drink per day for women. Obesity is defined as body mass index ≥ 30 kg/m^2. Sedentary behavior is indicated by six or more hours per day of sedentary activity. Moderate recreational activity requires at least 10 minutes per session and a minimum of one session per week. Diabetes status is based on

self-reported diagnosis. Sleep disorders are indicated by six or fewer hours of sleep per night. Depression is defined by a PHQ-9 score of 15 or higher.

These eight treatment variables represent a comprehensive set of modifiable risk factors that may interact in complex ways to influence blood pressure. The outcome variable is systolic blood pressure measured during the physical examination component of NHANES, using the average of up to three consecutive readings following standardized protocols.

Visualizing Complex Treatment Interactions

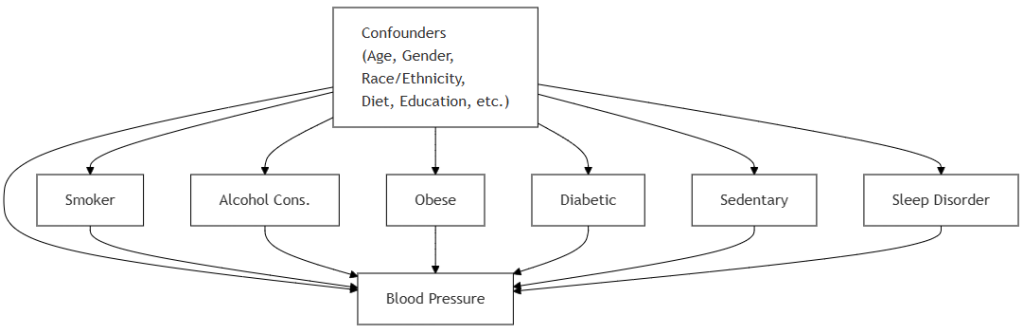

Figure 2.5: Basic directed acyclic graph showing multiple treatments affecting blood pressure with common confounding. This structure represents the standard multiple treatment scenario without accounting for treatment interactions.

The complexity of treatment interactions in this real-world setting is illustrated through directed acyclic graphs (DAGs). Figure 2.5 presents a simplified DAG showing the basic structure with multiple treatments affecting blood pressure, all influenced by common confounders such as age, genetics, and socioeconomic factors. This structure represents the standard multiple treatment scenario where each treatment has both direct effects and effects mediated through other treatments. Figure 2.6 extends this framework to visualise underlying treatment interactions, reflecting the reality that lifestyle factors often cluster and interact in complex ways. Smoking, alcohol consumption, and obesity form one cluster of interrelated behaviours, while diabetes and sleep disorders demonstrate another interaction pattern. These interactions create challenges for traditional estimation approaches that assume additive or independent treatment effects.

Methodological Challenge: Blood Pressure Medication

A methodological challenge arises from the presence of blood pressure medication in the study population. Medication can act simultaneously as a confounder, mediator, and collider. Lifestyle factors influence whether and when medication is prescribed, and current blood pressure is strongly correlated with past levels that triggered treatment. Simply controlling for medication would therefore introduce collider bias, while omitting it would violate the unconfoundedness assumption. A practical solution is to analyze subpopulations with and without medication separately. Here, we focus on individuals not taking blood pressure medication, as detailed information on specific medications is limited. Restricting the analysis to participants not taking blood pressure medication yields more plausible lifestyle-related differences in blood pressure, as shown in Figure 2.7. This restricted analysis better reflects the natural relationships between lifestyle factors and blood pressure

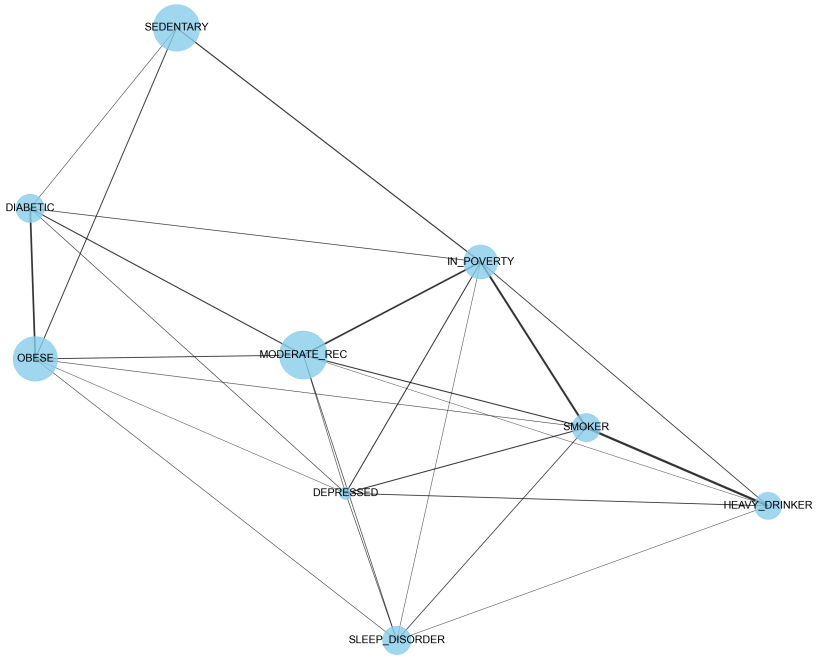

Figure 2.6: Treatment relationship network in NHANES. Node size represents prevalence; edges show correlations. Key patterns are: strong smoking-alcohol correlation, central obesity connecting multiple treatments, and sedentary-obesity linkage. These correlations challenge independent treatment assumptions.

in the absence of pharmacological intervention. This approach creates a study population that is younger and potentially healthier than the general hypertensive population.

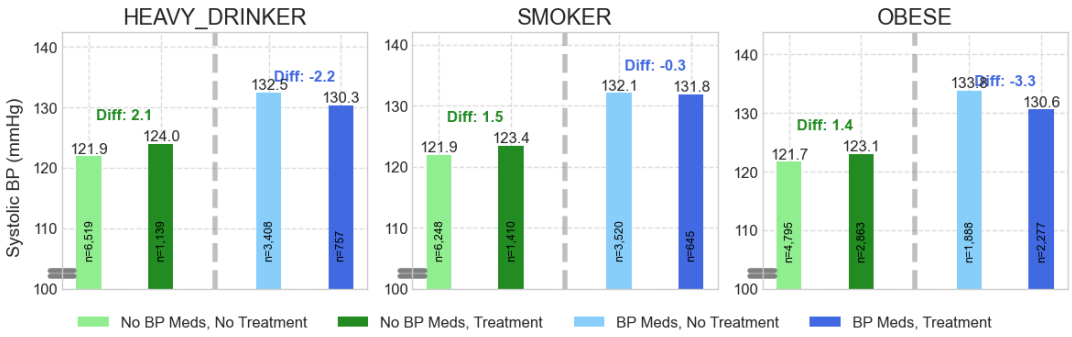

Figure 2.7: Simple outcome differences grouped by whether participants are taking blood pressure medication.

However, this subpopulation represents individuals who may be future candidates for blood pressure medication, making the identification of modifiable risk factors particularly relevant for prevention efforts. Figure 2.8 illustrates this challenge by showing the age distribution stratified by medication status. Participants taking blood pressure medication are substantially older (mean age 61.2 years) compared to those not on medication (mean age 46.7 years). This age difference reflects both the natural progression of hypertension with age and the clinical decision to initiate medication based on blood pressure levels and cardiovascular risk factors.

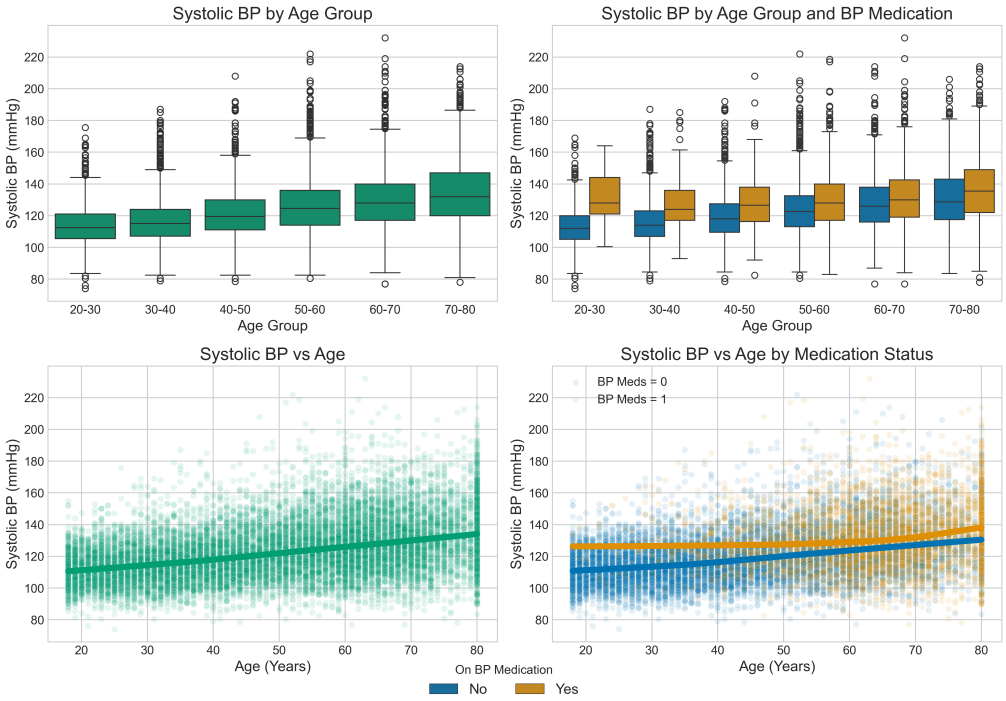

Figure 2.8: Age distribution of NHANES participants by blood pressure medication status. The substantial age difference highlights the confounding role of age and the selection bias introduced by conditioning on medication status.

Estimation Results

We apply our proposed simulated score estimator to estimate unconditional marginal average treatment effects for each lifestyle factor on systolic blood pressure. Results are highly consistent across all four outcome learners. Obesity shows the strongest and most robust association with systolic blood pressure, with effects ranging from about 1.84 mmHg (90% CI: 1.04, 2.63) to 2.28 mmHg (90% CI: 1.51, 3.05). Smoking also displays a stable positive effect across learners, with estimates between 1.17 mmHg (90% CI: 0.07, 2.26) and 1.29 mmHg (90% CI: 0.21, 2.38). Sedentary behavior shows a meaningful negative association in the linear interaction models; for example, -0.98 mmHg (90% CI: -1.73, -0.23). Depression exhibits a significant negative effect only in the neural learner: -2.81 mmHg (90% CI: -4.21, -1.41). Diabetes, heavy drinking, sleep disorders, and moderate recreational activity do not show statistically meaningful effects in any learner, as their 90% intervals include zero. Overall, the results are stable across learners, and the pattern of significant effects is consistent. The magnitudes are likely conservative because the analysis focuses on a younger and healthier subpopulation that is not yet on blood pressure medication.

These results highlight the varying contributions of different lifestyle factors to blood pressure elevation. The substantial effect of obesity is consistent with established physiological mechanisms linking adiposity to increased blood volume, cardiac output, and peripheral resistance.

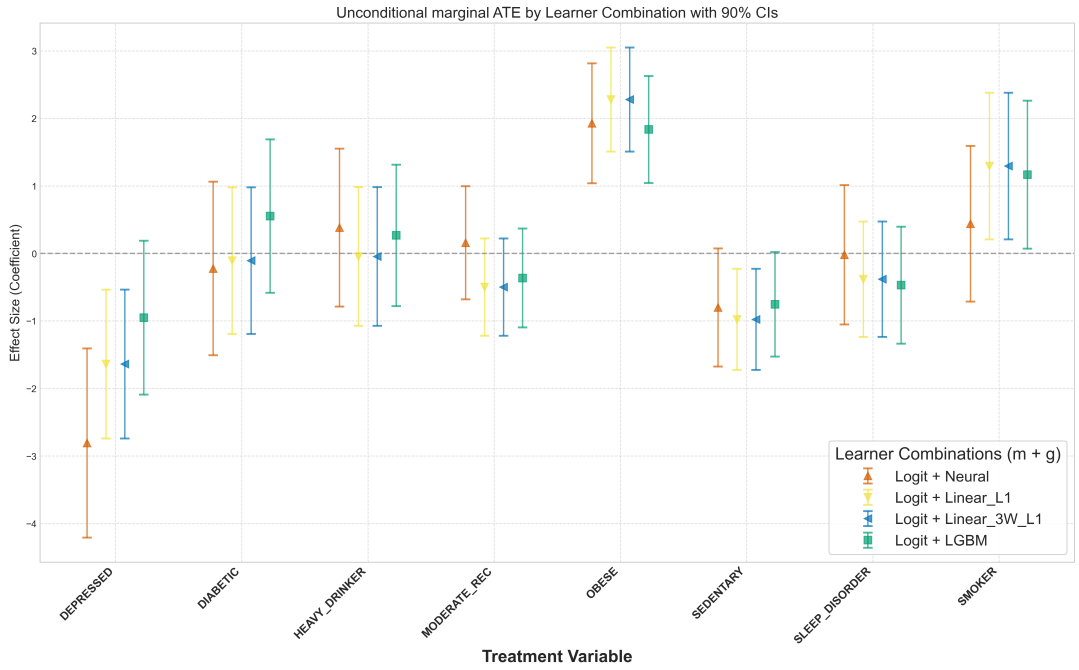

Figure 2.9: Estimated unconditional marginal average treatment effects of lifestyle factors on systolic blood pressure. Error bars represent 90% confidence intervals. Obesity shows the strongest effect, followed by diabetes and physical inactivity.

Practical Implementation Considerations

The application to NHANES data illustrates several practical considerations for implementing our proposed estimator in real-world settings. First, the choice of hyperparameters K (number of cross-fitting folds) and R (number of simulated subsets) requires careful consideration. Based on our experience, we recommend $K \geq 5$ to ensure adequate sample splitting and $R \geq \max(100, N^{1/2})$ to control simulation error. The condition $\frac{\tilde{\Sigma}_j}{KR} \ll \frac{\Sigma_j}{N}$ ensures that simulation variance does not dominate sampling variance. The parallelization across observations and simulation draws offers substantial computational benefits. The independence of subset simulations for different observations enables efficient parallel processing, making the method feasible for moderate to large sample sizes. For the NHANES analysis with 9,737 observations, 20 covariates, 8 treatments, and R=100 simulation subsets, computation time ranged from about 20 minutes for the linear models to approximately 1.8 hours for the neural network on the RRZ high-performance cluster. The neural model was run on a single node equipped with an AMD EPYC 9654 processor using 2 allocated 8-core CPUs.

2.6 Conclusion

This paper makes several methodological and theoretical contributions to the causal inference literature on multiple treatments. We develop a general combinatorial framework for treatment attribution that accommodates arbitrary treatment combinations without restrictive functional-form assumptions or prior knowledge of the causal relationships among treatments. Based on debiased machine learning, the approach offers a principled way

to decompose treatment effects in settings where treatments may interact in non-additive ways.

We further introduce a simulation-based computational strategy using the Method of Simulated Scores, which addresses the intractability of evaluating high-dimensional integrals over the treatment subset space while preserving theoretical guarantees. Building on this framework, we establish formal asymptotic properties for our estimators, including consistency and asymptotic normality, through Neyman-orthogonal scores that remain robust to nuisance estimation. Cross-fitting and simulation methods are combined to ensure valid inference even when flexible machine learning tools are used to estimate nuisance functions.

In an extensive simulation study involving challenging data-generating processes, the proposed T-DML estimator attains the lowest root mean squared error in most scenarios, along with competitive coverage probabilities and shorter confidence intervals. Our empirical application to NHANES data complements these findings, revealing meaningful patterns in the determinants of blood pressure, with obesity exhibiting the strongest marginal effect, followed by smoking, and a negative impact of depression.

2.7 Appendix

2.7.1 Proofs of the Theorems

Proof of Proposition 3. Let $\Delta(X) = \mu(A,X) - \mu_0(A,X)$ be the perturbation in the outcome model. By (2.20) and (2.22), the derivative of the expected score with respect to μ is:

$$\partial_\mu E[\psi_j] = E_X\left[\sum_{a\in\mathcal{A}_{-j}} \pi(a|X)(\Delta(a+e_j,X) - \Delta(a,X))\right] - E_P[\alpha_0(A,X)\Delta(A,X)] \tag{2.42}$$

$$= E_X[\Gamma(X;\Delta)] - E_P\left[\frac{\pi(A_{-j}|X)}{P(A|X)}(2A_j-1)\Delta(A,X)\right]. \tag{2.43}$$

We expand the second term (the observational expectation) by summing over $A_j \in \{0,1\}$:

$$E_P[\dots] = E_X\Bigg[\sum_{a\in\mathcal{A}_{-j}}\Bigg(P(a+e_j|X)\frac{\pi(a|X)}{P(a+e_j|X)}(+1)\Delta(a+e_j,X) + P(a|X)\frac{\pi(a|X)}{P(a|X)}(-1)\Delta(a,X)\Bigg)\Bigg] \tag{2.44}$$

$$= E_X\left[\sum_{a\in\mathcal{A}_{-j}} \pi(a|X)(\Delta(a+e_j,X) - \Delta(a,X))\right]. \tag{2.45}$$

This exactly matches the first term $E_X[\Gamma(X;\Delta)]$. Thus, the two terms cancel out, and the derivative is zero. A similar argument holds for perturbations in P. □

Proof of theorem 1. Let $\hat{\theta}_{j,N}$ be the estimator solving the empirical score equation

$$\mathsf{E}_N[\psi_j(W;\hat{\theta},\hat{\eta})] = 0,$$

where $\hat{\eta} = (\hat{\mu},\hat{P})$ are the machine learning nuisance estimators. Under the Assumption 6 that ψ_j is linear in θ, we expand the empirical score around the true parameter $\theta_j(\pi_{T,\delta})$ and true nuisance η_0:

$$\sqrt{N}(\hat{\theta}_{j,N} - \theta_j(\pi_{T,\delta})) = \underbrace{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_j(W_i;\theta_j(\pi_{T,\delta}),\eta_0)}_{\text{(I) oracle term}} + \underbrace{\sqrt{N}\mathsf{E}_N[\psi_j(W;\theta_j(\pi_{T,\delta}),\hat{\eta}) - \psi_j(W;\theta_j(\pi_{T,\delta}),\eta_0)]}_{\text{(II) nuisance error term}} + o_p(1).$$

We further decompose term (II) using the properties of cross-fitting and Neyman orthogo-

nality:

$$\text{term (II)} = \sqrt{N} \int \psi_j(w; \theta_j(\pi_{T,\delta}), \hat{\eta}) d(P - P_0)(w) + \sqrt{N} \left(\mathsf{E}_P[\psi_j(W; \theta_j(\pi_{T,\delta}), \hat{\eta})] - \mathsf{E}_P[\psi_j(W; \theta_j(\pi_{T,\delta}), \eta_0)] \right).$$

The first sub-term is an empirical process term which vanishes if the complexity of the function space (captured by the critical radius m_N) satisfies $\sqrt{N} m_N^2 \to 0$ in Assumption 4. The second sub-term represents the bias. Using a Taylor expansion (Gateaux derivative) in the direction $\hat{\eta} - \eta_0$:

$$\mathsf{E}_P[\psi_j(W; \theta_j(\pi_{T,\delta}), \hat{\eta})] = \partial_\eta \mathsf{E}_P[\psi_j(W; \theta_j(\pi_{T,\delta}), \eta_0)][\hat{\eta} - \eta_0] + O_P(\|\hat{\eta} - \eta_0\|^2).$$

By Proposition 3, the first-order derivative is zero. Thus, the bias is dominated by the second-order product of errors:

$$\text{Bias} \approx \sqrt{N} \mathsf{E}_P[(\hat{\alpha} - \alpha_0)(\mu_0 - \hat{\mu})].$$

Under the Mixed Bias rate condition in Assumption 3, this term is $o_p(1)$.

Indeed, the leading term converges to $\mathcal{N}(0, \Sigma)$ by the CLT. The variance $\Sigma_{j,T,\delta} = \text{Var}(\psi_j(W; \theta_j(\pi_{T,\delta}), \eta_0))$ is computed via the law of total variance:

$$\Sigma_{j,T,\delta} = \mathsf{E}[\text{Var}(\psi_j \mid X)] + \text{Var}(\mathsf{E}[\psi_j \mid X]).$$

1. **Conditional Variance:** Given $\psi_j = \Gamma(X) + \alpha(Y - \mu) - \theta$, the only random term given X is $\alpha(Y - \mu)$. Thus, $\text{Var}(\psi_j \mid X) = \mathsf{E}[\alpha_0^2 (Y - \mu_0(A, X))^2 \mid X]$. Substituting the signed density ratio:

$$\mathsf{E}[\alpha_0^2 \text{Var}(Y \mid A, X) \mid X] = \sum_{a \in \mathcal{A}} P(a|X) \left(\frac{\pi(a_{-j}|X)}{P(a|X)} \right)^2 \text{Var}(Y \mid A = a, X). \tag{2.46}$$

2. **Variance of Conditional Expectation:** $\mathsf{E}[\psi_j \mid X] = \Gamma(X; \mu_0) - \theta_j(\pi_{T,\delta})$, where $\Gamma(X; \mu_0)$ is the conditional marginal effect $\theta_j(X; \pi)$.

Summing these terms yields:

$$\Sigma_{j,T,\delta} = \mathsf{E}_P \left[\left(\frac{\pi(A_{-j}|X)}{P(A|X)} \right)^2 \text{Var}(Y \mid A, X) \right] + \text{Var}_P(\theta_j(X; \pi)). \tag{2.47}$$

Since $E[\alpha_0^2] < \infty$ by the square-integrability assumption by Assumption 5, the variance is finite. □

Proof of Theoerem 2. Let $\pi_{T,\delta}$ denote the truncated tempered policy and a^* be the target atomic configuration (the mode of $P(a|X)$). We aim to bound $|\mathbb{E}_{\pi_{T,\delta}}[\mu(A)] - \mu(a^*)|$.

We decompose the difference over the feasible support $\mathcal{A}_{feas} = \{a : P(a|X) \geq \delta\}$ and the infeasible region. Let π_T be the untruncated tempered policy. Using Triangle

inequality, we show

$$\begin{aligned}\text{Bias} &= \left|\sum_{a\in\mathcal{A}} \pi_{T,\delta}(a)\mu(a) - \mu(a^*)\right| \\ &\leq \underbrace{\sum_{a\in\mathcal{A}_{feas}} \pi_{T,\delta}(a)|\mu(a) - \mu(a^*)|}_{\text{Term I: Smoothing Error}} + \underbrace{\left|\sum_{a\in\mathcal{A}} (\pi_{T,\delta}(a) - \mathbb{I}(a = a^*))\mu(a)\right|}_{\text{Term II: Mass Shift}}.\end{aligned}$$

Using the Lipschitz Assumption 7:

$$|\mu(a) - \mu(a^*)| \leq L_{max} \cdot d_H(a, a^*).$$

Therefore, the smoothing error is bounded by the expected Hamming distance. We bound the expected distance $\mathbb{E}_{\pi_T}[d_H(A, a^*)]$ by analyzing the probability mass assigned to suboptimal configurations. Let $\rho = \frac{\max_{a\neq a^*} P(a)}{P(a^*)} < 1$ be the ratio of the second-best probability to the mode probability. For any suboptimal state $a \neq a^*$, the relative likelihood under the tempered policy is:

$$\frac{\pi_T(a)}{\pi_T(a^*)} = \left(\frac{P(a)}{P(a^*)}\right)^{1/T} \leq \rho^{1/T}.$$

Since probabilities sum to 1, we can loosely bound the total mass on all suboptimal states by summing this worst-case ratio over the 2^d possible configurations (ignoring the denominator normalization for the upper bound):

$$\pi_T(A \neq a^*) \leq |\mathcal{A}| \cdot \rho^{1/T}.$$

This can be rewritten in exponential form by defining the "probability gap" $\Delta = -\log(\rho) > 0$:

$$\pi_T(A \neq a^*) \leq |\mathcal{A}| \cdot \exp\left(-\frac{\Delta}{T}\right).$$

Combining this with the Lipschitz assumption (max distance is d):

$$\text{Bias} \leq L_{max} \cdot d \cdot |\mathcal{A}| \cdot \exp\left(-\frac{\Delta}{T}\right).$$

This confirms that the approximation error vanishes exponentially as $T \to 0$.

We bound the error introduced by truncating the policy support using Assumption 8. Let $\mu(a, x) = \mathbb{E}[Y|A = a, X = x]$ be the outcome regression function, bounded by $|\mu(a, x)| \leq B$. The truncation bias is defined as the absolute difference between the expectation under the full tempered policy π_T and the truncated policy π_δ:

$$\text{Bias}_{trunc} = \left|\mathbb{E}_{\pi_T}[\mu(A, X)] - \mathbb{E}_{\pi_\delta}[\mu(A, X)]\right|.$$

We express this expectation as an integral over the covariate space $\mathcal{X}$. Note that for the truncated policy, the weight is set to zero when $P(a|x) < \delta$. We can decompose

the integral over the feasible region $\mathcal{X}_{feas} = \{x : P(a|x) \geq \delta\}$ and the truncated region $\mathcal{X}_{trunc} = \{x : P(a|x) < \delta\}$:

$$\begin{aligned}\text{Bias}_{trunc} &= \left| \int_{\mathcal{X}} \sum_a \mu(a,x) \left(\pi_T(a|x) - \pi_\delta(a|x)\right) dP(x) \right| \\ &= \left| \int_{\mathcal{X}_{feas}} (\dots) dP(x) + \int_{\mathcal{X}_{trunc}} \sum_a \mu(a,x) \underbrace{\left(\pi_T(a|x) - 0\right)}_{\pi_\delta = 0 \text{ here}} dP(x) \right|.\end{aligned}$$

Assuming the renormalization factor in $\mathcal{X}_{feas}$ introduces negligible lower-order error compared to the complete loss of mass in $\mathcal{X}_{trunc}$, the dominant bias term comes from the truncated region.

We focus on the error from the truncated region. By bringing the absolute value inside the integral (Jensen's inequality) and applying the boundedness of the outcome ($|\mu| \leq B$):

$$\text{Bias}_{trunc} \leq \int_{\mathcal{X}_{trunc}} \sum_a |\mu(a,x)\pi_T(a|x)| \, dP(x) \tag{2.48}$$

$$\leq B \cdot \int_{\mathcal{X}_{trunc}} \underbrace{\sum_a \pi_T(a|x)}_{=1} \, dP(x). \tag{2.49}$$

The remaining integral is simply the probability measure of the truncated set. By Assumption 8, the probability mass of the region where propensity scores fall below δ decays polynomially:

$$\int_{\mathcal{X}_{trunc}} dP(x) = \mathbb{P}_X(P(A|X) < \delta) \leq c \cdot \delta^\gamma.$$

Substituting this back into the inequality, we obtain the final bound for the truncation cost:

$$\text{Bias}_{trunc} \leq B \cdot c \cdot \delta^\gamma.$$

This confirms that the bias is controlled by the tail behavior of the overlap distribution, as modeled by the parameters c and γ (following the framework of Tsybakov (2004) and Ma and Wang (2020)).

Summing the two bounds yields the result:

$$\text{Bias} \leq L_{max} d e^{-\Delta/T} + 2Bc\delta^\gamma.$$

□

Proof of Lemma 1. We aim to minimize the asymptotic order of the MSE, defined as the sum of the squared bias and the variance:

$$\text{MSE}(\delta) \asymp \text{Bias}^2(\delta) + \text{Variance}(\delta).$$

From Assumption 8 and the variance bound derived in Theorem 3:

$$\text{Bias}(\delta) = O(\delta^\gamma) \implies \text{Bias}^2(\delta) \asymp \delta^{2\gamma},$$
$$\text{Variance}(\delta) = O\left(\frac{1}{N\delta}\right).$$

To find the optimal δ, we minimize the rate function $R(\delta) = \delta^{2\gamma} + \frac{1}{N}\delta^{-1}$ with respect to δ. Taking the derivative:

$$\frac{d}{d\delta}R(\delta) = 2\gamma\delta^{2\gamma-1} - \frac{1}{N}\delta^{-2}.$$

Setting the derivative to zero to find the critical point:

$$\begin{aligned} 2\gamma\delta^{2\gamma-1} &= \frac{1}{N}\delta^{-2} \\ \delta^{2\gamma-1}\cdot\delta^2 &\propto N^{-1} \\ \delta^{2\gamma+1} &\asymp N^{-1}. \end{aligned}$$

Solving for δ, we obtain the optimal decay rate:

$$\delta_N \asymp N^{-\frac{1}{2\gamma+1}}.$$

Substituting this back into the MSE yields the minimax convergence rate of $N^{-\frac{2\gamma}{2\gamma+1}}$ (or $N^{-\frac{\gamma}{2\gamma+1}}$ for the Root-MSE). □

Proof of Theorem 3 . We aim to bound the total estimation error $|\hat{\theta}_{j,N} - \theta_j(a^*)|$. By the triangle inequality, we decompose this error into three components: the stochastic estimation error, the truncation bias, and the tempering bias.

$$|\hat{\theta}_{j,N} - \theta_j(a^*)| \leq \underbrace{|\hat{\theta}_{j,N} - \theta_j(\pi_{T_N,\delta_N})|}_{\text{(I) Stochastic Error}} + \underbrace{|\theta_j(\pi_{T_N,\delta_N}) - \theta_j(a^*)|}_{\text{(II) Approximation Bias}}. \tag{2.50}$$

We first study bounding the stochastic error via empirical process. Let $\hat{\theta}_{j,N}$ be the DML estimator solving the efficient score equation. Assuming the nuisance components are estimated consistently, the stochastic error is dominated by the empirical process of the efficient influence function ψ_π:

$$\hat{\theta}_{j,N} - \theta_j(\pi_{T_N,\delta_N}) = \frac{1}{N}\sum_{i=1}^{N}\left(\psi_\pi(Z_i) - \mathbb{E}[\psi_\pi(Z)]\right) + o_p(N^{-1/2})$$

where the influence function depends on the Regularized Riesz Representer: $\psi_\pi(Z) = \alpha_{T,\delta}(A,X)(Y-\mu) + \mu - \theta$. To obtain a tight bound, we apply Bernstein's inequality, which requires bounding both the supremum and the variance of ψ_π.

We invoke Proposition 2, which establishes the deterministic bounds of the weights. By Proposition 2, the Riesz representer is uniformly bounded by the inverse truncation threshold:

$$\sup_{a,x}|\alpha_{T_N,\delta_N}(a,x)| \leq \frac{1}{\delta_N}.$$

Assuming bounded outcomes $|Y| \leq B$, the influence function is uniformly bounded by a sequence M_N:

$$M_N := \sup_z |\psi_\pi(z)| \leq \frac{C}{\delta_N}.$$

The variance of the influence function is determined by the second moment of the Riesz representer. While the supremum scales with $1/\delta_N$, the expected variance scales more favorably under Assumption 8.

$$\sigma_N^2 := \text{Var}(\psi_\pi) \asymp \mathbb{E}\left[\alpha_{T_N,\delta_N}(A,X)^2\right] \asymp \mathbb{E}\left[\frac{1}{P(A|X)}\mathbb{I}(P \geq \delta_N)\right].$$

Integrating the inverse propensity score against the density $P(A|X)$ over the feasible region implies that the variance scales as the inverse overlap:

$$\sigma_N^2 \leq \frac{C'}{\delta_N}.$$

For the sum $S_N = \sum \psi_\pi(Z_i)$, Bernstein's concentration inequality for bounded variables states that for any $t > 0$:

$$P\left(\left|\frac{1}{N}S_N\right| > t\right) \leq 2\exp\left(-\frac{Nt^2}{2\sigma_N^2 + \frac{2}{3}M_N t}\right).$$

We seek the tightest rate t_N such that this probability remains bounded. The dominating term in the denominator is the variance $\sigma_N^2 \approx \delta_N^{-1}$. Setting the exponent to order $O(1)$ requires $Nt^2 \approx \sigma_N^2$:

$$t_N \asymp \sqrt{\frac{\sigma_N^2}{N}} \asymp \sqrt{\frac{1}{N\delta_N}}.$$

Checking the higher-order term with $M_N \approx \delta_N^{-1}$:

$$\frac{M_N t_N}{\sigma_N^2} \asymp \frac{\delta_N^{-1}(N\delta_N)^{-1/2}}{\delta_N^{-1}} = \frac{1}{\sqrt{N\delta_N}}.$$

Provided $N\delta_N \to \infty$, this term vanishes, confirming that the variance term dominates the tail behavior. Thus, we obtain the tight stochastic error rate:

$$|\hat{\theta}_{j,N} - \theta_j(\pi_{T_N,\delta_N})| = O_p\left(\frac{1}{\sqrt{N\delta_N}}\right).$$

We utilize the bound derived in Theorem 2. The approximation bias consists of the smoothing cost (dependent on T_N) and the truncation cost (dependent on δ_N).

$$\text{Bias} \leq C_1 \exp\left(-\frac{\Delta}{T_N}\right) + C_2\delta_N^\gamma.$$

We choose the decay rate of the temperature T_N such that the exponential term is negligible compared to the polynomial truncation term (e.g., $T_N \propto 1/\log N$). Thus, the dominant

bias term is:

$$\text{Bias} = O(\delta_N^\gamma)$$

Combining the stochastic and bias terms, the total error rate is:

$$\text{Error} \asymp \frac{1}{\sqrt{N\delta_N}} + \delta_N^\gamma.$$

To minimize the convergence rate, we equate the order of the squared variance and the squared bias (balancing the trade-off):

$$\begin{aligned}\frac{1}{N\delta_N} &\asymp \delta_N^{2\gamma} \\ \delta_N^{2\gamma+1} &\asymp N^{-1} \\ \delta_N^* &\asymp N^{-\frac{1}{2\gamma+1}}.\end{aligned}$$

Substituting this optimal δ_N back into the error expression:

$$\text{Rate} \asymp (\delta_N^*)^\gamma = \left(N^{-\frac{1}{2\gamma+1}}\right)^\gamma = N^{-\frac{\gamma}{2\gamma+1}}.$$

This confirms that the estimator converges to $\theta(a^*)$ in probability at the stated minimax rate. □

Proof of Theorem 4. We analyze the asymptotic distribution of the centered and scaled estimator for the j-th marginal effect. Let the total approximation bias be denoted by $\text{Bias}_N = \theta_j(\pi_{T_N,\delta_N}) - \theta_j(a^*)$. We decompose the error into the stochastic component and the deterministic bias:

$$\hat{\theta}_{j,N} - \theta_j(a^*) = (\hat{\theta}_{j,N} - \theta_j(\pi_{T_N,\delta_N})) + \text{Bias}_N.$$

Rearranging terms to isolate the stochastic fluctuation:

$$\hat{\theta}_{j,N} - \theta_j(a^*) - \text{Bias}_N = \hat{\theta}_{j,N} - \theta_j(\pi_{T_N,\delta_N}).$$

Multiplying by the scaling factor $\sqrt{N\delta_N}$:

$$Z_N := \sqrt{N\delta_N}\left(\hat{\theta}_{j,N} - \theta_j(\pi_{T_N,\delta_N})\right).$$

Step 1: Linearization via Efficient Influence Function. Under the DML assumptions (consistent nuisance estimation with product rates $o_p(N^{-1/2})$), the estimator is asymptotically equivalent to the sample average of the efficient influence function $\psi_{j,N}$ evaluated at the true nuisance parameters but with tuning parameters (T_N, δ_N):

$$\hat{\theta}_{j,N} - \theta_j(\pi_{T_N,\delta_N}) = \frac{1}{N}\sum_{i=1}^{N} \bar{\psi}_{j,N}(W_i) + o_p\left(\frac{1}{\sqrt{N\delta_N}}\right),$$

where $\bar{\psi}_{j,N}(W) = \psi_j(W; \eta_0, T_N, \delta_N) - \mathbb{E}[\psi_j]$ and $o_p\left(\frac{1}{\sqrt{N\delta_N}}\right)$ is from (2.33). Substituting this into the expression for Z_N:

$$Z_N = \sum_{i=1}^{N} \frac{\sqrt{\delta_N}}{\sqrt{N}} \bar{\psi}_{j,N}(W_i) + o_p(1).$$

This defines a triangular array of independent random variables $Y_{i,N} = \frac{\sqrt{\delta_N}}{\sqrt{N}} \bar{\psi}_{j,N}(W_i)$ with mean zero.

Step 2: Variance Stabilization. We examine the variance of the terms in the sum. The variance of the influence function is dominated by the squared Riesz weight $\alpha_{j,N}^2$. Using Assumption 8, the second moment scales as:

$$\text{Var}(\bar{\psi}_{j,N}) \asymp \mathbb{E}[\alpha_{j,N}^2] \asymp \mathbb{E}\left[\frac{1}{P(A|X)} \mathbb{I}(P \geq \delta_N)\right] \asymp \frac{1}{\delta_N}.$$

The variance of the sum Z_N is:

$$\text{Var}(Z_N) = \sum_{i=1}^{N} \text{Var}(Y_{i,N}) = N \cdot \frac{\delta_N}{N} \text{Var}(\bar{\psi}_{j,N}) = \delta_N \cdot \text{Var}(\bar{\psi}_{j,N}).$$

As $N \to \infty$, since $\text{Var}(\bar{\psi}_{j,N}) \asymp \delta_N^{-1}$, the product converges to a strictly positive constant:

$$\lim_{N \to \infty} \text{Var}(Z_N) = \sigma_\infty^2 > 0.$$

Step 3: Lyapunov Central Limit Theorem. To prove asymptotic normality for the triangular array (not identically distributed due to change the tuning parameters that leads to change of variance of regularised Riesz weights), we verify the Lyapunov condition using the third absolute moment. We must demonstrate that the Lyapunov ratio vanishes as $N \to \infty$:

$$\lim_{N \to \infty} \frac{\sum_{i=1}^{N} \mathbb{E}[|Y_{i,N}|^3]}{(\text{Var}(Z_N))^{3/2}} = 0.$$

First, we bound the third moment. Since the Riesz weights are strictly bounded by $1/\delta_N$ (Proposition 2), the influence function satisfies $|\bar{\psi}_{j,N}| \leq C/\delta_N$. The third moment of the unscaled influence function behaves as:

$$\mathbb{E}[|\bar{\psi}_{j,N}|^3] \leq \sup |\bar{\psi}_{j,N}| \cdot \mathbb{E}[|\bar{\psi}_{j,N}|^2] \asymp \frac{1}{\delta_N} \cdot \frac{1}{\delta_N} = \frac{1}{\delta_N^2}.$$

Now, substituting this into the ratio for the scaled variable $Y_{i,N} = \frac{\sqrt{\delta_N}}{\sqrt{N}} \bar{\psi}_{j,N}$:

$$\sum_{i=1}^{N} \mathbb{E}[|Y_{i,N}|^3] = N \cdot \left(\frac{\sqrt{\delta_N}}{\sqrt{N}}\right)^3 \cdot \mathbb{E}[|\bar{\psi}_{j,N}|^3] \asymp N \cdot \frac{\delta_N^{1.5}}{N^{1.5}} \cdot \frac{1}{\delta_N^2} = \frac{1}{\sqrt{N\delta_N}}.$$

Since the denominator $(\text{Var}(Z_N))^{3/2}$ converges to a strictly positive constant $(\sigma_\infty^2)^{3/2}$, the Lyapunov ratio is dominated by the numerator:

$$\text{Ratio} \asymp \frac{1}{\sqrt{N\delta_N}}.$$

By the Optimal Tuning Rates, $N\delta_N \to \infty$, so the ratio converges to zero. Thus, the Lyapunov condition is satisfied, implying $Z_N \xrightarrow{d} \mathcal{N}(0, \sigma_\infty^2)$. □

Proof Sketch of Theorem 5. Recall that the simulated estimator solves the linear equation $\frac{1}{N}\sum_{i=1}^N \tilde{\psi}_i(\hat{\theta}) = 0$. Since $\tilde{\psi}_i$ is affine in θ (i.e., $\tilde{\psi}_i(\theta) = \tilde{\psi}_i^* - \theta$), we can write the estimator as a sample average of the uncentered simulated scores: $\tilde{\theta}_{j,N} = \frac{1}{N}\sum_{i=1}^N \tilde{\psi}_{ij}^*$.

We decompose the total estimation error $\sqrt{N}(\tilde{\theta}_{j,N} - \theta_j(\pi_{T,\delta}))$ into three distinct components:

$$\sqrt{N}(\tilde{\theta}_{j,N} - \theta_j(\pi_{T,\delta})) = \underbrace{\frac{1}{\sqrt{N}}\sum_{i=1}^N \psi_j^*(W_i)}_{\text{(I) Oracle Sampling Process}} + \underbrace{\sqrt{N}\mathbb{E}_n[\hat{\psi} - \psi]}_{\text{(II) Nuisance Estimation Error}} + \underbrace{\frac{1}{\sqrt{N}}\sum_{i=1}^N (\tilde{\psi}_{ij}^* - \psi_{ij}^*)}_{\text{(III) Simulation Process}}. \tag{2.51}$$

Term (I) is exactly the sum of the efficient influence functions derived in Theorem 4. This term converges in distribution to $\mathcal{N}(0, \Sigma)$ by CLT. Term (II): Under the DML Assumptions (cross-fitting and product-bias condition), this term converges to zero in probability (Neyman Orthogonality). Term (III) represents the pure Monte Carlo noise. Conditional on the data W, $\tilde{\psi}_{ij}^* - \psi_{ij}^*$ are mean-zero (by the unbiased simulation condition) and independent across i. Its variance is $\frac{1}{N}\sum \text{Var}(\tilde{\psi} - \psi) = \frac{1}{M}\mathbb{E}[\text{Var}(\tilde{\Gamma}|W)]$. By a secondary CLT applied to the simulation process, this term converges to $\mathcal{N}(0, \frac{1}{M}\Sigma_{sim})$. By the independence of the sampling and simulation processes, the variances add linearly:

$$V_{total} = \Sigma + \frac{1}{M}\Sigma_{sim}.$$

□

2.7.2 Details on the Simulation Study

Data Generating Process (DGP) Details

This appendix provides a complete technical specification of the data generating processes used in the simulation study. The DGP is adapted from Belloni et al. (2017) and extended to multiple treatments.

General Structure

Let d denote the number of binary treatments, with $A = (A_1, \ldots, A_d)$ representing the treatment vector. For each observation $i = 1, \ldots, n$, the data are generated as follows.

Covariates. The p-dimensional covariate vector is drawn from a multivariate normal distribution:

$$X_i \sim \mathcal{N}(0, \Sigma), \quad \text{with} \quad \Sigma_{kj} = 0.5^{|j-k|}.$$

Treatment Assignment. Each treatment A_j ($j = 1, \ldots, d$) is assigned independently (for the independent DGPs) or with pairwise correlation (for the correlated DGPs) via a logistic model. For each treatment j, we generate a coefficient vector β_j of length p with a decaying structure. Let B be a $p \times d$ matrix with entries drawn independently from Uniform$(-1, 1)$. Define the decay vector w with $w_k = 1/k^2$ for $k = 1, \ldots, p$. Then set

$$\beta_j = B_{:,j} \circ w,$$

where $\circ$ denotes element-wise multiplication.

The linear predictor for treatment j is

$$\eta_{ij} = c_A \cdot X_i^\top \beta_j,$$

where c_A is a scaling factor that controls the strength of confounding. Define

$$\sigma_j^2 = \beta_j^\top \Sigma \beta_j,$$

and let

$$c_A = \sqrt{\frac{(\pi^2/3) R_A^2}{(1 - R_A^2) \cdot \frac{1}{d} \sum_{j=1}^{d} \sigma_j^2}}.$$

Here, R_A^2 is the desired proportion of variance in the treatment assignment explained by the covariates.

For each treatment j, generate an independent uniform variable $U_{ij} \sim$ Uniform$(0, 1)$ and set

$$A_{ij} = \mathbb{I}\left\{ \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})} > U_{ij} \right\}.$$

Inducing Pairwise Correlation (for Correlated DGPs). For the DGP variants *2-Way* and *Complex*, we induce correlation between adjacent treatment pairs (A_1, A_2), (A_3, A_4), etc., using a Gaussian copula with correlation $\rho = 0.5$. For each such pair $(2k - 1, 2k)$, we generate:

$$\begin{aligned} z_{i,2k-1} &\sim \mathcal{N}(0, 1), \\ z_{i,2k} &= \rho \, z_{i,2k-1} + \sqrt{1 - \rho^2} \, \varepsilon_{i,k}, \quad \varepsilon_{i,k} \sim \mathcal{N}(0, 1), \end{aligned}$$

and then set

$$U_{i,2k-1} = \Phi(z_{i,2k-1}), \quad U_{i,2k} = \Phi(z_{i,2k}),$$

where Φ is the standard normal cumulative distribution function. The remaining U_{ij} (if d is odd) are kept independent.

Outcome Generation. The outcome Y_i is generated as

$$Y_i = \sum_{j=1}^{d} \theta_j A_{ij} + f_{\text{int}}(A_i) + c_Y \cdot X_i^\top \beta_Y \cdot \sum_{j=1}^{d} A_{ij} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,1),$$

where θ_j are fixed baseline treatment effects (drawn once and held constant across replications), and β_Y is a coefficient vector with $\beta_{Y,k} = 1/k^2$. The scaling factor c_Y is given by

$$c_Y = \sqrt{\frac{R_Y^2}{(1-R_Y^2)\beta_Y^\top \Sigma \beta_Y}},$$

with R_Y^2 set to 0.9 in our simulations.

The interaction term $f_{\text{int}}(A_i)$ depends on the DGP variant:

- **Simple:** $f_{\text{int}}(A_i) = 0$.
- **2-Way:** $f_{\text{int}}(A_i) = \sum_{k=1}^{\lfloor d/2 \rfloor} A_{i,2k-1} A_{i,2k}$.
- **3-Way:** $f_{\text{int}}(A_i) = A_{i1}A_{i2}A_{i3} + A_{i1}A_{i4}A_{i6}$ (the second term included only when $d \geq 6$).
- **Complex:** $f_{\text{int}}(A_i) = \sum_{k=1}^{\lfloor d/2 \rfloor} A_{i,2k-1} A_{i,2k} + A_{i1}A_{i2}A_{i3} + A_{i1}A_{i4}A_{i6}$.

True Unconditional Marginal Treatment Effects. For each treatment j, the true unconditional marginal average treatment effect (ATE) in a given sample is computed by counterfactual comparison:

$$\theta_j^* = \frac{1}{n} \sum_{i=1}^{n} \Big[Y_i^{(j,1)}(A_{i,-j}) - Y_i^{(j,0)}(A_{i,-j}) \Big],$$

where $Y_i^{(j,1)}(A_{i,-j})$ and $Y_i^{(j,0)}(A_{i,-j})$ are the potential outcomes for unit i when treatment j is set to 1 or 0, respectively, while the other treatments $A_{i,-j}$ are fixed at their observed values. Because the DGP is fully known, these potential outcomes can be evaluated exactly by plugging the corresponding treatment vectors into the outcome equation. The resulting θ_j^* incorporates both the baseline coefficient θ_j and the contributions from all interaction terms that involve treatment j, averaged over the empirical distribution of the other treatments.

Summary of DGP Variants

The four DGP variants used in the simulation study are:

- **Simple:** Independent treatments, no outcome interactions.
- **2-Way:** Correlated treatments (adjacent pairs), two-way outcome interactions.
- **3-Way:** Independent treatments, three-way outcome interactions.

- **Complex:** Correlated treatments (adjacent pairs), two-way and three-way outcome interactions.

This design allows us to separately examine the impact of treatment dependence (on propensity score estimation) and outcome interaction complexity (on treatment effect estimation).

Parameter Settings

In the simulation study, we fix the following parameters unless otherwise noted:

- Number of observations: $n = 2,000$
- Covariate dimension: $p = 10$
- Number of treatments: $d = 6$
- Outcome signal strength: $R_Y^2 = 0.9$
- Propensity score signal strength: $R_A^2 = 0.5$

All simulations are repeated 100 times with different random seeds to obtain stable performance metrics.

Extended Simulation Results

Table 2.3: Results Overview (n=2,000, dim_a=6, dim_x = 10, R2_A = 0.5, R2_y=0.9)

		μ = Linear				μ = NeuralNet			
DGP	Method	RMSE	Std.dev.	Length	Cover	RMSE	Std.dev.	Length	Cover
Simple	T-DML	**0.06**	0.64	0.49	1.00	**0.11**	0.66	0.53	0.95
	T-Bench	0.09	0.64	0.35	0.88	0.28	0.73	0.27	0.29
	DML-IRM	0.35	0.80	0.36	0.31	0.13	0.66	0.65	0.95
	DML-PLR	0.55	0.01	0.2	0.08	0.53	0.05	0.28	0.15
	TMLE	0.23	0.72	1.02	0.91	0.18	0.67	0.65	0.86
2-Way	T-DML	**0.10**	0.64	0.50	0.98	**0.12**	0.67	0.54	0.93
	T-Bench	0.33	0.64	0.35	0.09	0.37	0.72	0.28	0.22
	DML-IRM	0.37	0.80	0.36	0.35	0.13	0.67	0.65	0.96
	DML-PLR	0.85	0.02	0.22	0.00	0.82	0.06	0.28	0.03
	TMLE	0.33	0.81	1.06	0.76	0.30	0.78	0.72	0.65
3-Way	T-DML	**0.07**	0.70	0.50	0.99	**0.12**	0.72	0.53	0.93
	T-Bench	0.35	0.63	0.35	0.20	0.42	0.73	0.27	0.21
	DML-IRM	0.34	0.83	0.37	0.27	0.13	0.72	0.66	0.96
	DML-PLR	0.69	0.02	0.22	0.20	0.67	0.05	0.28	0.25
	TMLE	0.24	0.78	1.04	0.91	0.19	0.73	0.67	0.85
Complex	T-DML	**0.12**	0.71	0.50	0.93	**0.13**	0.73	0.54	0.90
	T-Bench	0.66	0.64	0.35	0.02	0.60	0.72	0.28	0.09
	DML-IRM	0.34	0.84	0.37	0.36	**0.13**	0.73	0.65	0.96
	DML-PLR	1.11	0.02	0.22	0.00	1.08	0.06	0.28	0.01
	TMLE	0.37	0.87	1.08	0.72	0.36	0.85	0.74	0.58
		μ = Linear3W				μ = LGBM			
DGP	Method	RMSE	Std.dev.	Length	Cover	RMSE	Std.dev.	Length	Cover
Simple	T-DML	**0.06**	0.64	0.50	1.00	**0.14**	0.65	0.69	0.96
	T-Bench	0.10	0.65	0.35	0.84	0.27	0.64	0.18	0.20
	DML-IRM	0.34	0.79	0.37	0.33	0.32	0.76	1.34	0.88
	DML-PLR	0.55	0.02	0.22	0.09	0.48	0.14	0.38	0.25
	TMLE	0.23	0.72	1.01	0.91	0.19	0.62	0.45	0.65
2-Way	T-DML	**0.10**	0.65	0.50	0.98	**0.16**	0.65	0.70	0.88
	T-Bench	0.33	0.65	0.35	0.11	0.32	0.63	0.20	0.21
	DML-IRM	0.46	0.80	0.37	0.32	0.37	0.82	1.33	0.84
	DML-PLR	0.85	0.02	0.22	0.00	0.70	0.14	0.39	0.25
	TMLE	0.32	0.81	1.06	0.76	0.23	0.74	0.52	0.64
3-Way	T-DML	**0.08**	0.70	0.50	0.99	**0.15**	0.71	0.69	0.90
	T-Bench	0.27	0.65	0.35	0.34	0.37	0.66	0.18	0.15
	DML-IRM	0.37	0.84	0.37	0.36	0.32	0.81	1.34	0.87
	DML-PLR	0.69	0.02	0.22	0.19	0.59	0.14	0.39	0.35
	TMLE	0.24	0.78	1.03	0.91	0.20	0.67	0.46	0.63
Complex	T-DML	**0.12**	0.71	0.50	0.91	**0.19**	0.71	0.70	0.86
	T-Bench	0.59	0.65	0.35	0.01	0.48	0.65	0.20	0.11
	DML-IRM	0.51	0.85	0.37	0.27	0.38	0.88	1.33	0.83
	DML-PLR	1.11	0.02	0.22	0.00	0.92	0.15	0.39	0.14
	TMLE	0.37	0.87	1.08	0.71	0.25	0.80	0.53	0.61

RMSE by Outcome Learner

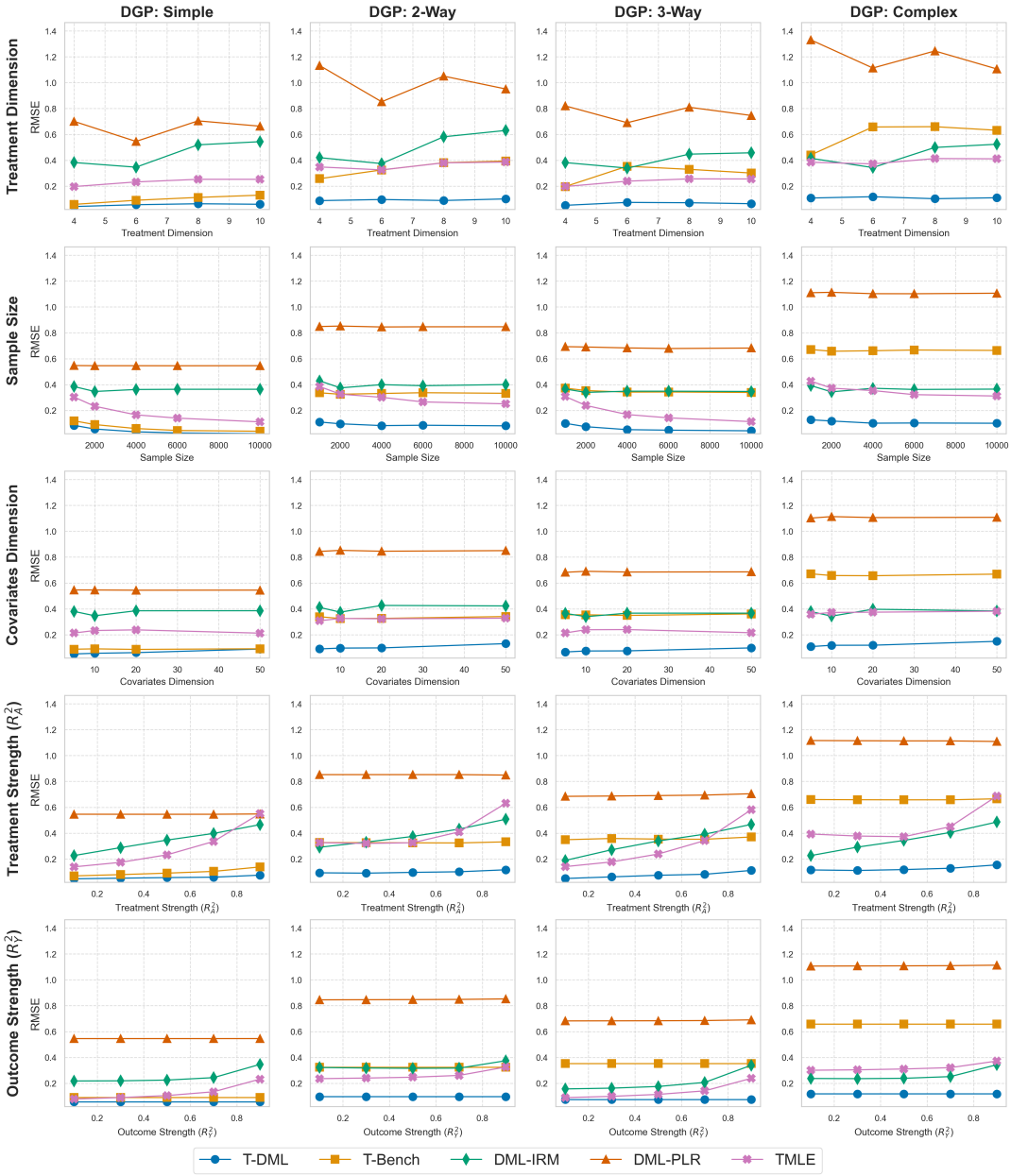

Figure 2.10: μ = Linear

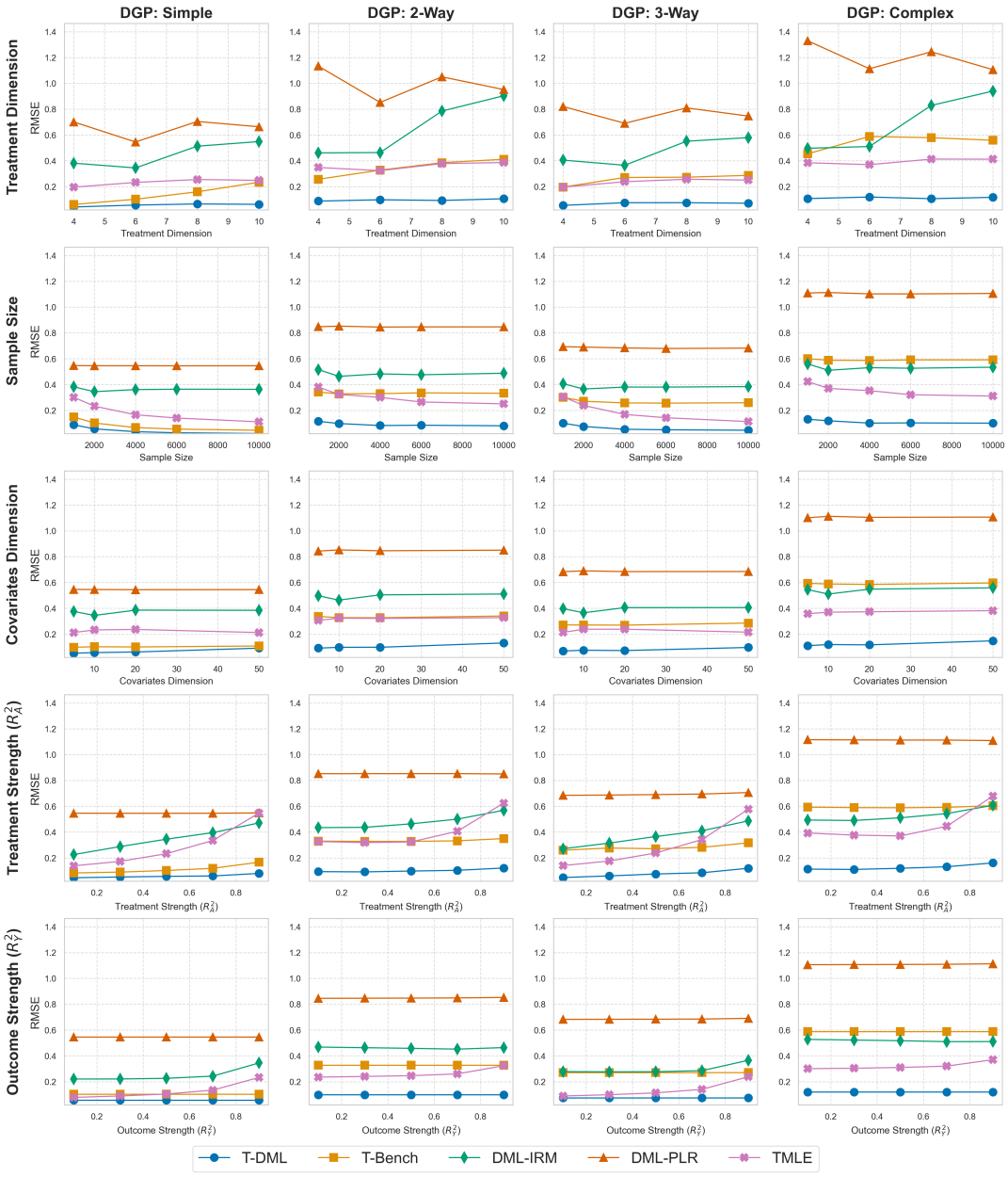

Figure 2.11: μ = Linear3W

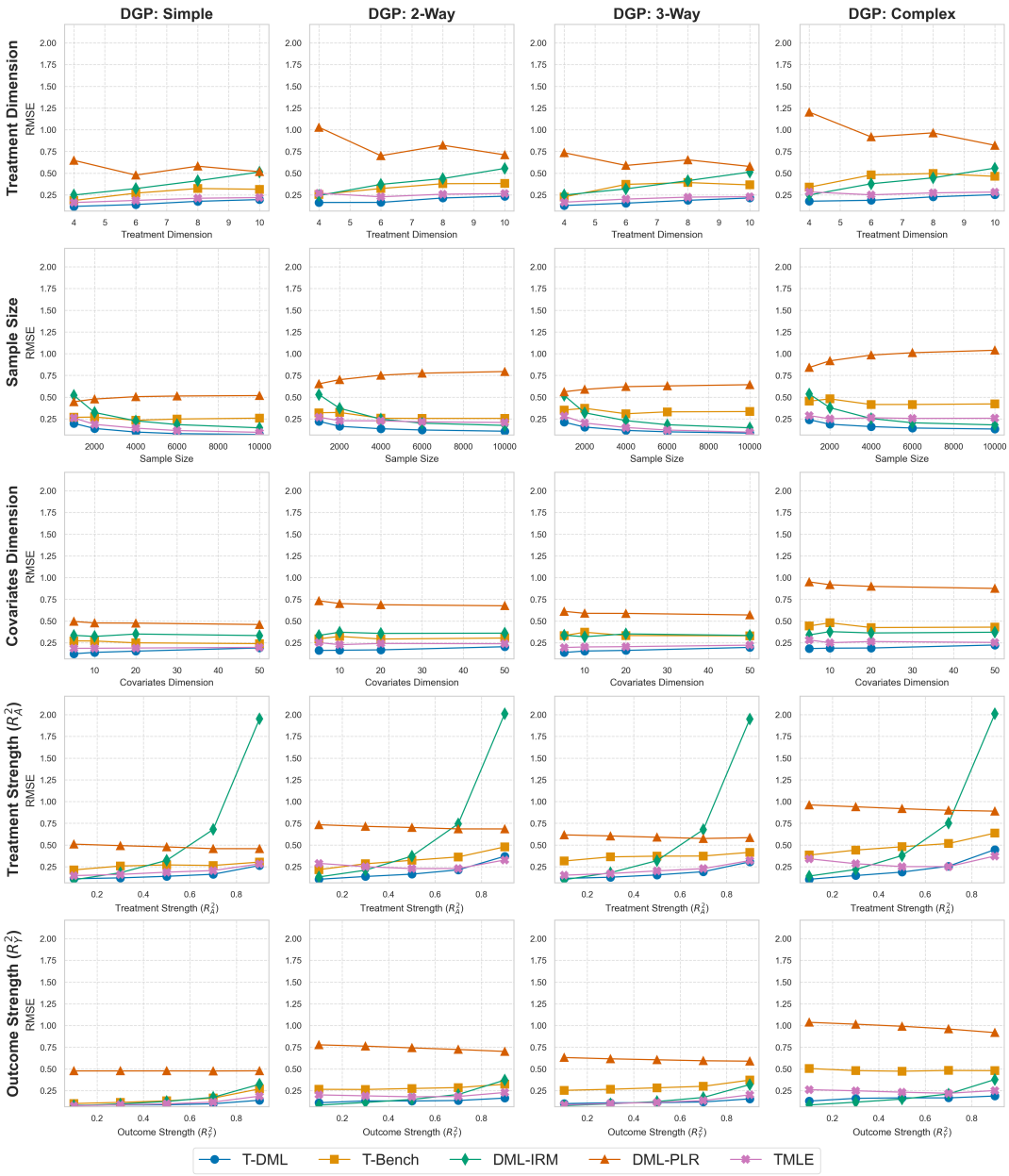

Figure 2.12: μ = LGBM

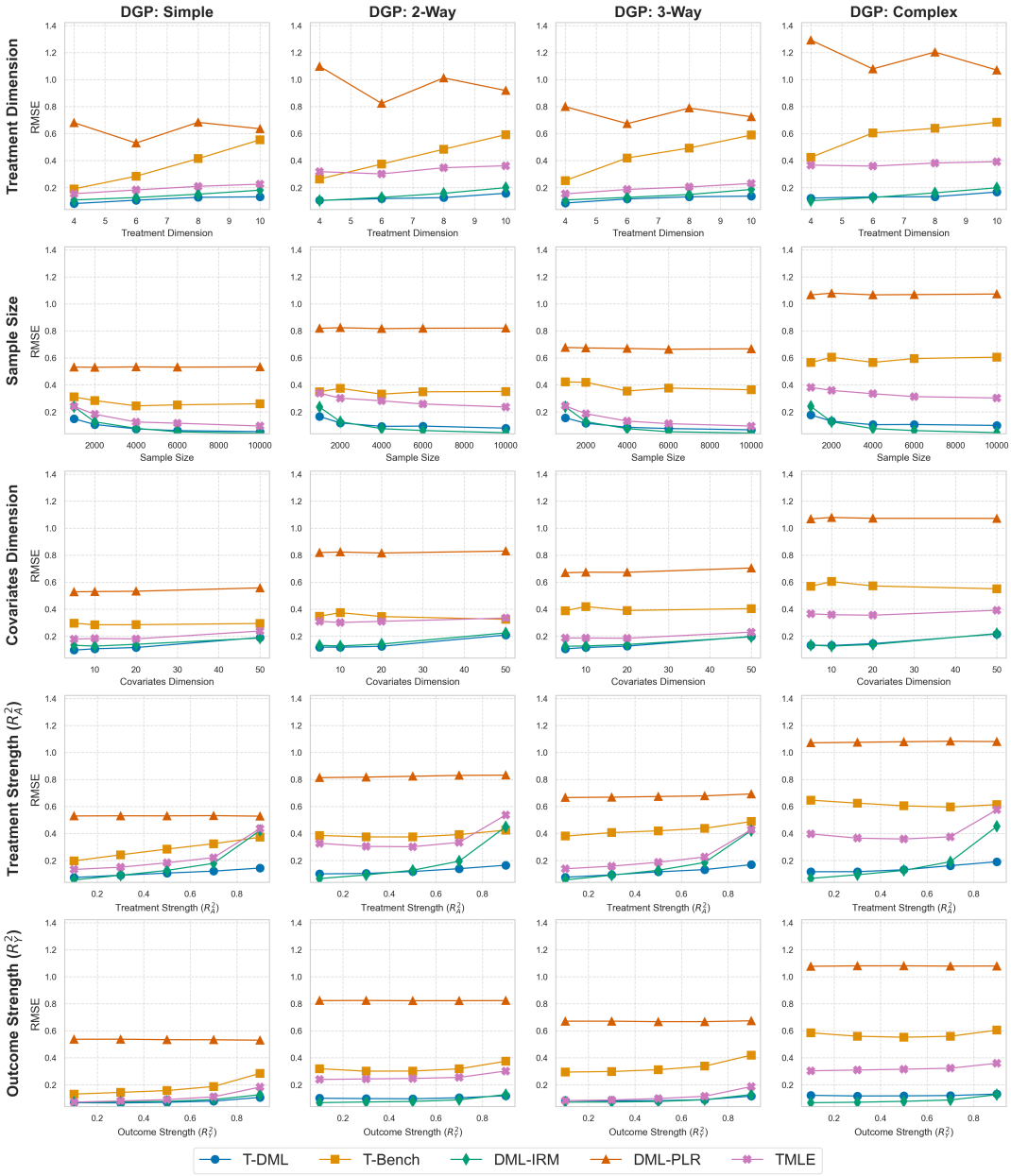

Figure 2.13: μ = NeuralNet

Coverage and CI Length by Outcome Learner

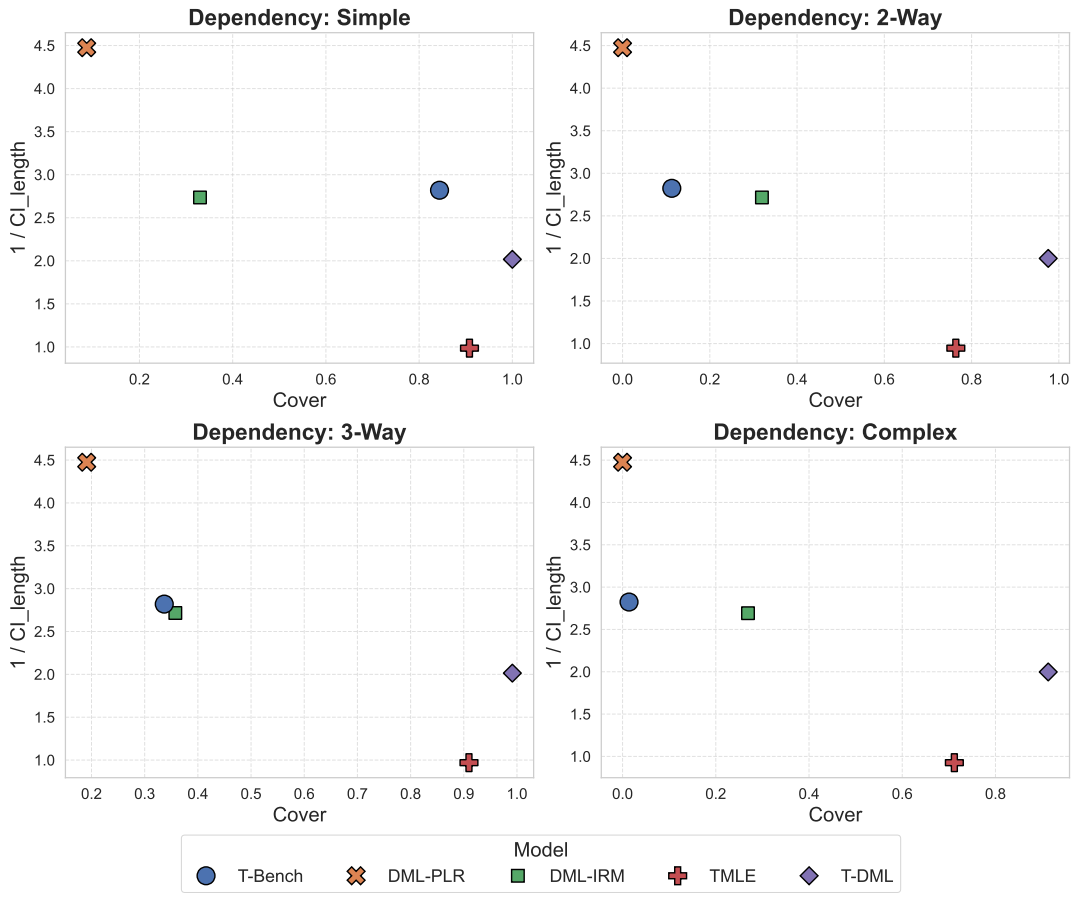

Figure 2.14: μ = Linear3W

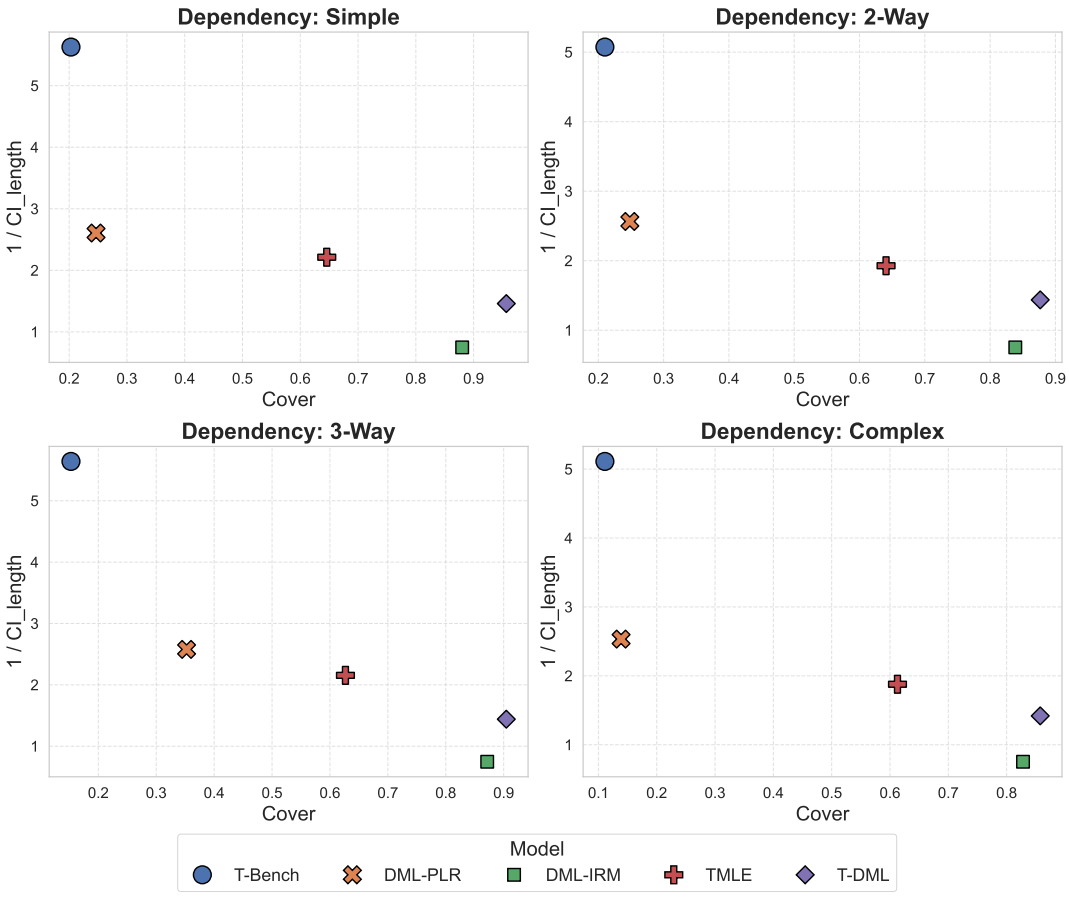

Figure 2.15: μ = LGBM

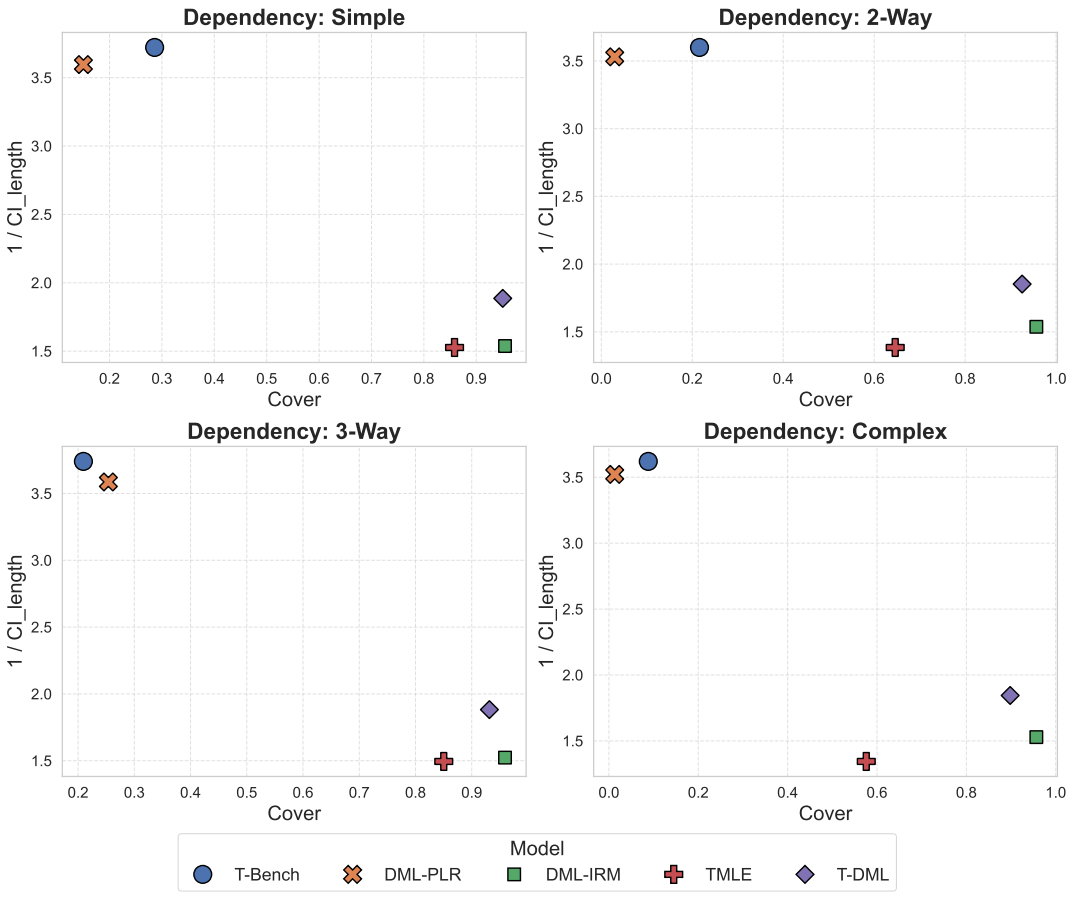

Figure 2.16: μ = NeuralNet

MAE by Outcome Learner

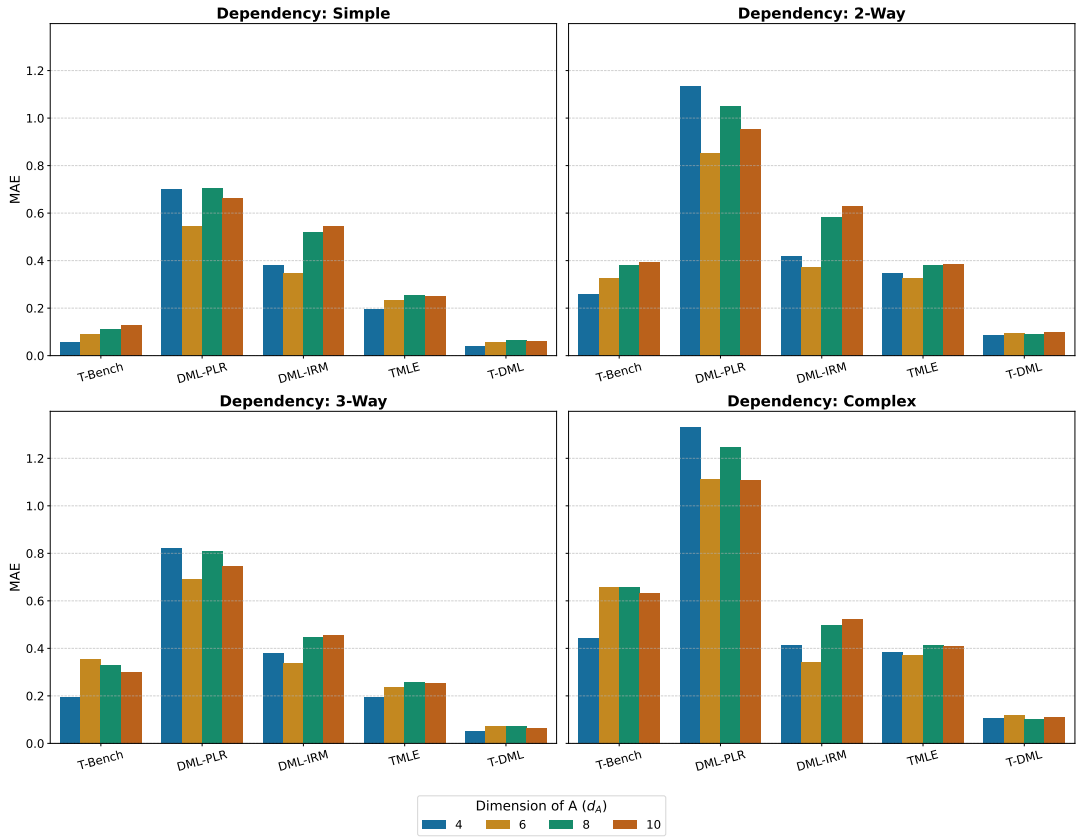

Figure 2.17: μ = Linear

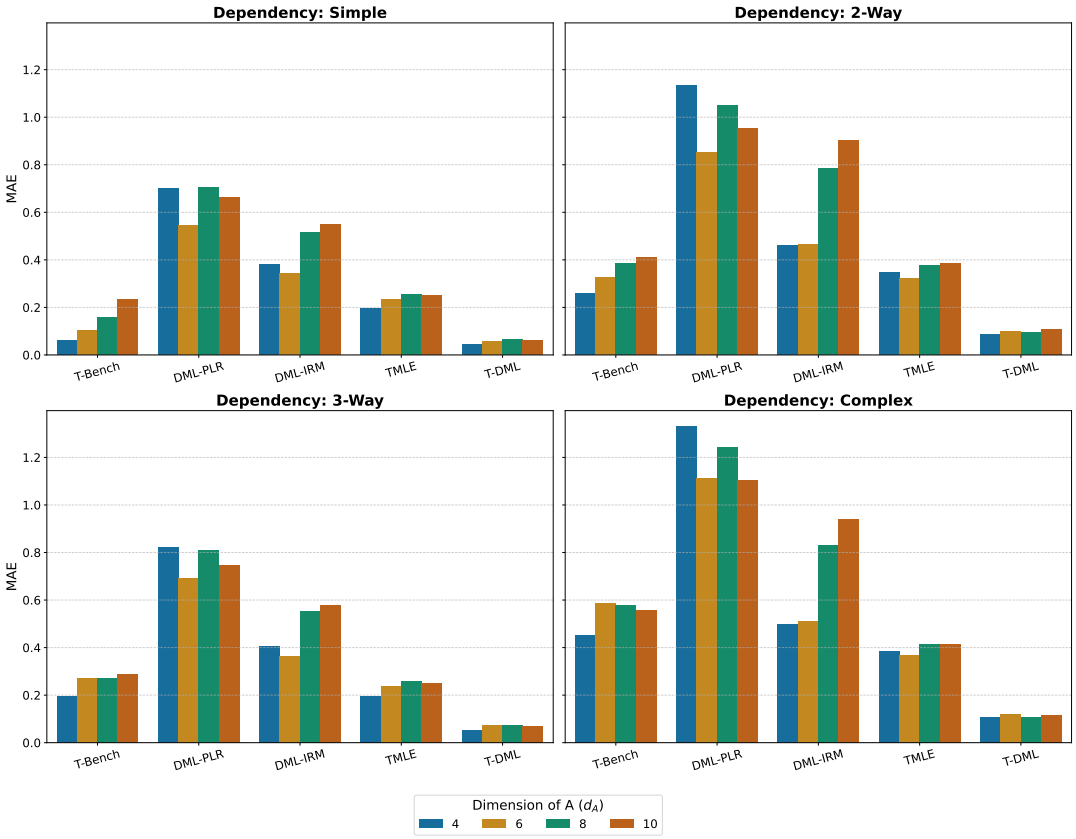

Figure 2.18: μ = Linear3W

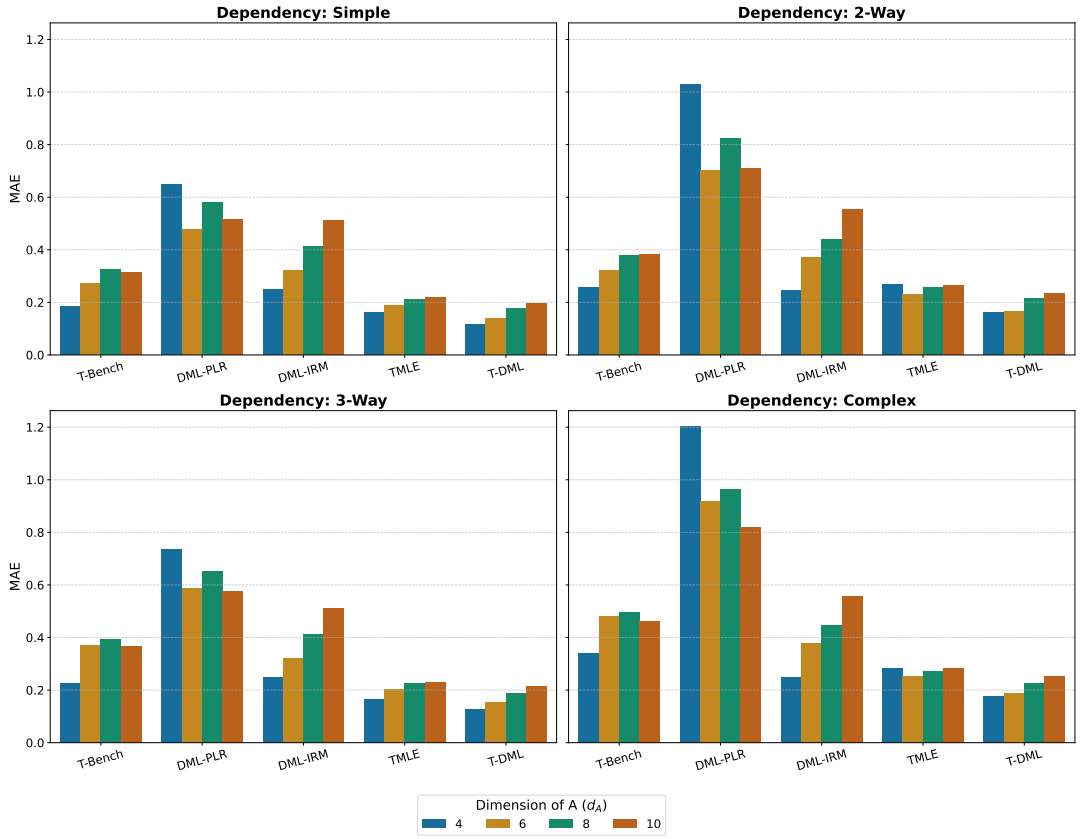

Figure 2.19: μ = LGBM

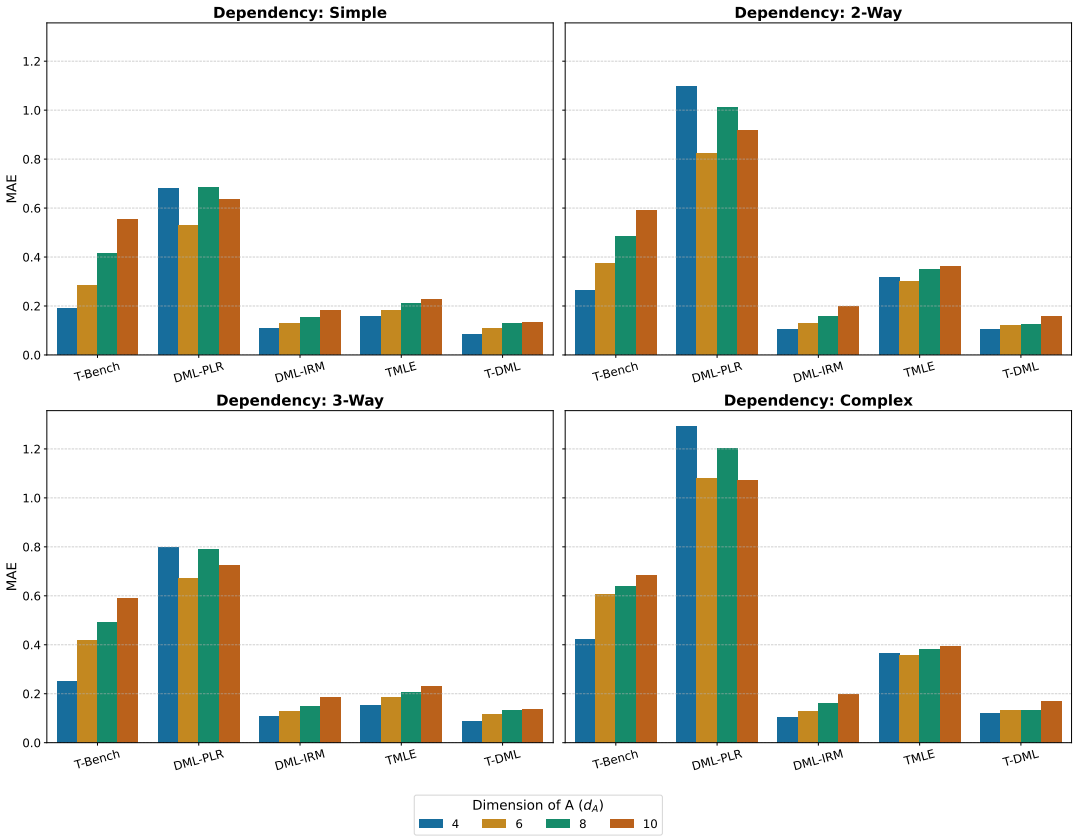

Figure 2.20: μ = NeuralNet

Propensity Diagnostics

Mean Effective Support

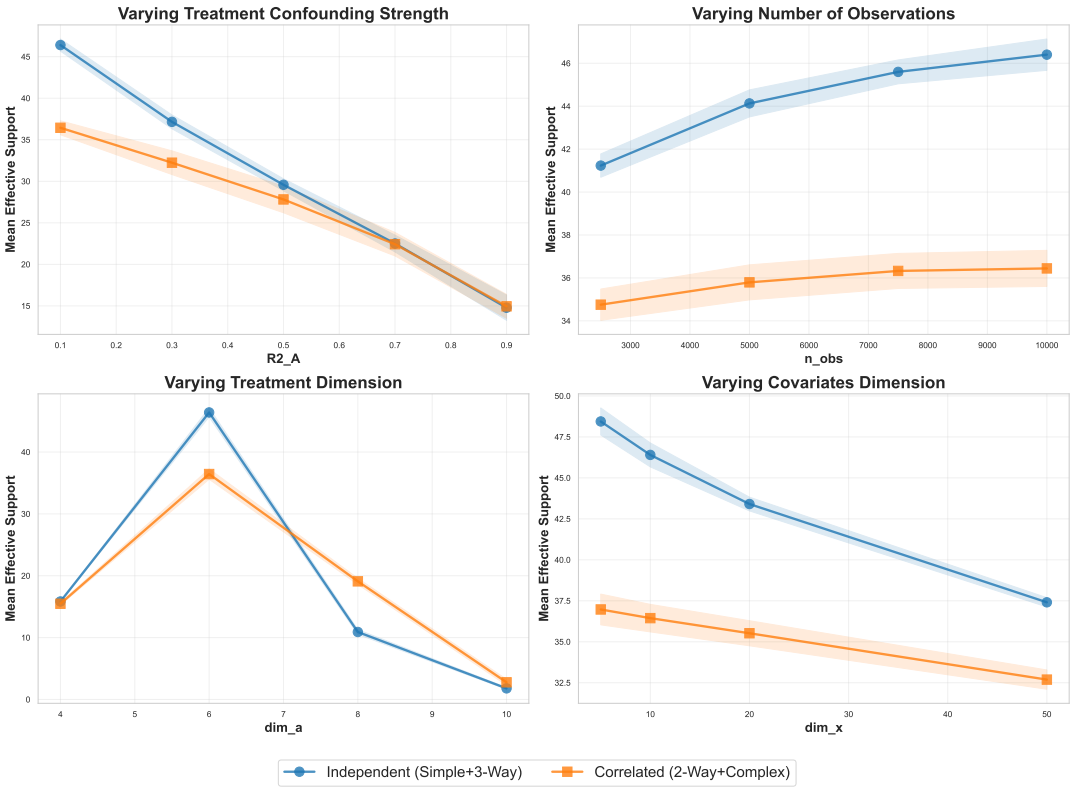

Figure 2.21: m = Logistic_L1

Detailed Propensity Metrics

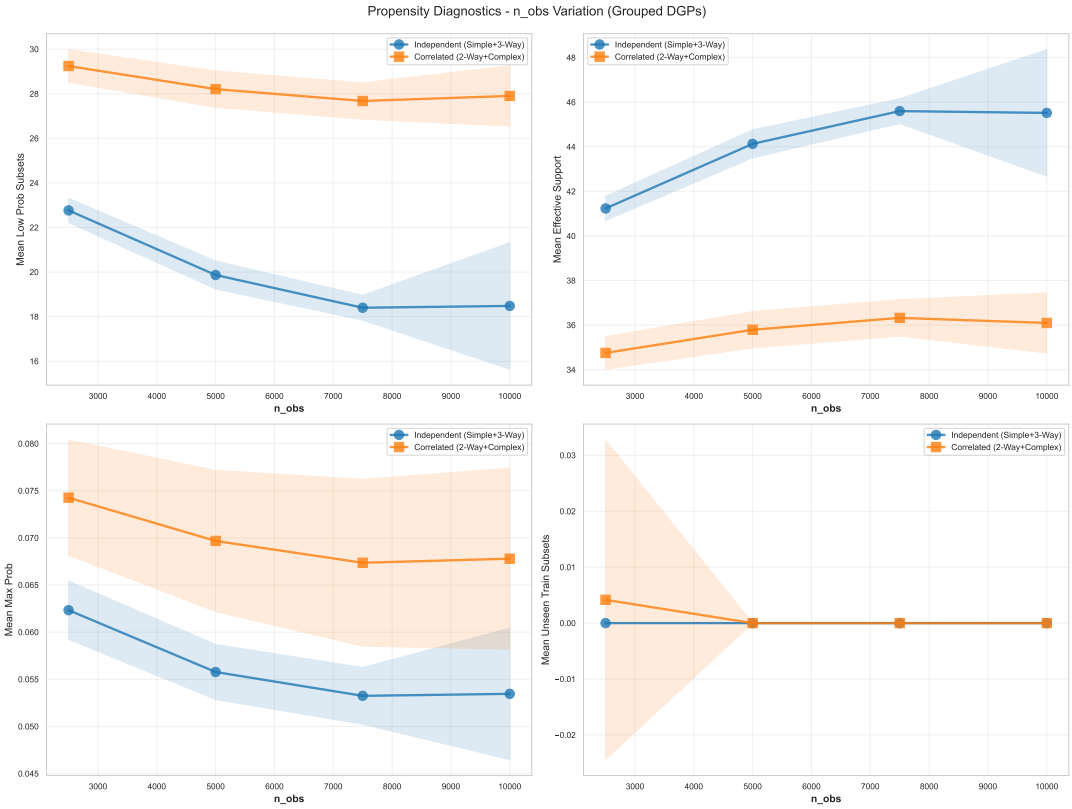

Figure 2.22: Propensity metrics by varying number of observations.

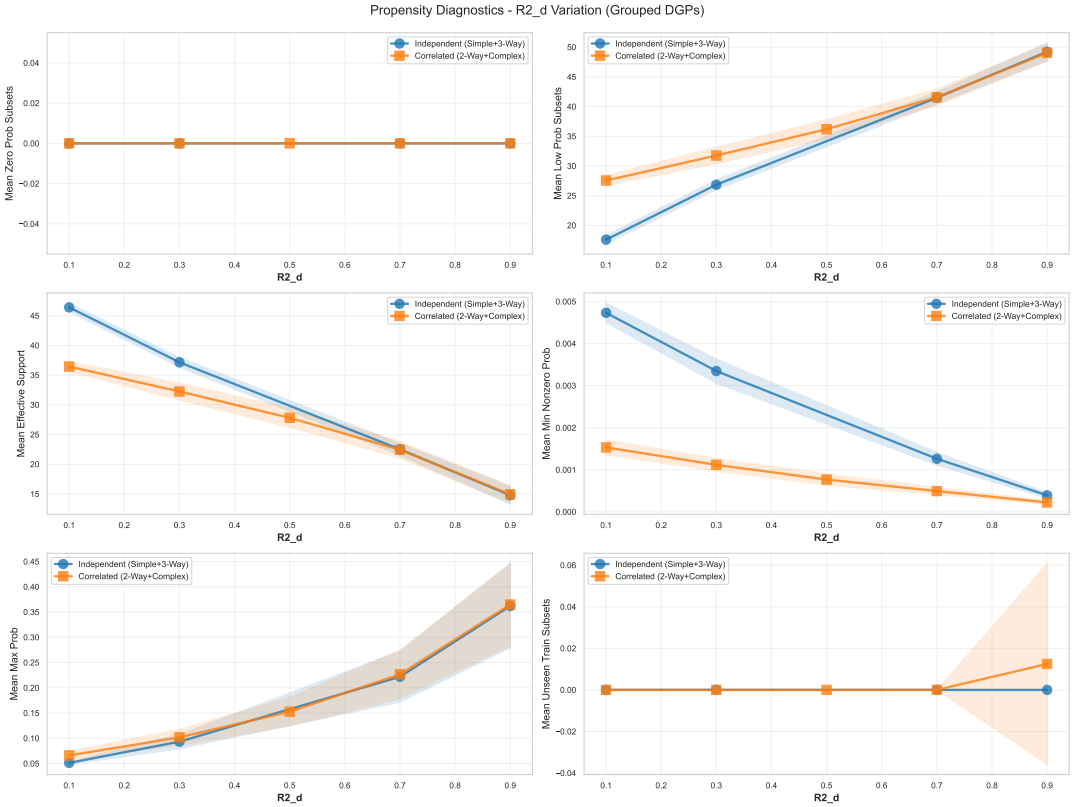

Figure 2.23: Propensity metrics by varying dimension of treatments.

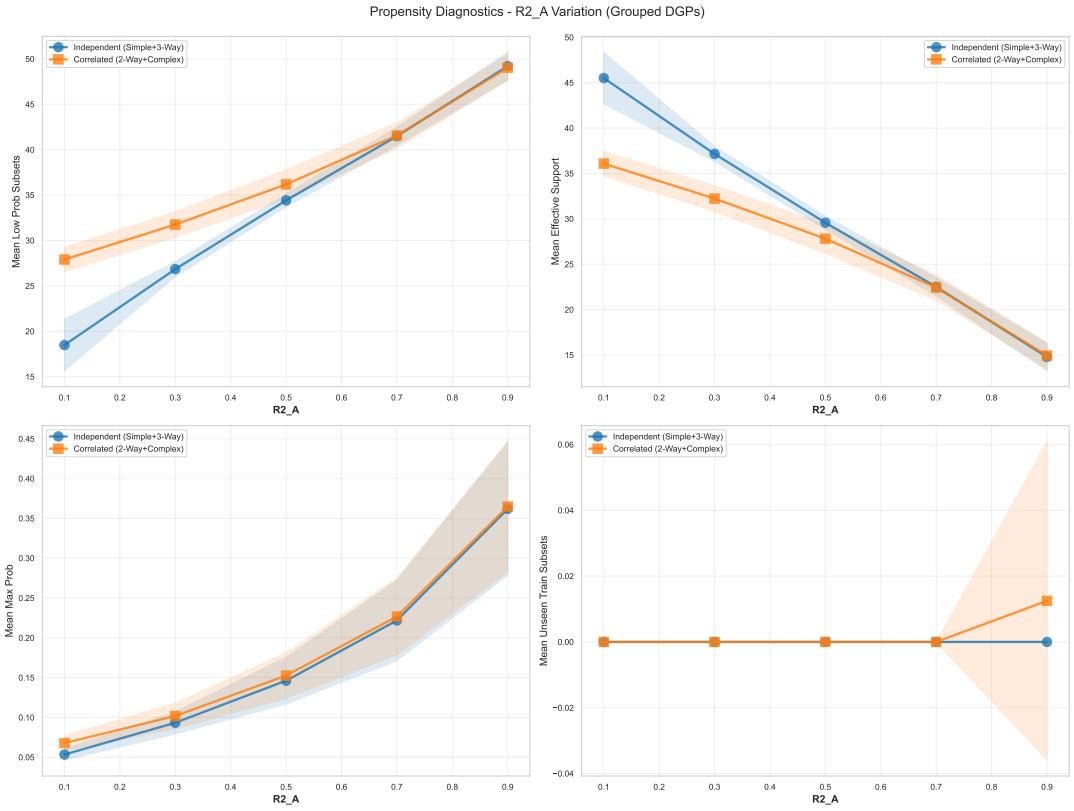

Figure 2.24: Propensity metrics by varying treatment confounding strength.

Detailed Sampling Metrics

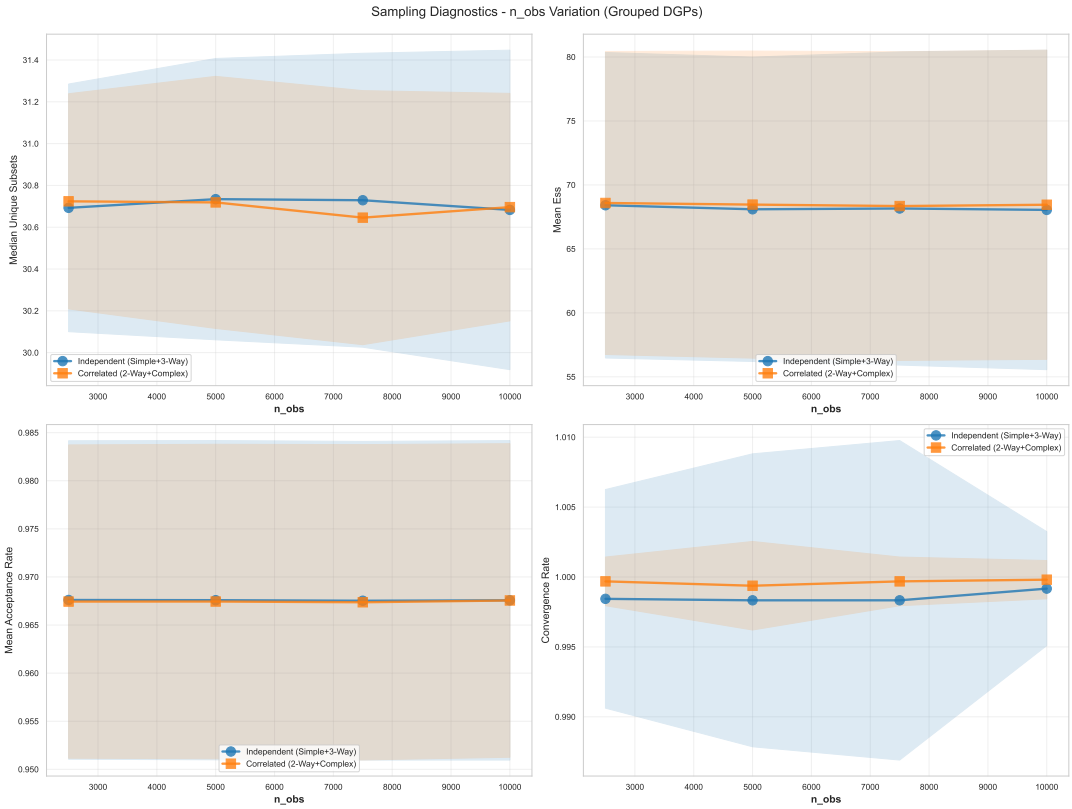

Figure 2.25: Sampling metrics by varying number of observations.

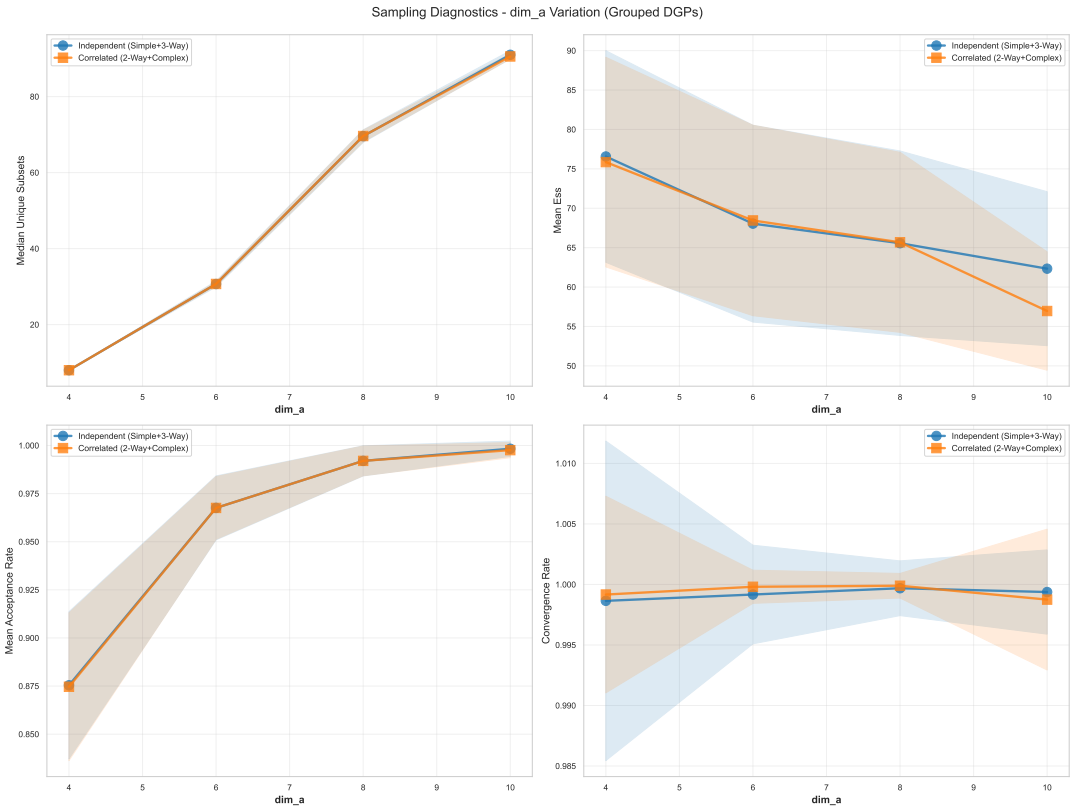

Figure 2.26: Sampling metrics by varying dimension of treatments.

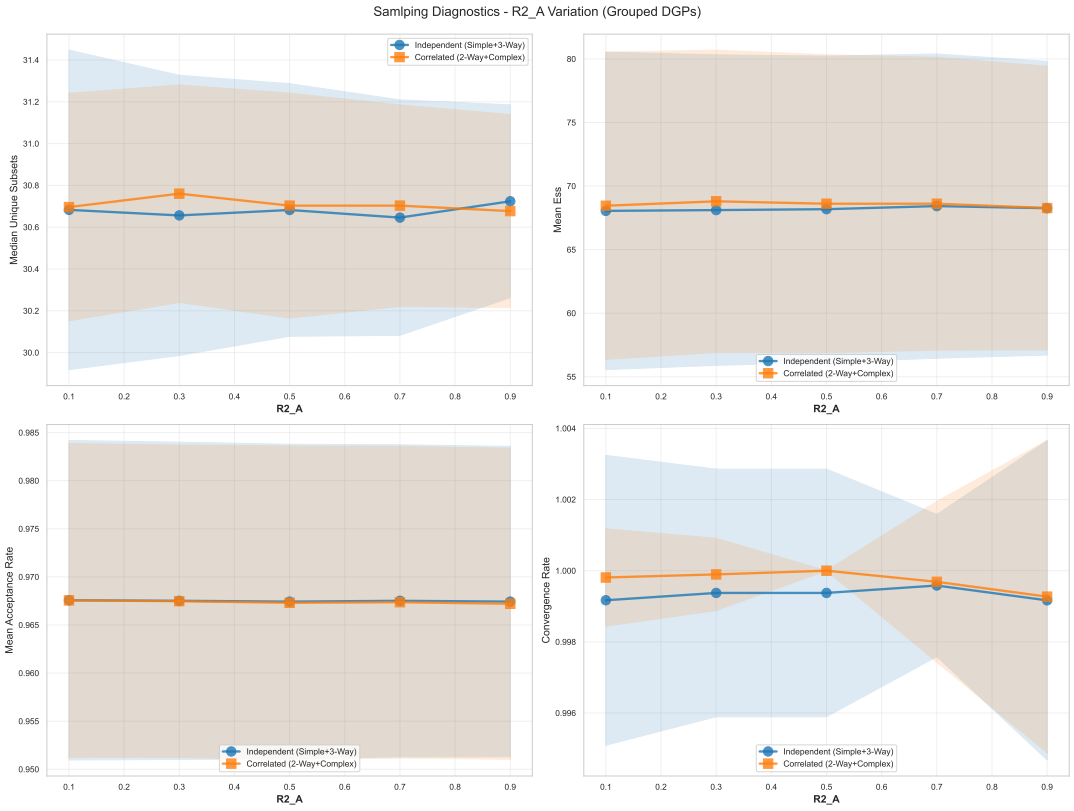

Figure 2.27: Sampling metrics by varying treatment confounding strength.

2.7.3 Details on the Empirical Application: NHANES

Data Set Description

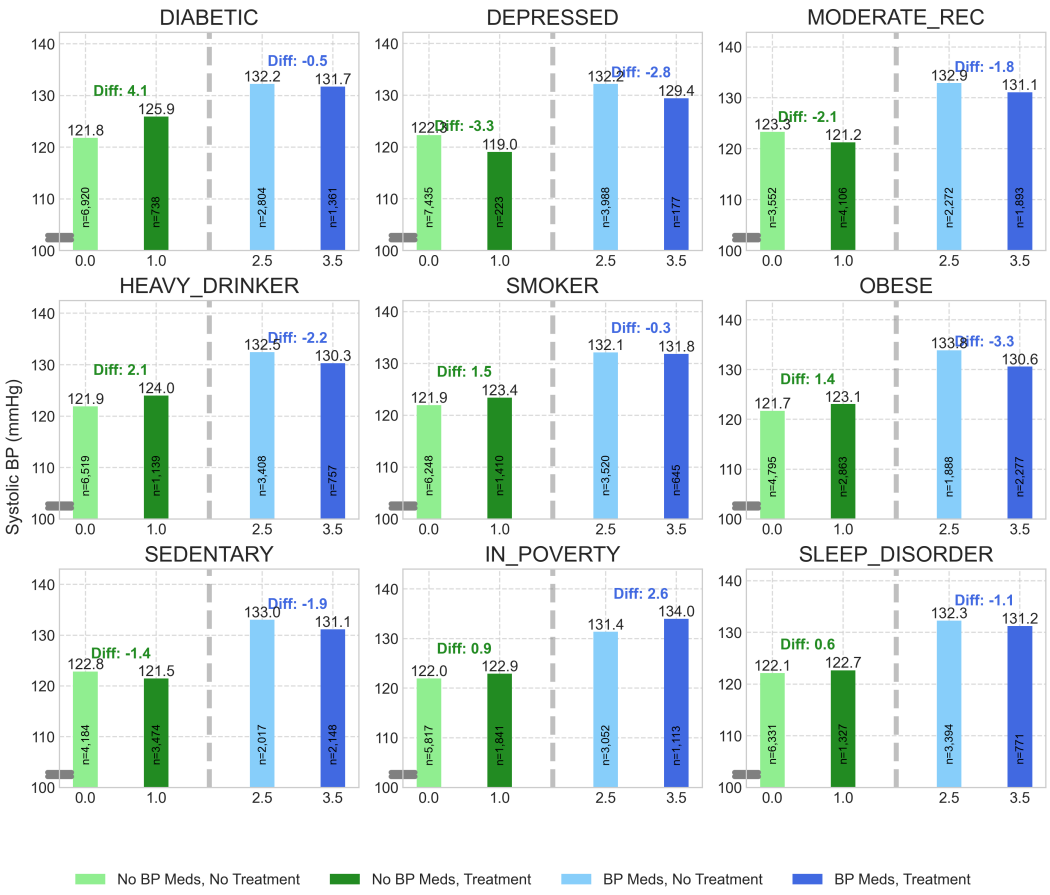

Figure 2.28: Outcome differences by treatment variables with and without blood pressure medication.

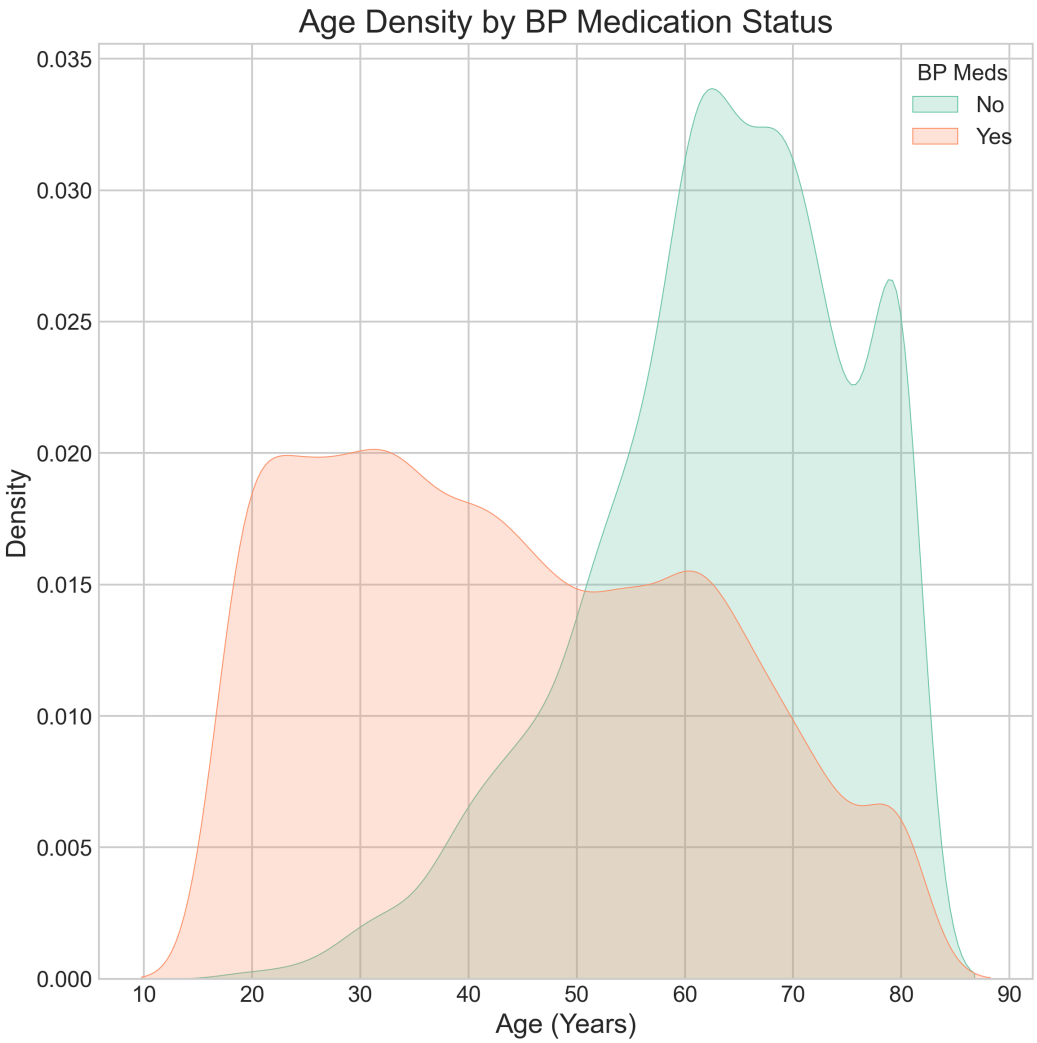

Figure 2.29: Age distribution displayed with and without blood pressure medication being taken.

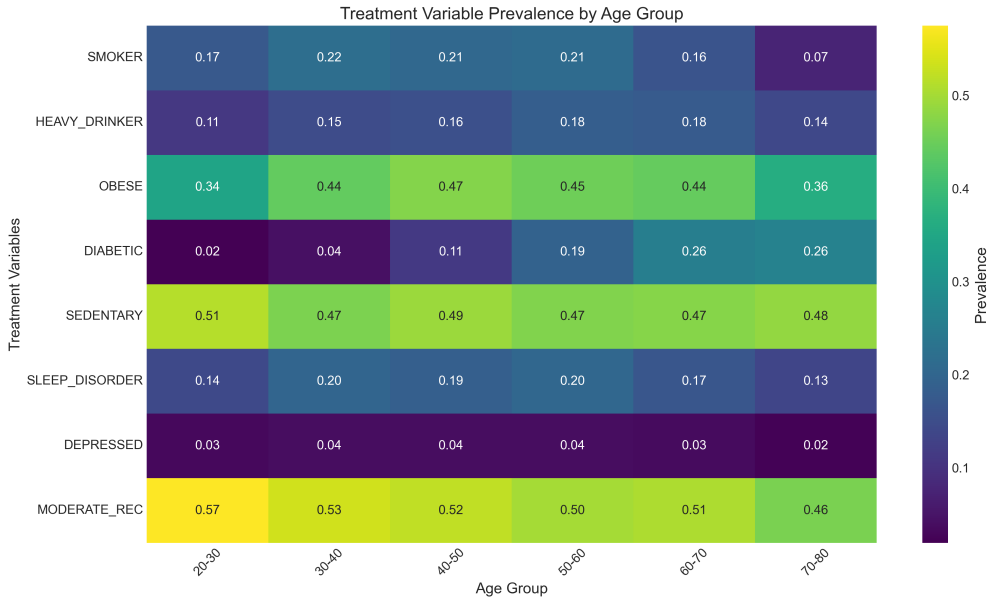

Figure 2.30: Treatment Prevalence by Age.

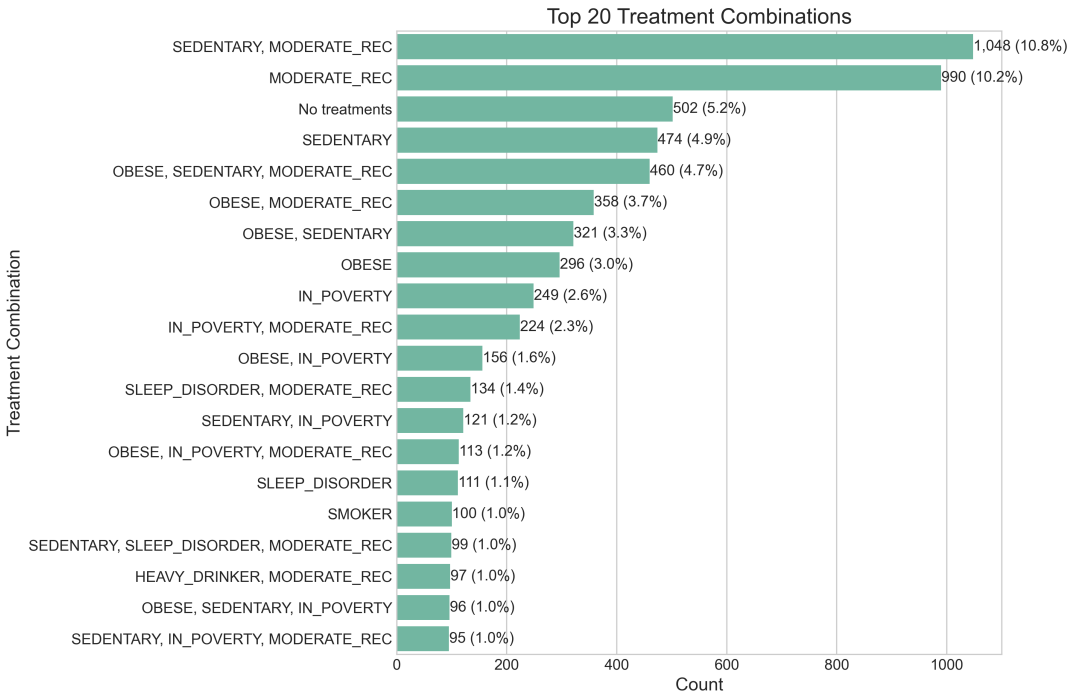

Figure 2.31: Most frequent treatment combinations.

Extended Application Results

Fitted Models by varying Outcome Learner

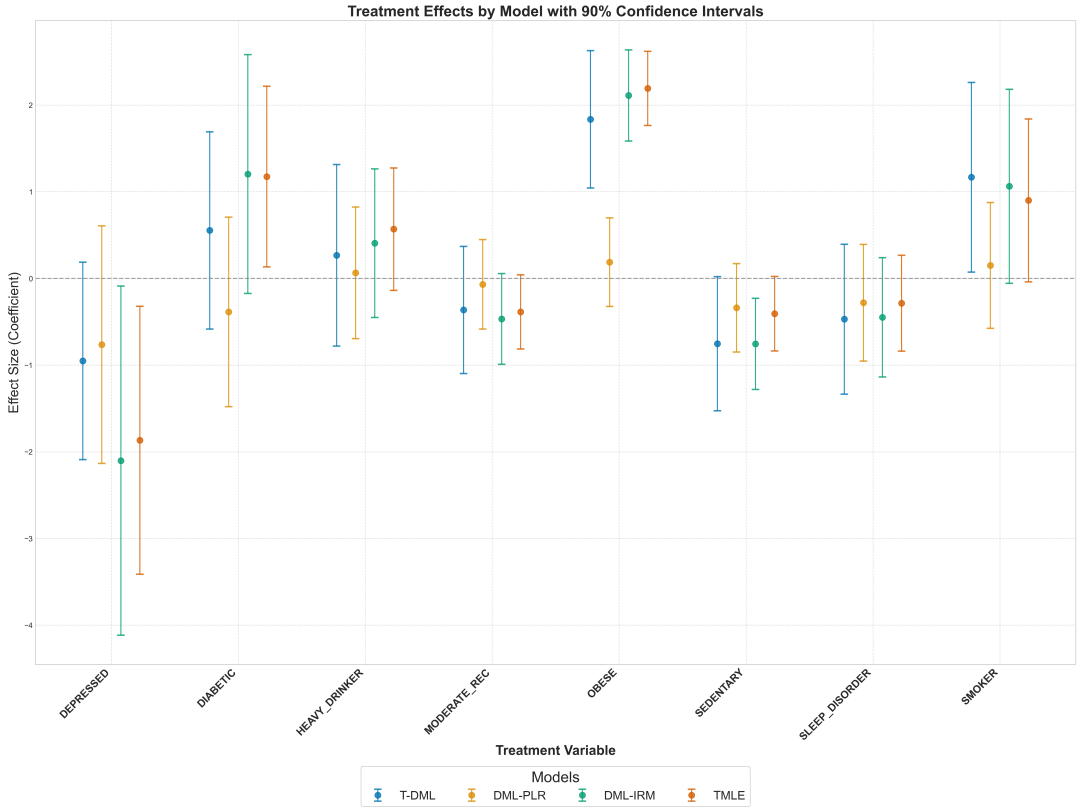

Figure 2.32: μ = LGBM (default)

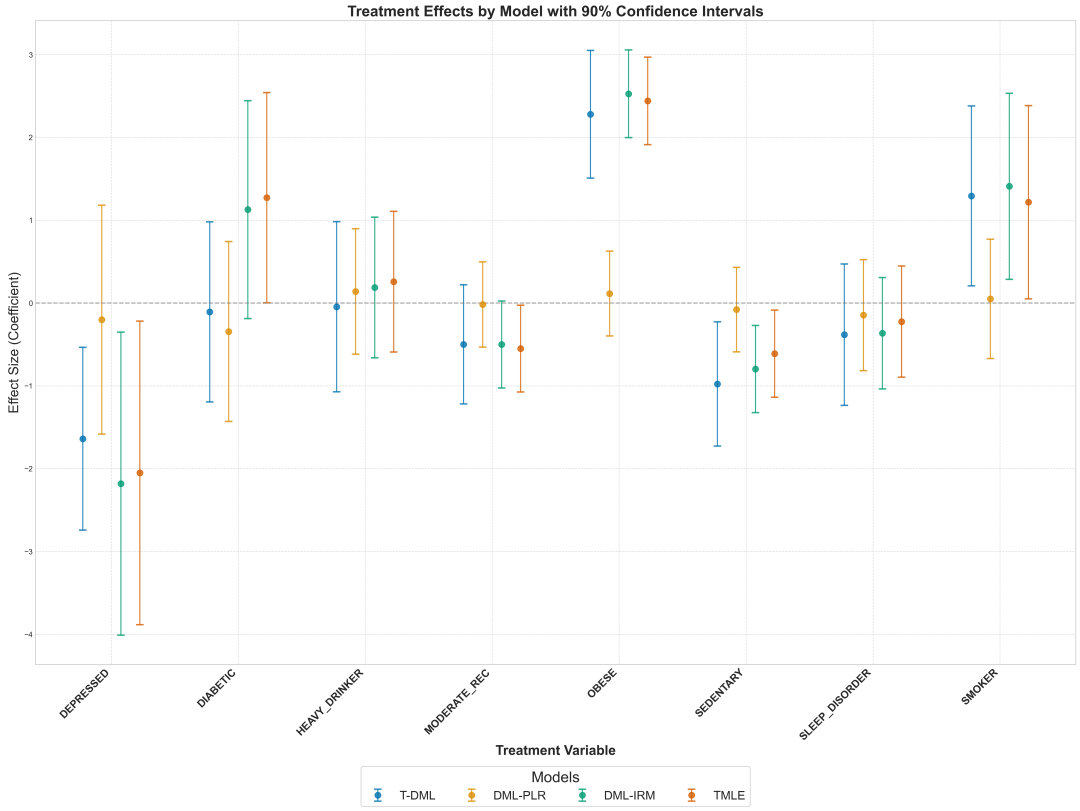

Figure 2.33: μ = Linear_L1

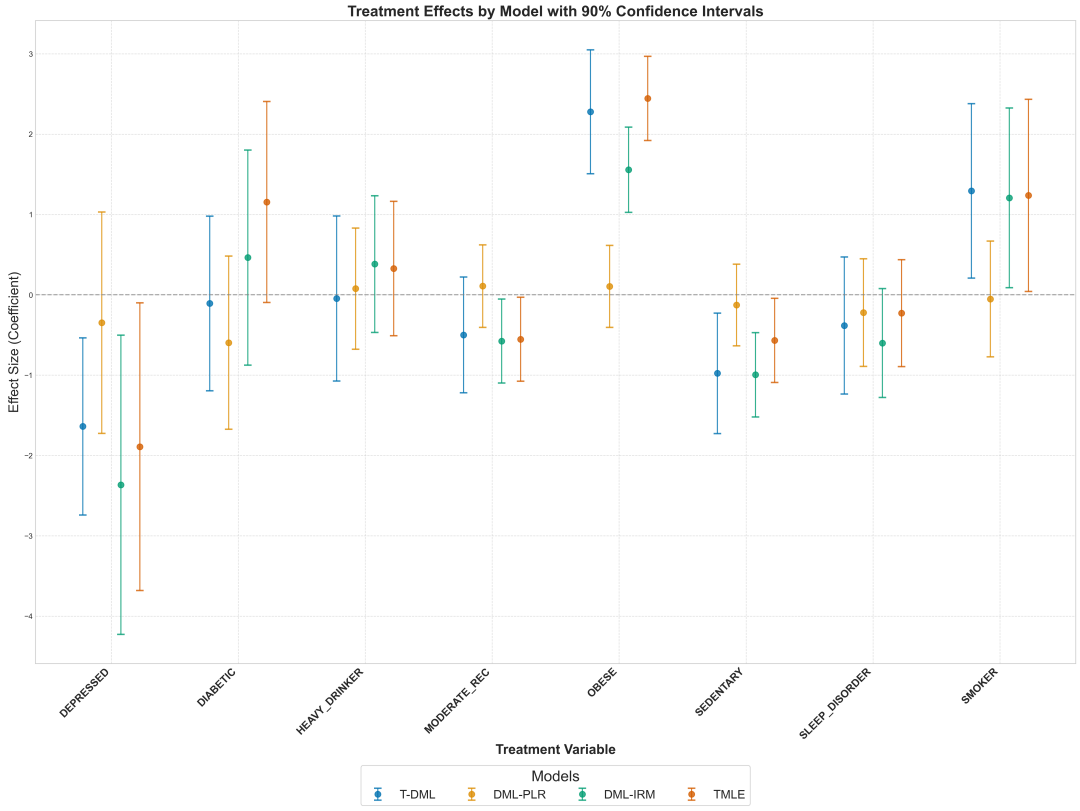

Figure 2.34: μ = Linear_3W_L1

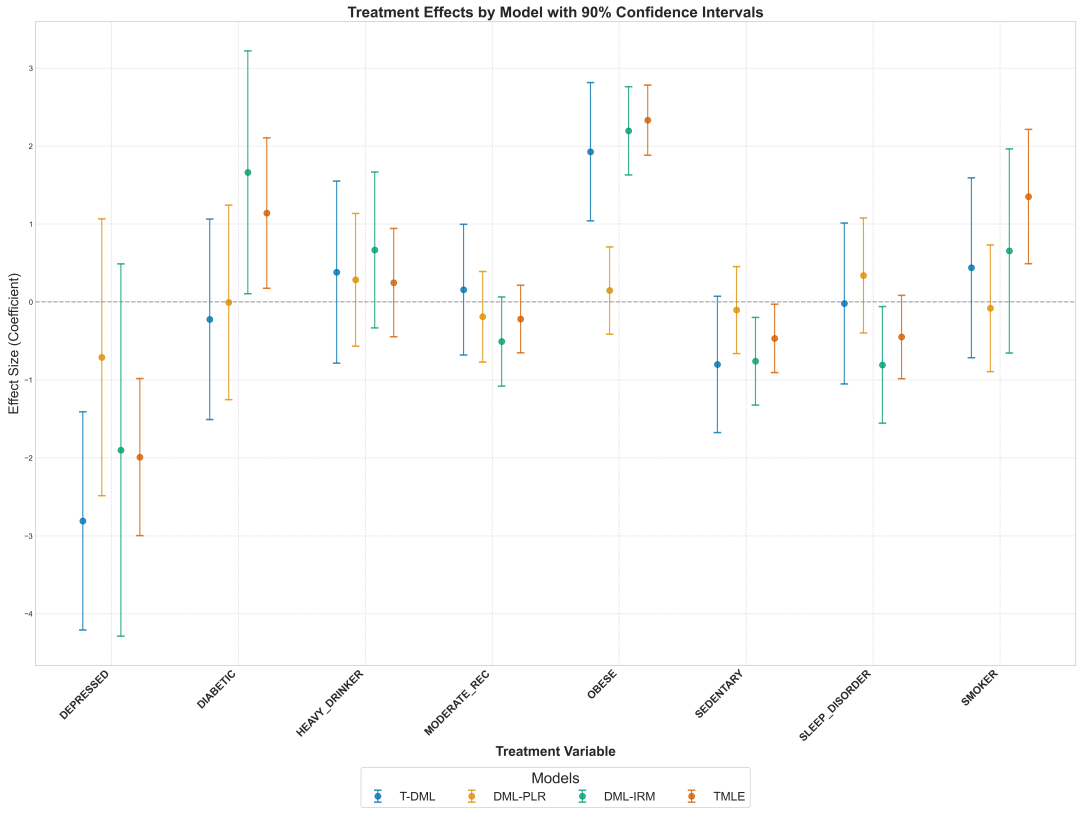

Figure 2.35: μ = NeuralNet

Chapter 3

Calibration Strategies for Robust Causal Estimation: Theoretical and Empirical Insights on Propensity Score-Based Estimators

Joint work with Sven Klaassen (Economic AI), Jannis Kueck (Heinrich Heine University Düsseldorf), and Philipp Bach (Freie Universität Berlin)

Abstract. The partitioning of data for estimation and calibration critically impacts the performance of propensity score based estimators like inverse probability weighting (IPW) and double/debiased machine learning (DML) frameworks. We extend recent advances in calibration techniques for propensity score estimation, improving the robustness of propensity scores in challenging settings such as limited overlap, small sample sizes, or unbalanced data. Our contributions are twofold: First, we provide a theoretical analysis of the properties of calibrated estimators in the context of DML. To this end, we refine existing calibration frameworks for propensity score models, with a particular emphasis on the role of sample-splitting schemes in ensuring valid causal inference. Second, through extensive simulations, we show that calibration reduces variance of inverse-based propensity score estimators while also mitigating bias in IPW, even in small-sample regimes. Notably, calibration improves stability for flexible learners (e.g., gradient boosting) while preserving the doubly robust properties of DML. A key insight is that, even when methods perform well without calibration, incorporating a calibration step does not degrade performance, provided that an appropriate sample-splitting approach is chosen.

Keywords: causal machine learning, calibration, double machine learning, sample splitting, balancing

3.1 Introduction

3.1.1 Motivation

In many settings of the causal inference literature, researchers are interested in the effect of a (binary) treatment $D \in \{0, 1\}$ on an outcome $Y \in \mathbb{R}$. If the treatment is not assigned randomly, a common assumption is the so-called unconfoundedness assumption

$Y(0), Y(1) \perp D \mid X$ assuming that the potential outcomes $Y(1)$ and $Y(0)$ are independent of the actual treatment status D conditional on control variables X. Let

$$m_0(x) := P(D = 1|X = x) = \mathsf{E}[D|X = x]. \tag{3.1}$$

be the propensity score. As famously shown in Rosenbaum and Rubin (1983) conditioning on the propensity score is sufficient to effectively account for the confounding through X

$$Y(0), Y(1) \perp D \mid m_0(X).$$

Consequently, propensity scores are a cornerstone of modern causal inference for addressing confounding in observational studies. They enable balancing treatment and control groups, allowing for unbiased estimation of treatment effects under unconfoundedness. Commonly, propensity scores are used in methods such as inverse probability weighting (IPW), matching, stratification, Bayesian causal inference and more recently double machine learning (DML). Moreover, effective propensity adjustment requires sufficient overlap for propensity score-based estimators. When overlap between treatment and control groups is limited, or treatment assignment is unbalanced, propensity scores can become extreme (i.e. close to 0 or 1) leading to instability of the causal estimates. In such cases, estimators can suffer from inflated variance as extreme weights in IPW disproportionately amplify small errors in propensity score estimation. Similarly, matching algorithms may struggle to find suitable matches, resulting in biased estimates. These challenges highlight the importance of robust and well-calibrated propensity score models to maintain the reliability of causal estimates.

To mitigate instability, researchers often enforce common support by trimming or bounding propensity scores. For example, observations with propensity scores below a certain threshold or outside a predefined range are excluded from analysis. While this approach can reduce variance, it does so at the expense of bias, as it discards valuable information and reduces sample size. Such trade-offs are particularly problematic in small-sample settings, where the exclusion of even a few observations can significantly impact the precision and validity of treatment effect estimates.

To use the propensity score for balancing, the property

$$m_0(X) = \mathsf{E}[D|m_0(X)] \tag{3.2}$$

is crucial, e.g. see Theorem 2 in Rosenbaum and Rubin (1983). In the classification literature, Equation (3.2) is known as the so-called calibration property.[1] Intuitively, a (binary) classifier $\hat{m}(\cdot)$ is well calibrated if the percentage of positive labels ($D = 1$) is approximately m for all instances with $\hat{m}(X) \approx m$. As the true propensity score is the conditional expectation it is calibrated

$$m_0(X) = \mathsf{E}[D|X] = \mathsf{E}[\mathsf{E}[D|X]|\mathsf{E}[D|X]] = \mathsf{E}[D|m_0(X)]. \quad P\text{-a.s.}$$

In most settings the true propensity score is not known such that it is typically estimated via some classification algorithm such as logistic regression, random forest, boosting

[1] To be precise, this notion of calibration is also referred to as conditional calibration, which is equivalent to probabilistic calibration for binary outcomes (Gneiting and Ranjan 2011).

methods or even deep neural networks. Consequently, the resulting estimator $\hat{m}(\cdot)$ of $m_0(\cdot)$ might not be calibrated, e.g. the percentage of treated units with $\hat{m}(X) \approx m$ might differ substantially from m for certain values of $m \in (0, 1)$. Accurately estimating treatment probabilities is crucial for valid causal inference. Inverse propensity score estimators aim to balance covariates between treatment and control groups. This balancing ensures more reliable and unbiased treatment effect estimates by aligning the covariate distributions, improving the comparability of the treatment groups in terms of observable characteristics. Whereas methods like isotonic regression and Platt scaling refer to the calibration of predictions, critical questions remain about the optimal integration of these calibration techniques in causal estimation workflows.

3.1.2 Related Literature

Recent advances in calibration for propensity score estimation and causal inference emphasize three interconnected themes: the adaptation of machine learning calibration techniques to causal settings, stabilization strategies for inverse probability weighting (IPW), and theoretical insights into finite-sample performance. Tan (2017) proposes calibrated estimation for logistic regression propensity scores. Instead of standard maximum likelihood, it directly enforces covariate balance: the method estimates separate propensity scores for the treated and control groups such that, when used for inverse probability weighting, the weighted covariate means in each group match the full-sample means. This approach, extendable with LASSO for high-dimensional data, enhances robustness for causal effect estimation, particularly under model misspecification. Deshpande and Kuleshov (2023) employed single split calibration on data with deterministic treatment assignments and complex settings with hidden confounders. They specifically highlight variance reduction properties and analyze the regret of recalibration of propensity scores. Their proofs are independent of the sample size N, as their calibration framework consists of splitting the data into training and calibration sets. This distinction is crucial for our work, as we focus on sample size dependent calibration algorithms.

Gutman et al. (2024) demonstrated that post-processing propensity scores using methods such as Platt scaling, referred to as post-calibration, improves treatment effect estimation by correcting propensity score distortions beyond covariate balancing. Calibration significantly benefits miscalibrated models like tree-based methods (e.g. gradient boosting), while logistic regression, contrary to the findings of Deshpande and Kuleshov (2023), shows minimal gains. This suggests that post-calibration allows flexible learners to retain their predictive accuracy while enhancing robustness to data challenges such as small sample sizes, model misspecification, class imbalance, or limited overlap, thus challenging the conventional trade-off between model complexity and calibration. Ballinari and Bearth (2025) use nested cross-fitting, reserving distinct data folds for propensity estimation and calibration, which is a straightforward application of the double machine learning theory of Chernozhukov et al. (2018). They reported instability in small samples, a limitation attributed to the reduced effective sample sizes for both steps. These studies collectively identify a tension between calibration flexibility and stability, particularly in finite-sample regimes. Efforts to address these challenges have taken different paths. For instance, van der Laan et al. (2024a) proposed a stratified calibration for treated and control groups to stabilize IPW weights. In another work, they extended calibration frameworks to esti-

mate heterogeneous treatment effects (van der Laan et al. 2023). Meanwhile, van der Laan et al. (2024b) extend automatic debiased machine learning (autoDML) (Chernozhukov et al. 2022b, 2024) by calibrating both outcome regression and Riesz representer. They demonstrate that the calibration step yields doubly robust asymptotically linear estimators, reducing nuisance rate requirements.

Theoretical work by Gamarnik (1998), Mammen and Yu (2007) and Wüthrich and Ziegel (2023) established consistency guarantees for isotonic regression, but practical implementations often struggle with convergence rates in small samples, as shown by Yang and Barber (2019). Furthermore, Yang and Barber (2019) established that isotonic projection is non-contractive under the ℓ_∞ norm, exacerbating edge instability in propensity score estimates. This theoretical insight justifies the empirical requirement of clipping extreme probabilities, which explains the unstable treatment effect estimates in Ballinari and Bearth (2025) using isotonic regression under limited overlap.

This paper systematically evaluates how calibration performance depends on data partitioning for propensity estimation and calibration. While existing studies fix specific splitting strategies (e.g., single-split or nested cross-fitting), we show that the choice of partitioning interacts critically with sample size, clipping thresholds, and complexity of the data generating process. For instance, van der Laan et al. (2024a) suggest calibration on the full-sample, which avoids reserving data exclusively for calibration. This approach can mitigate instability without sacrificing theoretical guarantees, a hypothesis we test across multiple data-generating processes (DGP). Similarly, we reconcile the debates about stratified versus pooled calibration by demonstrating that efficient reuse of cross-fitted propensity scores obviates the need for group-specific adjustments in many settings. Further, beyond the work of Ballinari and Bearth (2025), we provide a theoretical extension of the double machine learning theory to allow for different sample-splitting schemes. By synthesizing these insights, our work clarifies when and how calibration improves ATE estimation, providing a bridge between theoretical calibration properties and practical implementation challenges.

Plan of the Paper. The remainder of the paper is organized as follows. Section 3.2 introduces propensity score calibration, highlights different approaches, and details the properties of isotonic regression. Section 3.3 compares and proposes calibration algorithms for estimating the average treatment effect and establishes theoretical guarantees under double machine learning (DML), including convergence rates and asymptotic normality. Section 3.4 demonstrates robustness through simulations across diverse and challenging data-generating processes. The appendices 3.6.1 and 3.6.2 provide proofs for the main theoretical results. Sections 1 and 2 of Online Appendix 1 present further details on calibration and partially linear regression models, while Sections 3 and 4 cover implementation details for reproducibility, extended simulation results, and sensitivity analyses.

3.2 Propensity Score Calibration

Let $D \in \{0, 1\}$ be a binary treatment variable with covariates $X \in \mathcal{X} \subseteq \mathbb{R}^d$. The propensity score is defined as $m_0(x) := P(D = 1 \mid X = x) = \mathsf{E}[D \mid X = x]$. Since the propensity score represents a conditional expectation, it is calibrated such that $m_0(X) = \mathsf{E}[D \mid m_0(X)]$. The goal is to achieve a similar balancing property of an estimated version of $m_0(X)$. Given an estimate $\hat{m}(X)$ of $m_0(X)$, we consider popular

calibration methods such as isotonic regression. Generally, we consider calibration procedures based on the pseudo-sample $((D_1, \hat{m}(X_1)), \ldots, (D_N, \hat{m}(X_N)))$. The calibration algorithm approximates $\mathsf{E}[D|\hat{m}(X)]$ which typically differs from $\mathsf{E}[D|m_0(X)]$. In the following, we denote the calibrated propensity score by $\tilde{m} : \mathcal{X} \to [0,1], \quad x \mapsto \tilde{m}(x)$.

3.2.1 Rate Comparison $\hat{m}(\cdot)$ and $\tilde{m}(\cdot)$

For any estimate $\hat{m}(\cdot)$ of $m_0(\cdot)$ the mean-squared-error decomposes as

$$\|\hat{m}(X) - m_0(X)\|_{P,2} = \left(\mathbb{E}\left[\mathrm{Var}\big(m_0(X)|\hat{m}(X)\big)\right] + \|\mathbb{E}[m_0(X)|\hat{m}(X)] - \hat{m}(X)\|_{P,2}^2\right)^{1/2}. \tag{3.3}$$

The first term denotes the expected precision of $\hat{m}(\cdot)$, while the second term is the calibration error. Calibration procedures minimizing mean square error in the pseudo-sample approximate $\mathsf{E}[D|\hat{m}(X)]$, reducing this error.

Assumption 9. *Let $\tilde{m}(X)$ be an estimator of $\tilde{m}_0(X) := \mathsf{E}[D|\hat{m}(X)]$, with $\|\tilde{m}(X) - \tilde{m}_0(X)\|_{P,2} \leq \tilde{\varepsilon}_N$.*

Lemma 2. *Under Assumption 9:*

$$\|\tilde{m}(X) - m_0(X)\|_{P,2} \leq \left(\mathbb{E}\left[\mathit{Var}\big(m_0(X)|\hat{m}(X)\big)\right]\right)^{1/2} + \tilde{\varepsilon}_N. \tag{3.4}$$

Comparing (3.3) and (3.4) shows rate differences when the calibration error

$$\|\mathbb{E}[m_0(X)|\,\hat{m}(X)] - \hat{m}(X)\|_{P,2}^2$$

dominates. Improvements occur if $\tilde{\varepsilon}_N = o(\|\mathbb{E}[m_0(X)|\hat{m}(X)] - \hat{m}(X)\|_{P,2})$. This is relevant in double ML settings requiring $\|\hat{m}(X) - m_0(X)\|_{P,2} = o(N^{-1/4})$: If $\tilde{\varepsilon}_N = o(N^{-1/4})$, $\tilde{m}(\cdot)$ satisfies the requirement with potential rate improvements from faster convergence.

3.2.2 Calibration Methods

Model calibration has roots in meteorological forecasting to address reliability in probabilistic weather predictions (Toth et al. 2006, Dawid 2014, Gneiting 2014). Calibration is often achieved via post-hoc point calibrators that transform $\hat{m}(X)$ into $f(\hat{m}(X))$, balancing two aims: (1) calibration validity, and (2) preservation of predictive sharpness (i.e., $f \circ \hat{m}(X)$ approximates $\hat{m}$'s discriminative power) (Gneiting et al. 2007, Gupta et al. 2020). Common approaches include parametric methods like Platt scaling, which fits a logistic sigmoid to $f(X)$ (Platt 1999, Cox 1958); non-parametric methods such as histogram binning, partitioning predictions into fixed intervals (Zadrozny and Elkan 2001, Gupta and Ramdas 2021), and isotonic regression, learning a monotonic transform via empirical risk minimization (Zadrozny and Elkan 2002, Barlow and Brunk 1972); as well as conformal methods like Venn-Abers predictors, refining calibration through cross-conformal inference (Vovk and Petej 2014). Isotonic calibration, while distribution-free and tuning parameter-free, achieves asymptotic guarantees with $O_P(N^{-1/3})$ convergence (Zhang 2002, van der Laan et al. 2023). In contrast, histogram binning requires explicit

bin specification and trades flexibility for finite-sample validity (Gupta and Ramdas 2021). We focus on three common calibration methods, noting that calibration is an active area of research with potential for improvements and new proposals that require further simulation testing[2].

Isotonic Regression

Definition 1. *Given an estimate $\hat{m}(\cdot)$ of the propensity score, perform an isotonic regression as $f = \arg\min_{f \in \mathcal{F}_{iso}} \sum_{i=1}^{N} (D_i - (f \circ \hat{m})(X_i))^2$ with $\mathcal{F}_{iso}$ being the set of non-decreasing functions. The calibrated propensity score is then given by $\tilde{m} = f \circ \hat{m}$.*

Especially, the in-sample calibration property $\mathsf{E}_n[D|\tilde{m}(X_i)] = \tilde{m}(X_i)$ for $i \in \{1, \ldots, N\}$, seems to be desirable (e.g. Wüthrich and Ziegel (2023)). We consider an estimated propensity score $\hat{m}(\cdot)$ based on a separate sample such that $\hat{m}(\cdot)$ can be considered a fixed function. Let $U := \hat{m}(X)$ and define the pseudo sample as $Z := (D, U)$, where $(Z_i)_{i=1}^N$ are iid. copies of Z. Consider the following assumption

Assumption 10. *The regression function $\tilde{m}_0(u) := \mathsf{E}[D|U = u]$ is monotone.*

Remark 4. *Assumption 10 is substantially weaker than monotonicity in the original covariates X, and typically easier to satisfy: it only requires that after projecting X onto the preliminary propensity score $\hat{m}(X)$, the conditional expectation is (approximately) monotone. If $\hat{m}(\cdot)$ is non-injective, different values of X can map to the same U, and $\tilde{m}_0(U)$ averages over these X values—potentially smoothing out non-monotonicities in X and making monotonicity in U more plausible. Only if $\hat{m}(\cdot)$ is bijective does monotonicity of $\tilde{m}_0(U)$ become equivalent to monotonicity of $\mathbb{E}[D|X = x]$ under the reparametrization $u = \hat{m}(x)$. Importantly, even if $\tilde{m}_0(U)$ is not perfectly monotone but only approximately so (i.e., ϵ_{iso}-monotone in the sense of Yang and Barber 2019), isotonic regression still projects to the closest monotone function. The estimation error can then be bounded by the sum of the stochastic error and ϵ_{iso}-monotone, providing robustness to small violations of strict monotonicity in practical settings.*

Define $\tilde{m}$ as the estimator obtained by isotonic regression on the sample $(Z_i)_{i=1}^N$.

Lemma 3 (Convergence of Isotonic Regression)**.** *Under Assumption 10, the isotonic regression estimator $\tilde{m}$ satisfies: $\|\tilde{m}(U) - \tilde{m}_0(U)\|_{P,2} = O_P(N^{-1/3})$.*

The rate follows from the bracketing entropy bound in Theorem 2.7.5 of Vaart and Wellner (2023) for monotonic functions. Relying on Lemma 3.4.3 of Vaart and Wellner (2023) gives the desired rate as mentioned in the corresponding Section 3.4.3.2. Earlier work by Birman and Solomjak (1967) and van de Geer (2000) established foundational approximation and entropy arguments, extended to additive isotonic regression in Mammen and Yu (2007). For detailed convergence properties see e.g. Zhang (2002).

Notably, similar convergence rates hold in settings where the data is not strictly i.i.d., such as when using sample splitting or cross-fitting to estimate $\hat{m}(\cdot)$. Theorem 2 in van der Laan et al. (2024a) demonstrates that under cross-fitting regimes, where $\hat{m}(\cdot)$

[2]The description of Platt scaling and Venn-Abers calibration can be found in Online Appendix 3.6.3. As our theoretical results are built around isotonic regression, we introduce it here in more detail.

is trained on an independent sample and applied to the estimation sample, the pseudo-sample Z is still sufficiently weakly dependent for the $O_P(N^{-1/3})$ rate to hold. This aligns with the mean squared error (MSE) decomposition in (3.3), where the $O_P(N^{-1/3})$ $L_2(P)$ convergence rate of isotonic regression (Lemma 3) ensures that the calibration error term $\|\mathbb{E}[m_0(X)|\hat{m}(X)] - \hat{m}(X)\|_{P,2}^2$ decays as $O_P(N^{-2/3})$. Theorem 1 in van der Laan et al. (2024a) further establishes distribution-free calibration guarantees, bounding the calibration error by $O_P(N^{-2/3})$, irrespective of the smoothness of the inverse propensity score or the dimension of the covariates. Together, these results accommodate the calibration algorithms introduced in Section 3.3.

3.3 Calibration for Double Machine Learning

The following section combines calibration and double machine learning. The first part focuses on high-level conditions for different double machine learning algorithms, whereas the second part states explicit conditions for particular double machine learning models with isotonic regression.

3.3.1 Double Machine Learning Theory and Algorithms

In this section, we state conditions which enable a re-estimation step for nuisance estimators in the double machine learning framework if the complexity of the re-estimation procedure is not too large. As in Chernozhukov et al. (2018) we denote $\theta_0 \in \Theta \subset \mathbb{R}$ the parameter of interest. The leading example is the average treatment effect (ATE) $\theta_0 = \mathsf{E}[Y(1) - Y(0)]$. Further, we assume that θ_0 satisfies the moment condition,

$$\mathsf{E}_P[\psi(W; \theta_0, \eta_0)] = 0, \tag{3.5}$$

where ψ is a known score function, the data W is a random element in $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ with probability measure $P \in \mathcal{P}_N$ and η_0 is the true value of the nuisance parameter $\eta \in T$, where T is the convex subset of a normed vector space with norm $\|\cdot\|_T$.
The previous setting describes the standard double machine learning framework introduced in Chernozhukov et al. (2018). For simplicity, we restrict ourselves to the case of linear score functions, that is

$$\psi(w; \theta, \eta) = \psi^a(w; \eta)\theta + \psi^b(w; , \eta), \tag{3.6}$$

for all $w \in \mathcal{W}$, $\theta \in \Theta$ and $\eta \in T$. Further, since we would like to consider a scenario with a re-estimation or calibration step, which might not affect all nuisance parameters, we define $\eta_0 = (\eta_0^{(1)}, \eta_0^{(2)})$, where $\eta_0^{(2)}$ should be re-estimated as for example when $\eta_0^{(2)}$ is a propensity score to be calibrated. Correspondingly define $T = T^{(1)} \times T^{(2)}$.

Algorithm 1 recaps the standard version of the double machine learning algorithm based on cross-fitting (cf. Definition 3.2 in Chernozhukov et al. (2018)).
Let $(W_i)_{i=1}^N$ be iid. copies of W with probability measure P. To simplify notation, assume that N is divisible by K.

The standard DML 2 algorithm employs cross-fitting to handle the complexity of the estimated nuisance elements $\hat{\eta}_{0,k}$.

Algorithm 1 (uncalibrated) DML 2 Algorithm

1: **Input:** Data $(W_i)_{i=1}^N$. A K-fold random partition $(I_k)_{k=1}^K$ of $[N] = \{1, \dots, N\}$ such that each fold I_k is of size $n = N/K$. For each $k \in [K] = \{1, \dots, K\}$, define $I_k^c := \{1, \dots, N\} \setminus I_k$.
2: For each $k \in [K]$, fit a machine learning estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_0\big((W_i)_{i \in I_k^c}\big)$$

of η_0, where $\hat{\eta}_{0,k}$ is a random element in T, where the randomness only depends on the $(W_i)_{i \in I_k^c}$.
3: Construct the estimator $\hat{\theta}_0$ as the solution to

$$\frac{1}{K} \sum_{k=1}^{K} \mathsf{E}_{n,k}[\psi(W; \hat{\theta}_0, \hat{\eta}_{0,k})] = 0,$$

where $\mathsf{E}_{n,k}[\psi(W)] = n^{-1} \sum_{i \in I_k} \psi(W_i)$ is the empirical expectation over $(W_i)_{i \in I_k}$.

Remark 5. *Theorem 3.1 in Chernozhukov et al. (2018) shows that the estimator according to Algorithm 1 is asymptotically normally distributed. More specifically, it holds*

$$\sqrt{N}\sigma^{-1}(\hat{\theta}_0 - \theta_0) = \frac{1}{N} \sum_{i=1}^{N} \bar{\psi}(W_i) + O_P(\rho_N) \rightsquigarrow \mathsf{N}(0, 1) \tag{3.7}$$

uniformly over $P \in \mathcal{P}_N$, where the size of the remainder term obeys $\rho_N := N^{-1/2} + r_N + r'_N + N^{1/2}\lambda_N + N^{1/2}\lambda'_N \lesssim \delta_N$, with $\delta_N \geq N^{-1/2}$. Here, $\bar{\psi}(\cdot) := \sigma^{-1} J_0^{-1} \psi(\cdot; \theta_0, \eta_0)$ is the influence function and the approximate variance is $\sigma^2 := J_0^{-2} \mathsf{E}_P[\psi(W; \theta_0, \eta_0)^2]$.

In Remark 5 it is assumed that $\hat{\eta}_{0,k} = (\hat{\eta}_{0,k}^{(1)}, \hat{\eta}_{0,k}^{(2)}) \in \mathcal{T}_N$ with probability $1 - o(1)$, where $\mathcal{T}_N$ is a suitable nuisance realization set. To enable the re-estimation of nuisance elements $\hat{\eta}_{0,k}^{(2)}$ the algorithm and the nuisance realization set $\mathcal{T}_N$ has to be slightly adapted. Ballinari (2024) present a simple and straightforward adaptation using nested cross-fitting, which we refer to as Algorithm 2[3]. The standard double machine learning procedure in Algorithm 1 uses cross-fitting to handle the complexity of estimated nuisance elements $\hat{\eta}_{0,k}$. Algorithm 2 is a straightforward extension, which leaves the cross-fitting unchanged. The approach just employs a nested sample splitting procedure, such that the calibrated nuisance elements $\tilde{\eta}_{0,k}^{(2)}$ still depend only on the observations of the training sample $(W_i)_{i \in I_k^c}$. As a consequence, only the predictive performance of the calibrated nuisance estimators must be ensured.

Remark 6. *Let $\tilde{\theta}_0$ be the estimator according to Algorithm 2. Under the assumptions in Theorem 3.1 in Chernozhukov et al. (2018), Equation* (3.7) *in Remark 5 holds analogously for $\tilde{\theta}_0$.*

Typically, Assumption 3.2 in Theorem 3.1 in Chernozhukov et al. (2018) requires high-quality nuisance estimators. In particular, the re-estimation procedure (calibration) has to converge sufficiently fast, i.e. $\|\tilde{\eta}_{0,k}^{(2)} - \eta_0^{(2)}\|_{P,2} \lesssim \varepsilon_N = o(N^{-1/4})$. Since calibration properties are most important for small samples, further splitting of the training sample for calibration might not be desirable.

Therefore, we state Assumptions 3.1 and 3.2 of Chernozhukov et al. (2018) for an adapted nuisance realization set $\widetilde{\mathcal{T}}_N$. It is worth noting that in the following assumption the

[3]The pseudocode for **Algorithm 2** appears in Online Appendix 3.6.4.

calibrated nuisance elements $\tilde{\eta}_0^{(2)}((W_i)_{i\in[N]}))$ may depend on the full data. This allows us to introduce new estimation algorithms that rely on more sophisticated splitting rules for calibration.

Assumption 11. *Let $c_0 > 0$, $c_1 > 0$, and $q \geq 2$ be some finite constants such that $c_0 \leq c_1$, and let $\{\delta_N\}_{N\geq 1}$ and $\{\Delta_N\}_{N\geq 1}$ be some sequences of positive constants converging to zero such that $\delta_N \geq N^{-1/2}$. Also, let $K \geq 2$ be some fixed integer, and let $\{\mathcal{P}_N\}_{N\geq 1}$ be some sequence of sets of probability distributions P of $\mathcal{W}$ on W.*

Assumption 3.1 (Linear scores with approximate Neyman orthogonality) *For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold:* (i) *The true parameter value θ_0 obeys (3.5).* (ii) *The score ψ is linear in the sense of (3.6).* (iii) *The map $\eta \mapsto E_P[\psi(W;\theta,\eta)]$ is twice continuously Gateaux-differentiable on T.* (iv) *The score ψ obeys the Neyman orthogonality or, more generally, the Neyman λ_N near-orthogonality condition at (θ_0,η_0) with respect to the nuisance realization set $\widetilde{\mathcal{T}}_N \subset T$ for $\lambda_N := \sup_{\eta\in\mathcal{T}_N} |\partial_\eta \mathsf{E}_P[\psi(W;\theta_0,\eta_0)][\eta-\eta_0]| \leq \delta_N N^{-1/2}$.* (v) *The identification condition holds; namely, the singular values of the matrix $J_0 := \mathsf{E}_P[\psi^a(W;\eta_0)]$ are between c_0 and c_1.*

Assumption 3.2 (Score regularity and quality of nuisance parameter estimators) *For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold:*

(a) Given a random subset I of $[N]$ of size $n = N/K$, the nuisance parameter estimator $\tilde{\eta}_0 := (\hat{\eta}_0^{(1)}((W_i)_{i\in I^c}), \tilde{\eta}_0^{(2)}((W_i)_{i\in[N]}))$ belongs to the realization set $\widetilde{\mathcal{T}}_N$ with probability at least $1-\Delta_N$, where $\widetilde{\mathcal{T}}_N = \widetilde{\mathcal{T}}_N^{(1)} \times \widetilde{\mathcal{T}}_N^{(2)}$ contains η_0 and is constrained by the following conditions.

(b) The moment conditions hold:

$$m_N := \sup_{\eta\in\widetilde{\mathcal{T}}_N} \mathsf{E}_P[|\psi(W;\theta_0,\eta)|^q]^{1/q} \leq c_1; \quad m'_N := \sup_{\eta\in\widetilde{\mathcal{T}}_N} \mathsf{E}_P[|\psi^a(W;\eta)|^q]^{1/q} \leq c_1.$$

(c) The following conditions on the statistical rates r_N, r'_N, and λ'_N hold:

$$r_N := \sup_{\eta\in\widetilde{\mathcal{T}}_N} |\mathsf{E}_P[\psi^a(W;\eta)] - \mathsf{E}_P[\psi^a(W;\eta_0)]| \leq \delta_N,$$

$$r'_N := \sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \left(\mathsf{E}_P\left[\left(\psi\big(W;\theta_0,(\eta^{(1)},\eta_0^{(2)})\big) - \psi\big(W;\theta_0,(\eta_0^{(1)},\eta_0^{(2)})\big)\right)^2\right]\right)^{1/2} \leq \delta_N,$$

$$\lambda'_N := \sup_{r\in(0,1),\eta\in\widetilde{\mathcal{T}}_N} \left|\partial_r^2 \mathsf{E}_P[\psi(W;\theta_0,\eta_0 + r(\eta-\eta_0))]\right| \leq \delta_N/\sqrt{N}.$$

(d) The variance of the score ψ is non-degenerate: $c_0 \leq \mathsf{E}_P[\psi(W;\theta_0,\eta_0)^2]$.

Indeed, if the calibration method is not too complex (see Assumption 12), for example when isotonic regression is used for calibration, the additional sample split in Algorithm 2 can be avoided by calibrating the predictions on each "test"-fold I_k which are used to estimate the target parameter θ_0. This procedure is described in Algorithm 3. Although the calibrated nuisance estimator depends on the full data in this case, we will show an analog result as in Theorem 3.1 in Chernozhukov et al. (2018) under the Assumptions

Algorithm 3 (k-fold cross-fitting calibration) DML 2 Algorithm

1: **Input:** Data $(W_i)_{i=1}^N$. A K-fold random partition $(I_k)_{k=1}^K$ of $[N] = \{1, \dots, N\}$ such that each fold I_k is of size $n = N/K$. For each $k \in [K] = \{1, \dots, K\}$ define $I_k^c := \{1, \dots, N\} \setminus I_k$.
2: For each $k \in [K]$, fit a machine learning estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_0\big((W_i)_{i \in I_k^c}\big)$$

of η_0, where $\hat{\eta}_{0,k}$ is a random element in T, where the randomness only depends on the $(W_i)_{i \in I_k^c}$.
3: For each $k \in [K]$, rely on estimated nuisance element $\hat{\eta}_{0,k}^{(2)}$ to fit a re-estimation procedure

$$\tilde{\eta}_{0,k}^{(2)} = \tilde{\eta}_0^{(2)}\big((W_i)_{i \in I_k}, \hat{\eta}_{0,k}^{(2)}\big)$$

of $\eta_0^{(2)}$, where $\tilde{\eta}_{0,k}^{(2)}$ is a random element in $T^{(2)}$.
4: Construct the estimator $\tilde{\theta}_0$ as the solution to

$$\frac{1}{K} \sum_{k=1}^{K} \mathsf{E}_{n,k}\big[\psi\big(W; \tilde{\theta}_0, (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)})\big)\big] = 0,$$

where $\mathsf{E}_{n,k}[\psi(W)] = n^{-1} \sum_{i \in I_k} \psi(W_i)$ is the empirical expectation over $(W_i)_{i \in I_k}$.

11 and 12 in Theorem 1. As a slight modification of Algorithm 3 one can use different K-fold cross-fitting procedures for the estimated nuisance elements. For example, 2-fold cross-fitting as described in Algorithm 4, uses half of the data for nuisance estimation and the other half for calibration[4]. Consequently, the calibration step might be more stable. Deshpande and Kuleshov (2023) employ a non-cross-fitted version of Algorithm 4, where the data is split for calibration only.

Another option is to simultaneously calibrate all cross-fitted predictions $\hat{\eta}_0^{(2)}$ as described in Algorithm 5. The main difference between Algorithm 3 and 5 is the dependency structure of the data used to calibrate the nuisance elements. In Algorithm 3 the recalibration is fitted on i.i.d. samples conditional on the corresponding "training"-fold I_k^c, whereas in Algorithm 5 samples used for the calibration step have a complex dependency structure.

Assumption 12 (Calibration Complexity)**.** *Let $\{\tilde{r}_N\}_{N \geq 1}$ and $\{\tilde{r}_N^a\}_{N \geq 1}$ be some sequences of positive constants converging to zero. The estimator $\tilde{\eta} = (\hat{\eta}^{(1)}, \tilde{\eta}^{(2)})$ is a random element in T with nuisance realization set $\widetilde{\mathcal{T}}_N \subseteq \widetilde{\mathcal{T}}_N^{(1)} \times \widetilde{\mathcal{T}}_N^{(2)}$ such that*

(i) Let $\eta^{(1)}$ is a fixed element in $\widetilde{\mathcal{T}}_N^{(1)}$ and define

$$\mathsf{F}_2(\eta^{(1)}) := \left\{ \psi\left(\cdot; \theta_0, (\eta^{(1)}, \eta^{(2)})\right) - \psi\left(\cdot; \theta_0, (\eta^{(1)}, \eta_0^{(2)})\right) \Big| (\eta^{(1)}, \eta^{(2)}) \in \widetilde{\mathcal{T}}_N \right\}.$$

Further let $F_2(\eta^{(1)})(\cdot)$ be a measurable envelope for $\mathsf{F}_2(\eta^{(1)})$ such that

$$\sup_{\eta^{(1)} \in \widetilde{\mathcal{T}}_N^{(1)}} \|F_2(\eta^{(1)})(W_i)\|_{P,q} \equiv \sup_{\eta^{(1)} \in \widetilde{\mathcal{T}}_N^{(1)}} \|F_2(\eta^{(1)})\|_{P,q} \leq V_n$$

[4]The pseudo-code for **Algorithm 4** is provided in the Online Appendix 3.6.4

Algorithm 5 (full-sample calibration) DML 2 Algorithm

1: **Input:** Data $(W_i)_{i=1}^N$. A K-fold random partition $(I_k)_{k=1}^K$ of $[N] = \{1, \dots, N\}$ such that each fold I_k is of size $n = N/K$. For each $k \in [K] = \{1, \dots, K\}$ define $I_k^c := \{1, \dots, N\} \setminus I_k$.

2: For each $k \in [K]$, fit a machine learning estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_0\big((W_i)_{i \in I_k^c}\big)$$

of η_0, where $\hat{\eta}_{0,k}$ is a random element in T, where the randomness only depends on the $(W_i)_{i \in I_k^c}$.

3: Combine all estimated nuisance elements $\hat{\eta}_{0,k}^{(2)}$ to fit a re-estimation procedure

$$\tilde{\eta}_0^{(2)} = \tilde{\eta}_0^{(2)}\big((W_i)_{i=1}^N, (\hat{\eta}_{0,k}^{(2)})_{k \in [K]}\big)$$

of $\eta_0^{(2)}$, where $\tilde{\eta}_0^{(2)}$ is a random element in $T^{(2)}$.

4: Construct the estimator $\tilde{\theta}_0$ as the solution to

$$\frac{1}{K} \sum_{k=1}^K \mathsf{E}_{n,k}\big[\psi\big(W; \tilde{\theta}_0, (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_0^{(2)})\big)\big] = 0,$$

where $\mathsf{E}_{n,k}[\psi(W)] = n^{-1} \sum_{i \in I_k} \psi(W_i)$ is the empirical expectation over $(W_i)_{i \in I_k}$.

for $q > 2$. For each $\eta^{(1)} \in \widetilde{\mathcal{T}}_N^{(1)}$ define σ_n^2 as a sequence converging to zero such that

$$\sup_{f \in \mathsf{F}_2(\eta^{(1)})} \mathsf{E}[f^2] \leq \sigma_n^2 \leq \|F_2(\eta^{(1)})\|_{P,2}^2$$

and u_n such that

$$\sup_{\eta^{(1)} \in \widetilde{\mathcal{T}}_N^{(1)}} J\left(\sigma_n / \|F_2(\eta^{(1)})\|_{P,2}, \mathsf{F}_2(\eta^{(1)}), F_2(\eta^{(1)})\right) \leq u_n.$$

Finally, assume the following growth condition is satisfied

$$\sup_{\eta^{(1)} \in \widetilde{\mathcal{T}}_N^{(1)}} \left(u_n \|F_2(\eta^{(1)})\|_{P,2} + \sigma_n \sqrt{\log(n)} \right.$$
$$\left. + n^{1/q - 1/2} V_n \left(u_n^2 \frac{\|F_2(\eta^{(1)})\|_{P,2}^2}{\sigma_n^2} \vee \log(n) \right) \right) \lesssim \tilde{r}_N$$

(ii) It holds

$$\sup_{\eta^{(1)} \in \widetilde{\mathcal{T}}_N^{(1)}} \left| \mathsf{E}\left[\psi^a\left(W; (\eta^{(1)}, \eta_0^{(2)})\right) - \psi^a\left(W; (\eta_0^{(1)}, \eta_0^{(2)})\right) \right] \right| \leq \tilde{r}_N^a$$

with probability converging to one. Further, the entropy conditions above also need to hold for ψ^a:
Let $\eta^{(1)}$ is a fixed element in $\widetilde{\mathcal{T}}_N^{(1)}$ and define

$$\mathsf{F}_2^a(\eta^{(1)}) := \left\{ \psi^a\left(\cdot; (\eta^{(1)}, \eta^{(2)})\right) - \psi^a\left(\cdot; (\eta^{(1)}, \eta_0^{(2)})\right) \,\middle|\, (\eta^{(1)}, \eta^{(2)}) \in \widetilde{\mathcal{T}}_N \right\}.$$

Further let $F_2^a(\eta^{(1)})(\cdot)$ be a measurable envelope for $\mathsf{F}_2^a(\eta^{(1)})$ such that

$$\sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \|F_2^a(\eta^{(1)})(W_i)\|_{P,q} \equiv \sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \|F_2^a(\eta^{(1)})\|_{P,q} \leq V_{n,a}$$

for $q > 2$. For each $\eta^{(1)} \in \widetilde{\mathcal{T}}_N^{(1)}$ define $\sigma_{n,a}^2$ as a sequence converging to zero such that

$$0 < \sup_{f\in\mathsf{F}_2^a(\eta^{(1)})} \mathsf{E}[f^2] \leq \sigma_{n,a}^2 \leq \|F_2^a(\eta^{(1)})\|_{P,2}^2$$

and $u_{n,a}$ such that

$$\sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} J\left(\sigma_{n,a}/\|F_2^a(\eta^{(1)})\|_{P,2}, \mathsf{F}_2^a(\eta^{(1)}), F_2^a(\eta^{(1)})\right) \leq u_{n,a}.$$

Finally, assume the following growth condition is satisfied

$$u_{n,a} \sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \|F_2^a(\eta^{(1)})\|_{P,2} + \sigma_{n,a}\sqrt{\log(n)}$$
$$+ n^{1/q-1/2} V_{n,a} \left(u_{n,a}^2 \frac{\sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \|F_2^a(\eta^{(1)})\|_{P,2}^2}{\sigma_{n,a}^2} \vee \log(n) \right) \lesssim \tilde{r}_N^a.$$

Assumption 12 imposes high-level assumptions on the complexity of the calibration step. Assumption 12 (i) restricts the complexity of the class $\mathsf{F}_2(\eta^{(1)})$ via standard complexity measures. If the function class is suitably measurable and the uniform entropy integral obeys $\log \sup_Q N\left(\epsilon\|F_2(\eta^{(1)})\|_{Q,2}, \mathsf{F}_2(\eta^{(1)}), \|\cdot\|_{Q,2}\right) \leq C$, it holds

$$J\left(\sigma_n/\|F_2(\eta^{(1)})\|_{P,2}, \mathsf{F}_2(\eta^{(1)}), F_2(\eta^{(1)})\right) \lesssim \sigma_n^{1/2}\|F_2(\eta^{(1)})\|_{P,2}^{-1/2},$$

since

$$J\left(\delta, \mathsf{F}_2(\eta^{(1)}), F_2(\eta^{(1)})\right) := \int_0^\delta \sup_Q \sqrt{1 + \log N(\epsilon\|\mathsf{F}_2(\eta^{(1)})\|_{Q,2}, \mathsf{F}_2(\eta^{(1)}), \|\cdot\|_{Q,2})}d\epsilon$$
$$\leq \int_0^\delta \sqrt{1 + C\epsilon^{-1}}d\epsilon \leq \delta + \sqrt{C}\int_0^\delta \epsilon^{-1/2}d\epsilon \lesssim \sqrt{\delta}$$

for any δ small enough and probability measure Q. The first part of Assumption 12 (ii) imposes a Lipschitz continuity condition on ψ^a which is the first part of the linear score defined in Equation (3.6). The second part of Assumption 12 (ii) provides similar complexity assumptions as for the function class $\mathsf{F}_2(\eta^{(1)})$ in Assumption 12 (i) and also the required growth rates. Remark that the conditions in Assumption 12 are quite similar to Belloni et al. (2018), but we build upon standard Donsker conditions. Again, it is worth noting that Assumption 12 (ii) depends only on the score ψ^a. In the case of a nonparametric causal model, also known as *interactive regression model* (IRM), considered in Section 3.3.2, the first part of the linear score is given by $\psi^a = -1$, see Equation (3.9), and therefore $\sigma_{n,a}^2 = 0$. Hence, in the interactive regression model, Theorem 1 below holds

with $\tilde{r}_N^a = 0$ and only Assumption 12 (i) is required.

Theorem 1. *Let $\tilde{\theta}_0$ be the estimator according to Algorithm 3 to 5. Assume $\delta_N \geq N^{-1/2}$ for all $N \geq 1$. Under Assumption 11 and 12, equation* (3.7) *in Remark 5 holds with updated remainder, that is*

$$\sqrt{N}\sigma^{-1}(\hat{\theta}_0 - \theta_0) = \frac{1}{N}\sum_{i=1}^{N} \bar{\psi}(W_i) + O_P(\tilde{\rho}_N) \rightsquigarrow \mathsf{N}(0,1)$$

uniformly over $P \in \mathcal{P}_N$, where the size of the remainder term obeys

$$\tilde{\rho}_N := \rho_N + \tilde{r}_N + \tilde{r}_N^a = N^{-1/2} + r_N + r'_N + \tilde{r}_N + \tilde{r}_N^a + N^{1/2}\lambda_N + N^{1/2}\lambda'_N \lesssim \delta_N.$$

The crucial difference between Algorithm 3 and 5 lies in the assumptions on the nuisance realization set $\tilde{\mathcal{T}}_N$, which require convergence rates of the recalibration procedure. In Algorithm 3, estimation properties are well known, e.g. for isotonic regression see Section 3.2.2. These proofs heavily rely on the i.i.d. assumption of the samples used for recalibration, which is violated for Algorithm 5. Nevertheless, cross-fitting might result in only weak dependencies between different samples, such that the convergence rates might still be sufficient for the calibration with Algorithm 5. This algorithm is closely related to the IC-IPW approach described by van der Laan et al. (2024a), in which calibration on the full sample is applied to the treated and control groups separately, while retaining the theoretical guarantees established in van der Laan et al. (2024a) (Theorems 1–2).

3.3.2 Calibration for Double Machine Learning Models

The results of Section 3.3 can be applied directly to different regression models and causal parameters of interest. In standard settings, a convergence rate of $\|\hat{m}_0 - m_0\|_{P,2} = o_P(N^{-1/4})$ is assumed. If isotonic regression is used for calibration, Lemma 3 directly implies that the convergence rate of the calibrated propensity score will still satisfy the rate condition $\|\tilde{m}_0 - m_0\|_{P,2} = o_P(N^{-1/4})$. Furthermore, we state explicit assumptions to restrict the complexity of the nuisance calibration to avoid additional sample splitting.

Consider the fully heterogeneous or interactive regression model (IRM) as in Chernozhukov et al. (2018). This nonparametric regression model is often considered when augmented inverse probability weighting (AIPW) is used for estimation. Let $D \in \{0,1\}$ be a binary treatment variable and $W = (Y, D, X)$, where

$$Y = g_0(D,X) + U,\ \mathsf{E}[U|D,X] = 0, \qquad D = m_0(X) + V,\ \mathsf{E}[V|X] = 0. \tag{3.8}$$

A common parameter of interest is the average treatment effect $\theta_0 := \mathsf{E}[g_0(1,X) - g_0(0,X)]$. Let the score function for the augmented inverse probability weighted estimator be

$$\psi(W;\theta,\eta) := (g(1,X) - g(0,X)) + \frac{D}{m(X)}(Y - g(1,X)) - \frac{1-D}{1-m(X)}(Y - g(0,X)) - \theta \tag{3.9}$$

where $\eta = (g,m)$ denotes the nuisance functions for the outcome regression $g_0(D,X) = E[Y|D,X]$ and propensity score $m_0(X) = E[D|X]$.

Assumption 13 (cf. Assumption 5.1 in Chernozhukov et al. (2018)). *Let* $\{\delta_N\}, \{\Delta_N\} \searrow 0$*;* $c, \epsilon, C > 0$*,* $q > 4$*,* $K \geq 2$ *fixed;* $N/K \in \mathbb{N}$*. For* $\eta = (\eta_1, \ldots, \eta_\ell)$*, define* $\|\eta\|_{p,q} := \max_{1 \leq j \leq \ell} \|\eta_j\|_{P,q}$*. For all* $P \in \mathcal{P}$*, the following hold: (a) Equations* (3.8) *are satisfied, (b)* $\|Y\|_{P,q} \leq C$*, (c)* $P(\epsilon \leq m_0(X) \leq 1 - \epsilon) = 1$*, (d)* $\|U\|_{P,2} \geq c$*, (e)* $\|\mathsf{E}_P[U^2|X]\|_{P,\infty} \leq C$ *and (f) for a random subset* $I \subset [N]$ *of size* $n = N/K$*, the nuisance parameter estimator* $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i \in I^c})$ *satisfies, with* P*-probability* $\geq 1 - \Delta_N$*:*

(i) $\|\hat{\eta}_0 - \eta_0\|_{P,2} \leq \delta_N$, $\|\hat{\eta}_0 - \eta_0\|_{P,q} \leq C$
(ii) $\|\hat{m}_0 - m_0\|_{P,2} \times \|\hat{g}_0 - g_0\|_{P,2} \leq \delta_N N^{-1/2}$ *with* $\|\hat{g}_0 - g_0\|_{P,\infty} \leq C$, $\|\hat{m}_0 - 1/2\|_{P,\infty} \leq 1/2 - \epsilon$

Under Assumption 13, Remark 5 holds (Chernozhukov et al. 2018, Theorem 5.1). Our assumption strengthens the original by requiring $q > 4$ rather than $q > 2$. For Theorem 1, we add:

Assumption 14 (Calibration rate/complexity). *The following assumptions hold:* (i) *With* P*-probability* $\geq 1 - \Delta_N$*, we have:* $\|\tilde{m}(X) - m_0(X)\|_{P,2} \lesssim \varepsilon_N \leq \log^{-1/2} N$*,* $\varepsilon_N(\cdot)\|\hat{g}_0 - g_0\|_{P,2} \leq \delta_N N^{-1/2}$*. Further, the predictions are well separated from zero and one,* $\|\tilde{m}(X) - 1/2\|_{P,\infty} \leq 1/2 - \epsilon$*.* (ii) *Let* $\tilde{m}(\cdot) \in \mathcal{M}$*, such that the covering numbers obey* $\sup_Q N(\epsilon, \mathcal{M}, L_2(Q)) \leq C\epsilon^{-1}$*.*

Assumption 14 imposes mild conditions on the calibration procedure. Assumption 14(i) ensures that the convergence rate of the calibrated propensity score $\tilde{m}(\cdot)$ is still sufficiently fast, while Assumption 14(ii) restricts the complexity of calibration, so that additional cross-fitting can be avoided.

Theorem 2. *Under Assumptions 13 and 14(i) Remark 6 is valid. If additionally Assumption 14(ii) is satisfied, Theorem 1 holds.*

Convergence rates of $\|\hat{m}_0 - m_0\|_{P,2} = o_P(N^{-1/4})$ and $\|\hat{g}_0 - g_0\|_{P,2} = o_P(N^{-1/4})$ imply the conditions of Assumption 13(f)(ii). Considering Lemma 3 this immediately implies Assumption 14 if isotonic regression is used for calibration as the convergence rate of the calibrated propensity score is given by $\|\tilde{m}_0 - m_0\|_{P,2} = o_P(N^{-1/4})$ and the complexity of monotone functions satisfies Assumption 14(ii). The proof of Theorem 2 is provided in Appendix 3.6.2. Isotonic calibration delivers two key improvements:

(i) **Fast Convergence and Complexity Control**: Under weak monotonicity (Assumption 10), isotonic regression achieves $\|\tilde{m} - m_0\|_{P,2} = O_P(N^{-1/3})$ (Lemma 3), satisfying $\|\tilde{m} - m_0\|_{P,2} \cdot \|\hat{g}_0 - g_0\|_{P,2} = o_P(N^{-1/2})$ when $\|\hat{g}_0 - g_0\|_{P,2} = o_P(N^{-1/4})$. The complexity condition (Assumption 14(ii)) holds via *bracketing entropy* $N(\epsilon, \mathcal{M}, L_2(P)) \leq C\epsilon^{-1}$ for monotone functions (van der Vaart & Wellner, 1996).

(ii) **Boundedness and Regularization:** By construction, isotonic calibration adjusts the estimated propensity scores toward values consistent with empirical treatment frequencies, thus improving probabilistic calibration. It enforces $\epsilon \leq \tilde{m}(X) \leq 1 - \epsilon$, which bounds the inverse propensity weights in $\psi(W; \theta, \eta)$. The resulting step-function form regularizes potentially overfitted ML estimates into piecewise constant regions, reducing variance while preserving the rank ordering of observations.

3.4 Simulation Study

In this section, we investigate the introduced calibrated propensity score models from Section 3.3.2 through an extensive simulation study[5]. We evaluate the impact of calibration methods (Venn-ABERS, Platt scaling, isotonic regression) on the performance of causal estimators (IPW, DML), supplemented by analyses of weight normalization and a comparison to covariate-balancing reweighting estimators (e.g., entropy balancing). Performance is assessed using calibration diagnostics (e.g., calibration plots, expected calibration error) and causal estimation metrics (RMSE, MAE, and variance) to unravel the interplay between robustness, forecast accuracy, and covariate balance.

To assess the contribution of potentially miss-calibrated propensity scores, we briefly introduce the causal estimators considered. The inverse probability weighting (IPW) estimator uses estimates of the propensity scores $\hat{m}(D = 1|X)$ directly. Here, an estimate $\hat{\theta}$ of the ATE is computed as $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D^{(i)}Y^{(i)}}{\hat{m}(D=1|X^{(i)})} - \frac{(1-D^{(i)})Y^{(i)}}{1-\hat{m}(D=1|X^{(i)})}\right)$. Especially treated units with low propensity scores and non-treated units with high propensity scores have extreme contributions. This can be critical if the underlying propensity score model is misspecified or overconfident.

The interactive regression model (IRM, Section 3.3.2) allows for heterogeneous treatment effects without strong form assumptions. Contrary, the partially linear regression model (PLR, Online Appendix 3.6.4) imposes an additive structure[6]. One obstacle in evaluating causal ATE models is that the true value of the causal parameter θ_0 is not observed in observational studies. For a fair evaluation, we have selected four external sources for the data generating processes (DGPs), each proposing different challenges for the models. An overview and detailed descriptions of the DGPs are provided in the Online Appendix 3.6.5. The DGPs are characterized by varying levels of noise (DGP 1, Belloni et al. (2017)), different dimensionality of observed covariates (DGP 1), and underlying nonlinearities (DGP 2, Deshpande and Kuleshov (2023); DGP 3, van der Laan et al. (2023)), overlap violations (DGP 2), or unbalancedness (DGP 4, Ballinari (2024), Nie and Wager (2020)). The DGPs satisfy the unconfoundedness assumption, $Y(d) \perp D \mid X$, with $Y(d)$ indicating the potential outcome under treatment $D = d$. Hence, these settings allow for identification of the average treatment effect, $\theta_0 = \mathbb{E}[Y(d = 1) - Y(d = 0)]$.

3.4.1 Learners and Calibration Methods

We test different learners for the outcome regression and the propensity score estimation. For the outcome regression, we consider a simple linear regression along with the tree-based Machine Learning algorithms LightGBM (LGBM) (Ke et al. 2017) and random forest. Both are flexible machine learning algorithms that perform well across a wide variety of datasets. For the propensity score estimation, we consider logistic regression, LGBM classifier, and random forest classifier. All models are employed within their default settings. Employing different fine-tuning schemes could benefit either approach and distort the comparison.

For propensity calibration, we consider the three approaches introduced in 3.2.2. First,

[5]The simulation is executed on an HPC cluster in parallel, using different seeds for the DGPs.

[6]Both, the IRM and PLR model, are implemented via the `DoubleML` package (Bach et al. 2022, 2024a).

we utilize `IsotonicRegression` from the `scikit-learn` package (Pedregosa et al. 2018). In addition, we employ the Inductive Venn–ABERS predictor (VAP) introduced by Vovk et al. (2015) and available at Petej (2024). VAP builds on the groupings in the outcome space made by isotonic regression. It utilizes potential labels to fit separate isotonic regressions. Thus, simple isotonic regression receives a coarser partitioning of the outcome space. Lastly, we incorporate Platt scaling, implemented via the `CalibratedClassifierCV` in the `scikit-learn` package.

3.4.2 Calibration Metrics

For binary classification, the ℓ_p Expected Calibration Error (ECE) (Naeini et al. 2015, Sun et al. 2024), for $p \geq 1$, is defined as: $\text{ECE}_p := \mathbb{E}\left[\mathbb{E}\left[\|D - m(X)\|^p \mid m(X)\right]\right]^{\frac{1}{p}}$. To approximate the expected calibration error (ECE), the estimated propensity scores $\hat{m}(X)$ are divided across the probabilistic output range $[0, 1]$ into equally spaced intervals (Naeini et al. 2015) $\{I_0, I_1, ..., I_M\}$ or quantiles (Nguyen and O'Connor 2015) of $\hat{m}(x)$. This allows us to generate buckets $\{B_i\}_{i=1}^{M}$, where $B_i = \{(X, D) \mid m(D = 1 \mid X) \in I_i\}$. Each predicted probability is assigned to the appropriate bin. The calibration error is then defined as the difference between the fraction of correct predictions (accuracy) and the mean predicted probability (confidence) within each bin: $\text{ECE}_p = \sum_{i=1}^{M} \frac{n_i}{N} \left\|\text{acc}_i(B_i) - \text{conf}_i(B_i)\right\|_p$, where $\text{acc}_i(B_i) = \frac{1}{|B_i|} \sum_{j=1}^{|B_i|} D_j$ and $\text{conf}_i(B_i) = \frac{1}{|B_i|} \sum_{j=1}^{|B_i|} m(D = 1 \mid X_j)$. In Figure 3.1, we can observe that the uncalibrated Algorithm 1, as well as the nested k-fold cross-fit Algorithm 2 are poorly calibrated for small sample sizes. Additionally, a version of Algorithm 1 clipped at the one percent level is included. This helps neglect some of the miscalibration, but still performs worse than Algorithms 4 and 5 for all sample sizes. In propensity weighting, severe deviations in the middle of the propensity score distribution are not particularly critical. However, deviations at the boundaries are crucial because they can lead to exploding weights. Therefore, it is generally advisable to include visualizations to assess both the overlap in propensity scores and their calibration properties. The overlap ratio plot, based on the reliability diagram, splits the probability space into equal parts. For each propensity bin, the plot displays the actual proportion of treated and untreated units separately. No clear violation of the overlap assumption can be seen for the true underlying propensities displayed in the left panel of Figure 3.2. The black dotted lines represent perfect calibration. The ratios illustrate the deviations for both the treated and untreated groups in the uncalibrated Algorithm 1 in the middle panel. Notably critical are the substantial proportions of treated observations with estimated propensity scores near zero and untreated observations with propensities close to one. In contrast, the perfect calibration property of Algorithm 5 is displayed in the right panel.

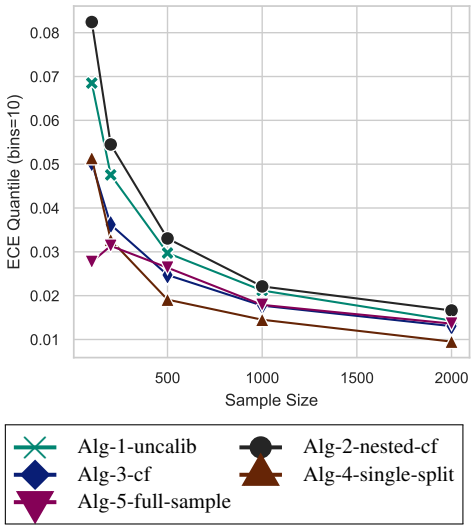

Figure 3.1: Quantile ECE, DGP 1, m = LGBM, n = 2000, p = 20

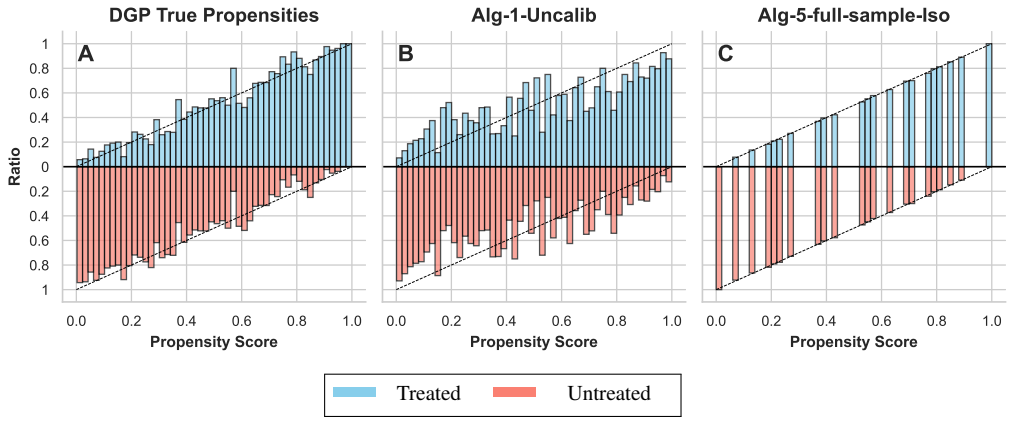

Figure 3.2: Overlap Ratios, DGP 1, n = 2000, p = 20, m = LGBM

3.4.3 General Findings

As expected, the nested cross-fitting Algorithm 2 faces stability challenges in small sample size settings. Venn-Abers calibration relies on isotonic regression combined with sample splitting. Consequently, Algorithm 3, when used with Venn-Abers, encountered similar instability. Additionally, the two-fold calibration Algorithm 4, combined with isotonic regression, required clipping at the 1-percent level to maintain stability. This instability arises from the known limitation of isotonic regression, which is prone to overfitting, especially with small calibration sets (van der Laan and Alaa 2024). To ensure a fair comparison, the uncalibrated Algorithm 1 is presented both unclipped and with a restriction at the 1-percent threshold (Alg-1-Clipped). Algorithms 3 and 5 were only clipped at a threshold of 10^{-12}. Given our sample sizes, any breach of this limit is effectively impossible. This serves more as a general recommendation, as the added clipping bias is negligible. Table 3.1 provides a summary of the results across all treatment models and the algorithms discussed, in combination with isotonic regression. Across all DGPs and settings, we can observe that calibration improves the inverse propensity-based IPW and IRM especially in combination with the tree-based propensity learners. The PLR model produces stable results across all settings, with minimal improvement from clipping or calibration for most DGPs. In general, the doubly-robust calibrated IRM and the PLR outperform the IPW. Algorithm 4 is biased in the PLR model for the tree-based methods

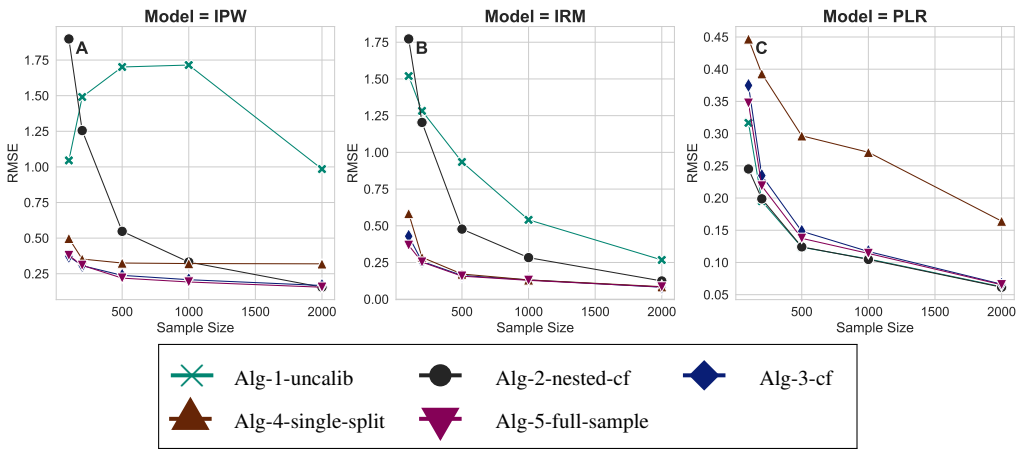

Figure 3.3: DGP 1, n =2000, p = 20, R2D = 0.5, m = LGBM, g = LGBM

random forest and LGBM in combination with VAP or isotonic regression[7]. This bias appears to persist regardless of the sample size, as demonstrated on the right-hand side of Figure 3.3. The impact of calibration is strongly dependent on the underlying propensity score learner.

Table 3.1: Results Overview

DGP	Model	Method	m = Logit			m = Random Forest			m = LGBM		
			MAE	RMSE	Std. dev.	MAE	RMSE	Std. dev.	MAE	RMSE	Std. dev.
1	IRM	Alg-1-Clipped	0.07	0.10	0.10	0.06	0.08	0.07	0.22	0.27	0.24
	IRM	Alg-1-Uncalib	0.08	0.13	0.13	1.85e+06	1.85e+07	1.84e+07	0.48	0.60	0.54
	IRM	Alg-2-nested-cf	0.12	0.15	0.14	0.10	0.12	0.12	0.10	0.12	0.12
	IRM	Alg-3-cf	0.06	0.08	0.07	0.06	0.08	0.08	0.06	0.08	0.07
	IRM	Alg-4-single-split	0.10	0.13	0.12	0.07	0.08	0.06	0.07	0.09	0.06
	IRM	Alg-5-full-sample	0.06	0.08	0.08	0.06	0.08	0.07	0.07	0.08	0.08
	IPW	Alg-1-Clipped	0.08	0.11	0.11	0.18	0.19	0.06	0.94	0.98	0.31
	IPW	Alg-1-Uncalib	0.10	0.16	0.16	1.09e+06	1.09e+07	1.09e+07	1.57	1.77	0.83
	IPW	Alg-5-full-sample	0.07	0.09	0.08	0.12	0.14	0.06	0.14	0.16	0.07
	PLR	Alg-1-Clipped	0.05	0.06	0.06	0.05	0.06	0.06	0.05	0.06	0.06
	PLR	Alg-1-Uncalib	0.05	0.06	0.06	0.05	0.06	0.06	0.05	0.06	0.06
	PLR	Alg-5-full-sample	0.05	0.06	0.06	0.05	0.06	0.06	0.05	0.07	0.07
2	IRM	Alg-1-Clipped	0.09	0.11	0.11	0.31	0.39	0.39	0.20	0.26	0.26
	IRM	Alg-1-Uncalib	0.09	0.11	0.11	2.22e+09	2.89e+09	2.86e+09	0.24	0.32	0.32
	IRM	Alg-2-nested-cf	0.18	0.23	0.22	0.16	0.22	0.22	0.19	0.24	0.24
	IRM	Alg-3-cf	0.09	0.11	0.11	0.09	0.12	0.12	0.10	0.12	0.12
	IRM	Alg-4-single-split	0.15	0.20	0.20	0.09	0.11	0.11	0.09	0.11	0.11
	IRM	Alg-5-full-sample	0.09	0.11	0.11	0.09	0.11	0.11	0.09	0.12	0.12
	IPW	Alg-1-Clipped	0.09	0.12	0.11	4.45	4.59	1.11	2.67	2.74	0.61
	IPW	Alg-1-Uncalib	0.09	0.12	0.11	1.17e+10	1.37e+10	7.23e+09	2.86	2.98	0.81
	IPW	Alg-5-full-sample	0.09	0.11	0.11	0.20	0.23	0.10	0.18	0.20	0.10
	PLR	Alg-1-Clipped	0.09	0.11	0.10	0.12	0.14	0.10	0.10	0.11	0.10
	PLR	Alg-1-Uncalib	0.09	0.11	0.10	0.12	0.14	0.10	0.10	0.11	0.10
	PLR	Alg-5-full-sample	0.09	0.11	0.10	0.09	0.11	0.10	0.09	0.10	0.10
3	IRM	Alg-1-Clipped	0.05	0.07	0.07	0.08	0.10	0.10	0.11	0.13	0.13
	IRM	Alg-1-Uncalib	0.05	0.07	0.07	9.97e+07	2.74e+08	2.73e+08	0.11	0.14	0.14
	IRM	Alg-2-nested-cf	0.10	0.13	0.12	0.08	0.10	0.10	0.10	0.13	0.13
	IRM	Alg-3-cf	0.05	0.07	0.07	0.06	0.07	0.07	0.05	0.07	0.06
	IRM	Alg-4-single-split	0.10	0.12	0.11	0.05	0.07	0.06	0.05	0.07	0.06
	IRM	Alg-5-full-sample	0.05	0.07	0.07	0.06	0.07	0.07	0.05	0.07	0.07
	IPW	Alg-1-Clipped	0.06	0.08	0.08	0.54	0.58	0.21	2.00	2.02	0.31
	IPW	Alg-1-Uncalib	0.06	0.08	0.08	3.07e+08	8.49e+08	7.91e+08	2.05	2.07	0.34
	IPW	Alg-5-full-sample	0.06	0.08	0.08	0.37	0.37	0.07	0.42	0.43	0.07
	PLR	Alg-1-Clipped	0.06	0.08	0.08	0.06	0.07	0.07	0.08	0.10	0.07
	PLR	Alg-1-Uncalib	0.06	0.08	0.08	0.06	0.07	0.07	0.08	0.10	0.07
	PLR	Alg-5-full-sample	0.07	0.09	0.08	0.06	0.07	0.07	0.06	0.07	0.07
4	IRM	Alg-1-Clipped	0.04	0.06	0.06	0.07	0.09	0.09	0.17	0.21	0.20
	IRM	Alg-1-Uncalib	0.04	0.06	0.06	1.74e+08	2.70e+08	2.68e+08	0.36	0.46	0.43
	IRM	Alg-2-nested-cf	0.05	0.07	0.07	0.05	0.07	0.07	0.06	0.07	0.07
	IRM	Alg-3-cf	0.04	0.06	0.06	0.05	0.06	0.06	0.05	0.06	0.06
	IRM	Alg-4-single-split	0.05	0.06	0.06	0.05	0.06	0.06	0.04	0.06	0.06
	IRM	Alg-5-full-sample	0.04	0.06	0.06	0.05	0.06	0.06	0.04	0.06	0.06
	IPW	Alg-1-Clipped	0.13	0.15	0.06	0.54	0.55	0.13	6.14	6.17	0.52
	IPW	Alg-1-Uncalib	0.13	0.15	0.06	4.64e+08	6.89e+08	5.11e+08	8.37	8.45	1.21
	IPW	Alg-5-full-sample	0.07	0.08	0.05	0.11	0.12	0.05	0.10	0.11	0.06
	PLR	Alg-1-Clipped	0.08	0.09	0.06	0.05	0.06	0.05	0.05	0.06	0.05
	PLR	Alg-1-Uncalib	0.08	0.09	0.06	0.05	0.06	0.05	0.05	0.06	0.05
	PLR	Alg-5-full-sample	0.08	0.09	0.06	0.07	0.09	0.05	0.08	0.09	0.05

For all DGPs: g = LGBM, and for Algorithms 2 - 5: Calibration = Isotonic Regression; DGP 1: n = 2000, p = 20, R2_d = 0.5; DGP 2: n = 2000, p = 3, overlap = 0.5; DGP 3: n = 2000, p = 4; DGP 4: n = 4000, p = 20, share treated = 0.1

In contrast to Błasiok et al. (2023), we generally observe good calibration properties for

[7]For more details on the influence of sample size on the proposed Algorithms, we refer to the Online Appendix 3.6.6, Figures 3.35, 3.37, 3.39, 3.41.

logistic regression and the least improvements through calibration. Random forest can be both under-confident, as seen in Figure 3.4 for DGP 1, and over-confident for DGPs 2 and 3[8].

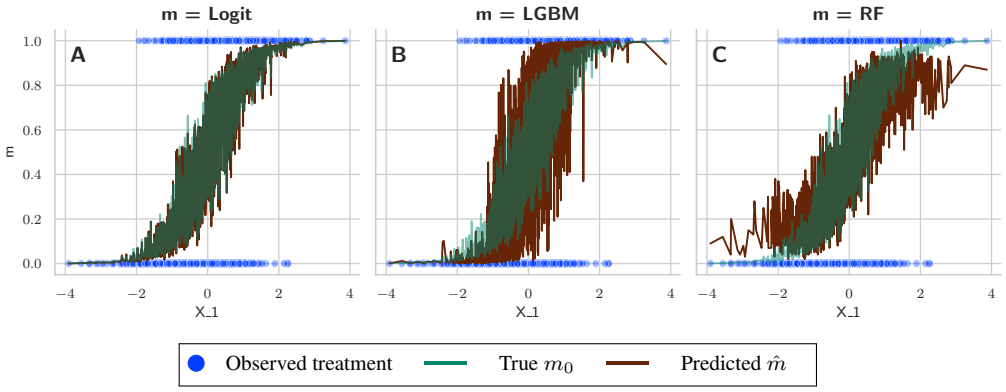

Figure 3.4: DGP 1, n =2000, p = 20, R2D = 0.5

As shown by Johansson et al. (2023), random forest tends to be under-confident for the minority class in unbalanced settings (DGP 4, Figure 3.8 in the Online Appendix 1). In general, the combination of random forest with calibration performs well across various settings. On the other hand, boosting-based methods, such as LGBM, tend to be over-confident. Figure 3.5 displays the distribution of ATE estimates for different propensity score learners under the IRM across 100 repetitions with different seeds in DGP 1. Calibration can correct for both under-confident and over-confident learners. However, the impact appears strongest for over-confident learners. For more details, we refer to the Online Appendix 1, where the robustness of our algorithms is tested with respect to the propensity and outcome learners, different clipping thresholds, and various levels of signal-to-noise ratio (DGP 1), overlap (DGP 2), and share of treated units (DGP 4).

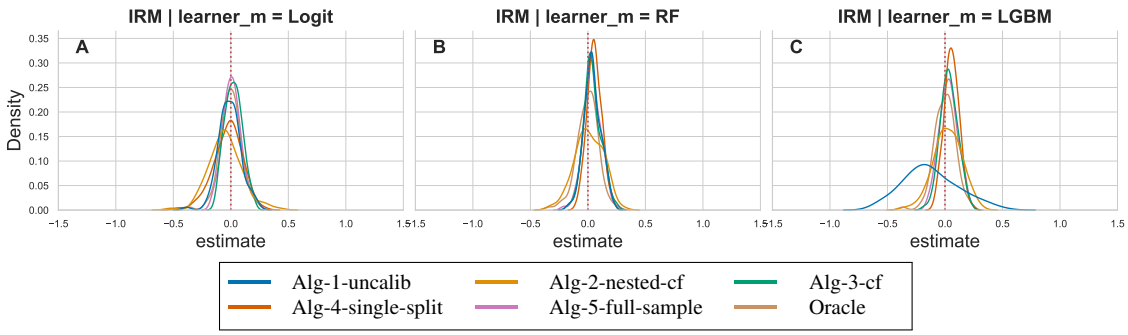

Figure 3.5: DGP 1, n =2000, p = 20, R2D = 0.5, m = LGBM, g = LGBM

3.4.4 Desirable Properties

Modern causal inference methods using weighting estimators address two core challenges: (1) achieving covariate balance, where the weighted covariate distributions satisfy $\mathbb{E}[w\mathbf{X} \mid D = 1] = \mathbb{E}[w\mathbf{X} \mid D = 0]$ for treatment $D \in \{0, 1\}$ and covariates $\mathbf{X}$, and (2) ensuring normalization $\sum_{i:D_i=1} w_i = 1$ and $\sum_{i:D_i=0} w_i = 1$ to stabilize weights (Busso et al. 2014).

[8]The corresponding figures are located in the Online Appendix 1, Figures 3.8 and 3.9.

Covariate balance is quantified via standardized mean differences (SMD):

$$\text{SMD}_k = \frac{\bar{X}^w_{k,D=1} - \bar{X}^w_{k,D=0}}{\sqrt{(s^w_{k,D=1})^2 + (s^w_{k,D=0})^2/2}},$$

where $\bar{X}^w_{k,D}$ and $s^w_{k,D}$, with $D \in \{0, 1\}$, are the weighted means and standard deviations of X_k in the treated and control groups, respectively. The SMD expresses an imbalance in standard deviation units, with $|\text{SMD}_k| < 0.1$ indicating an adequate balance (Austin and Stuart 2015). Figure 3.6 demonstrates SMD reduction across methods. Entropy balancing (EBAL) minimizes KL/ Rényi divergence from uniform weights $q_i = 1/n$ under balance constraints $\sum_{i:D_i=0} w_i \phi(\mathbf{X}_i) = \frac{1}{n_1} \sum_{i:D_i=1} \phi(\mathbf{X}_i)$ (Hainmueller 2012). OptimWeight solves $\min_w \sum_{i:D_i=0}(w_i - 1/n_0)^2$ with ℓ_∞-norm balance constraints $\|\frac{1}{n_1} \sum_{i:D_i=1} \mathbf{X}_i - \sum_{i:D_i=0} w_i \mathbf{X}_i\|_\infty \leq \delta$ (Zubizarreta 2015). Covariate balancing propensity score (CBPS) estimates the propensity score $m(\mathbf{X}_i; \beta) = \text{expit}(\mathbf{X}_i^T \beta)$ via GMM, combining score equations $\sum_{i=1}^n [D_i - m(\mathbf{X}_i; \beta)]\mathbf{X}_i = 0$ with balancing moments Imai and Ratkovic (2014). Inverse probability tilting (IPT) solves dual-moment conditions for normalization $\sum D_i / m(\mathbf{X}_i; \theta) = n$ and balance $\sum D_i \mathbf{X}_i / m(\mathbf{X}_i; \theta) = \sum (1 - D_i)\mathbf{X}_i / (1 - m(\mathbf{X}_i; \theta))$ (Graham et al. 2012). GLM uses $m(\mathbf{X}_i) = \text{expit}(\mathbf{X}_i^T \hat{\beta}_{\text{MLE}})$ with normalized IPW weights. To ensure a fair comparison, Alg-3-cf and Alg-5-full-sample used a linear outcome function without further fine-tuning, as excessive fine-tuning could bias the comparison. This means that the outcome equation is misspecified for all methods. Figure 3.6 shows that methods from the R package

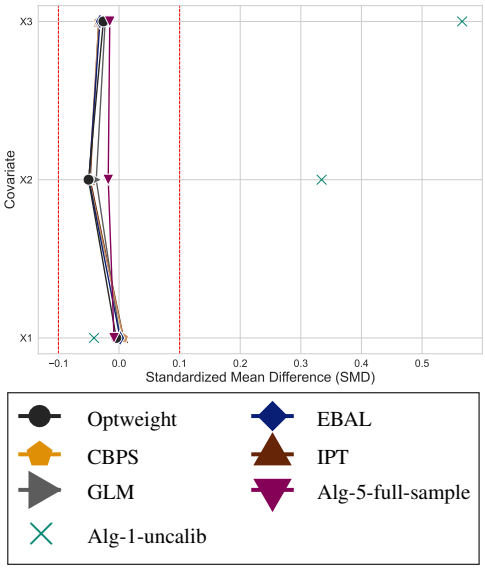

Figure 3.6: SMD across covariates for DGP 2. Dashed line indicates |SMD| = 0.1.

Table 3.2: Comparison under covariate balance

Method		m	Coverage	CI Length	Norm $D=1$	Norm $D=0$	RMSE	Std. dev.	MAE
Alg-3-cf	isotonic	Logit	0.960	0.413	0.996	0.987	0.098	0.097	0.077
Alg-3-cf	isotonic	RF	0.960	0.389	0.997	0.994	0.096	0.094	0.075
Alg-3-cf	platt	Logit	0.960	0.388	1.003	0.984	0.094	0.093	0.075
Alg-3-cf	platt	RF	0.940	0.358	0.984	0.961	0.092	0.091	0.072
Alg-5-full-sample	isotonic	Logit	0.970	0.404	0.999	0.998	0.097	0.095	0.077
Alg-5-full-sample	isotonic	RF	0.940	0.386	1.000	1.000	0.095	0.094	0.075
Alg-5-full-sample	platt	Logit	0.960	0.397	1.003	0.996	0.094	0.093	0.075
Alg-5-full-sample	platt	RF	0.950	0.373	0.998	0.988	0.094	0.093	0.074
Cbps	weighted	Logit	0.950	0.371	1.000	1.000	0.094	0.093	0.074
Ebal	weighted	-	0.950	0.376	1.000	1.000	0.094	0.093	0.074
Glm	weighted	Logit	0.950	0.371	1.000	1.000	0.094	0.093	0.075
Ipt	weighted	Logit	0.950	0.371	1.000	1.000	0.094	0.093	0.075
Optweight	weighted	-	0.960	0.376	1.000	1.000	0.093	0.092	0.074

DGP 2: n = 2000, p = 3, overlap = 0.5, Clip = 1e-12 g = Linear

`weightit` (Greifer 2025) (EBAL, OptWeight, CBPS, GLM, IPT) and the full-sample calibrated IRM achieve covariate balance, unlike the unadjasted IRM. Table 3.2 compares

performance metrics and normalization. All `weightit` methods enforce exact normalization via $\mathbb{E}[D/m(\mathbf{X})] = 1$ and $\mathbb{E}[(1-D)/(1-m(\mathbf{X}))] = 1$ (columns 'Norm D = 1' and 'Norm D = 0'), ensuring weight stability. Full-sample calibration achieves higher levels of normalization compared to their cross-fitted counterparts. However, the remaining small deviations from optimally normalized weights do not deteriorate performance.

3.5 Discussion

We extended and adapted some of the simulation settings to gain additional insights from the studies implemented by Deshpande and Kuleshov (2023), Gutman et al. (2024), Ballinari (2024), and van der Laan et al. (2024a). In our simulation study, our objective was to determine whether calibration works and to explore the best methods to achieve effective calibration. Deshpande and Kuleshov (2023) achieve significant improvements with their non-cross-fitted version of Algorithm 4, even in logistic regression. This result seems counter-intuitive. However, given their deterministic propensity scores in the drug effectiveness DGP, it is unsurprising that logistic regression is miscalibrated. In our adapted version, DGP 2, with nondeterministic propensity scores, such large improvements are not observed. In particular, Algorithm 4 without clipping, or in combination with PLR, is not recommended based on our findings.

Ballinari (2024) implements the nested cross-fitting Algorithm 2. As demonstrated, nested cross-fitting combined with isotonic regression is effective only when clipping is applied. Generally, Algorithm 2 shows poor calibration properties and is not recommended for small sample sizes, where calibration has the most significant impact. The unbalanced and nonlinear settings labeled "difficult" and "extreme" for DGPs 4 and 5 by Ballinari (2024) were tested under DGP 4 in this study. We show that either using a larger share of observations for calibration or adding 1-percent clipping allows isotonic regression to lower RMSEs for boosting-based methods and remain stable for other learners. The instability observed in the results of Ballinari (2024) was also addressed by van der Laan et al. (2024a). The authors concluded that, in a discretized version, it is important to calibrate the treated and untreated observations separately. However, as we demonstrate, the instability of isotonic regression is more likely due to small sample size issues. Our calibration algorithms, 3 and 5, are both stable without clipping or separate calibration for treated and untreated units. The latter follows van der Laan et al. (2024a)'s recommendation to calibrate on the full sample using cross-fitted propensity scores. In summary, our findings emphasize the critical role of method selection and sample size in calibration procedures. While (nested) cross-fitting for the calibration step is not always necessary, it may require supplemental clipping in settings prone to overfitting. Crucially, propensity score calibration enhances the robustness of inverse propensity-weighted ATE estimates.

3.6 Appendix

3.6.1 Proof of Lemma 2

In this appendix we prove the following Lemma from Section 3.2.

It holds

$$\begin{aligned}
&\|\hat{m}(X) - m_0(X)\|_{P,2}^2 \\
&= \mathbb{E}\left[\left(m_0(X) - \mathbb{E}[m_0(X)|\hat{m}(X)] + \mathbb{E}[m_0(X)|\hat{m}(X)] - \hat{m}(X)\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(m_0(X) - \mathbb{E}[m_0(X)|\hat{m}(X)] + \mathbb{E}[m_0(X)|\hat{m}(X)] - \hat{m}(X)\right)^2|\hat{m}(X)\right]\right] \\
&= \mathbb{E}\left[\mathrm{Var}\big(m_0(X)|\hat{m}(X)\big)\right] + \|\mathbb{E}[m_0(X)|\hat{m}(X)] - \hat{m}(X)\|_{P,2}^2 \\
&\quad + 2\mathbb{E}\left[\big(\mathbb{E}[m_0(X)|\hat{m}(X)] - \hat{m}(X)\big)\underbrace{\mathbb{E}\left[\big(m_0(X) - \mathbb{E}[m_0(X)|\hat{m}(X)]\big)|\hat{m}(X)\right]}_{=0}\right]
\end{aligned}$$

such that

$$\|\hat{m}(X) - m_0(X)\|_{P,2} = \left(\mathbb{E}\left[\mathrm{Var}\big(m_0(X)|\hat{m}(X)\big)\right] + \|\mathbb{E}[m_0(X)|\hat{m}(X)] - \hat{m}(X)\|_{P,2}^2\right)^{1/2}.$$

Proof of Lemma 1.

Under Assumption 1 we can decompose the root-mean-squared-error as follows

$$\begin{aligned}
\|\tilde{m}(X) - m_0(X)\|_{P,2} &\leq \underbrace{\|\tilde{m}(X) - \mathbb{E}[D|\hat{m}(X)]\|_{P,2}}_{\leq \tilde{\varepsilon}_N} + \|\mathbb{E}[D|\hat{m}(X)] - m_0(X)\|_{P,2} \\
&\leq \left(\mathbb{E}\left[\mathsf{V}\big(m_0(X)|\hat{m}(X)\big)\right]\right)^{1/2} + \tilde{\varepsilon}_N
\end{aligned}$$

due to

$$\begin{aligned}
\|\mathbb{E}[D|\hat{m}(X)] - m_0(X)\|_{P,2}^2 &= \|\mathbb{E}[m_0(X)|\hat{m}(X)] - m_0(X)\|_{P,2}^2 \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(m_0(X) - \mathbb{E}[m_0(X)|\hat{m}(X)]\right)^2\right]|\hat{m}(X)\right] \\
&= \mathbb{E}\left[\mathsf{V}\big(m_0(X)|\hat{m}(X)\big)\right].
\end{aligned}$$

Here, we used that

$$\sigma(\hat{m}(X)) \subseteq \sigma(X),$$

such that

$$\mathbb{E}[m_0(X)|\hat{m}(X)] = \mathbb{E}[\mathbb{E}[D|X]|\hat{m}(X)] = \mathbb{E}[D|\hat{m}(X)].$$

□

3.6.2 Proof of the Theorems

In this appendix we prove the following Theorems from Section 3.3.

For any $k \in \{1, \ldots, K\}$, we use the following empirical process notation

$$\mathsf{G}_{n,k}\left[\phi\left(W\right)\right] = \frac{1}{\sqrt{n}} \sum_{i \in I_k} \left(\phi(W_i) - \int \phi(w) dP_N \right)$$

for any P_N-integrable function ϕ on $\mathcal{W}$.

Proof of Theorem 1.

This proof adjusts the proof of Theorem 3.1 in Chernozhukov et al. (2018). Upon inspecting the Step 1 (DML2 case) of the proof of Theorem 3.1 in Chernozhukov et al. (2018), it is worth noting that the only terms which are affected by the calibration are

$$\begin{aligned}
\tilde{R}_{N,1} :&= \frac{1}{K} \sum_{k=1}^{K} \mathsf{E}_{n,k}\left[\psi^a\left(W; (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)})\right)\right] - \mathsf{E}\left[\psi^a\left(W_i; (\eta_{0,k}^{(1)}, \eta_{0,k}^{(2)})\right)\right] \\
\tilde{R}_{N,2} :&= \frac{1}{K} \sum_{k=1}^{K} \mathsf{E}_{n,k}\left[\psi\left(W; \theta_0, (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)})\right)\right] - \frac{1}{N} \sum_{i=1}^{N} \psi\left(W_i; \theta_0, (\eta_{0,k}^{(1)}, \eta_{0,k}^{(2)})\right)
\end{aligned}$$

as the other terms do not depend on the calibrated element $\tilde{\eta}_{0,k}^{(2)}$. First, we focus on the term $\tilde{R}_{N,2}$. Following Step 3 of the proof of Theorem 3.1, we obtain by triangle inequality

$$\left| \mathsf{E}_{n,k}\left[\psi\left(W; \theta_0, (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)})\right)\right] - \frac{1}{n} \sum_{i \in I_k} \psi\left(W_i; \theta_0, (\eta_{0,k}^{(1)}, \eta_{0,k}^{(2)})\right) \right| \leq \frac{\mathcal{I}_{3,k} + \mathcal{I}_{4,k}}{\sqrt{n}},$$

where

$$\begin{aligned}
\mathcal{I}_{3,k} &:= \left| \mathsf{G}_{n,k}\left[\psi\left(W; \theta_0, (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)})\right)\right] - \mathsf{G}_{n,k}\left[\psi\left(W; \theta_0, (\eta_{0,k}^{(1)}, \eta_{0,k}^{(2)})\right)\right] \right| \\
\mathcal{I}_{4,k} &:= \sqrt{n} \left| \mathsf{E}\left[\psi\left(W; \theta_0, (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)})\right) | (W_i)_{i \in I_k^c}\right] - \mathsf{E}\left[\psi\left(W; \theta_0, (\eta_{0,k}^{(1)}, \eta_{0,k}^{(2)})\right)\right] \right|.
\end{aligned}$$

Due to Assumption 11 it holds $(\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)}) \in \widetilde{\mathcal{T}}_N$ with probability $\geq 1 - \Delta_N$. Define $\mathcal{E}_N$ as the event that $(\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)}) \in \widetilde{\mathcal{T}}_N$ for all $k \in [K]$. Therefore, $P(\mathcal{E}_N) \geq 1 - K\Delta_N$. Consequently, the whole argument of Chernozhukov et al. (2018) is still valid which implies

$$\mathcal{I}_{4,k} = O_{P_N}(\sqrt{n}(\lambda_N + \lambda_N')).$$

Next, due to the triangle inequality

$$\begin{aligned}
\mathcal{I}_{3,k} \leq& \left| \mathsf{G}_{n,k}\left[\psi\left(W; \theta_0, (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)})\right)\right] - \mathsf{G}_{n,k}\left[\psi\left(W; \theta_0, (\hat{\eta}_{0,k}^{(1)}, \eta_{0,k}^{(2)})\right)\right] \right| \\
&+ \left| \mathsf{G}_{n,k}\left[\psi\left(W; \theta_0, (\hat{\eta}_{0,k}^{(1)}, \eta_{0,k}^{(2)})\right)\right] - \mathsf{G}_{n,k}\left[\psi\left(W; \theta_0, (\eta_{0,k}^{(1)}, \eta_{0,k}^{(2)})\right)\right] \right| \\
=:& \ \mathcal{I}_{3,k}^{(1)} + \mathcal{I}_{3,k}^{(2)},
\end{aligned}$$

where $\mathcal{I}_{3,k}^{(2)} = O_{P_N}(r'_N)$ by Assumption 11.2, see Chernozhukov et al. (2018). To bound $\mathcal{I}_{3,k}^{(1)}$, we rely on classical empirical process theory. Since conditionally on $(W_i)_{i\in I_k^c}$ the $\hat{\eta}_{0,k}^{(1)}$ is non-stochastic, remark that

$$
\begin{aligned}
&\mathsf{E}\left[\mathcal{I}_{3,k}^{(1)} | (W_i)_{i\in I_k^c}\right] \\
=&\mathsf{E}\left[\left|\mathsf{G}_{n,k}\left[\psi\left(W;\theta_0,(\hat{\eta}_{0,k}^{(1)},\tilde{\eta}_{0,k}^{(2)})\right)\right] - \mathsf{G}_{n,k}\left[\psi\left(W;\theta_0,(\hat{\eta}_{0,k}^{(1)},\eta_{0,k}^{(2)})\right)\right]\right| \middle| (W_i)_{i\in I_k^c}\right] \\
\leq& \sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \mathsf{E}\Bigg[\sup_{\eta^{(2)}:(\eta^{(1)},\eta^{(2)})\in\widetilde{\mathcal{T}}_N} \Big|\mathsf{G}_{n,k}\left[\psi\left(W;\theta_0,(\eta^{(1)},\eta^{(2)})\right)\right] \\
&\quad - \mathsf{G}_{n,k}\left[\psi\left(W;\theta_0,(\eta^{(1)},\eta_{0,k}^{(2)})\right)\right]\Big| \Big| (W_i)_{i\in I_k^c}\Bigg] \\
\leq& \sup_{\eta^{(1)}\in\tilde{\mathcal{T}_N}^{(1)}} \mathsf{E}\Big[\sup_{f\in\mathsf{F}_2(\eta^{(1)})} |\mathsf{G}_{n,k}(f)| \Big| (W_i)_{i\in I_k^c}\Big] \\
\leq& \sup_{\eta^{(1)}\in\tilde{\mathcal{T}_N}^{(1)}} \mathsf{E}[\sup_{f\in\mathsf{F}_2(\eta^{(1)})} |\mathsf{G}_{n,k}(f)|] \\
=& \sup_{\eta^{(1)}\in\tilde{\mathcal{T}_N}^{(1)}} \|\mathsf{G}_{n,k}\|_{\mathsf{F}_2(\eta^{(1)})}
\end{aligned}
$$

where $\mathcal{F}_2$ is defined in Assumption 12. Let $\eta^{(1)}$ be a any element of $\widetilde{\mathcal{T}}_N^{(1)}$. Relying on Theorem 5.2 of Chernozhukov et al. (2014), it holds

$$
\begin{aligned}
\mathsf{E}\left[\|\mathsf{G}_{n,k}\|_{\mathsf{F}_2(\eta^{(1)})}\right] \lesssim\ & J\left(\delta_n, \mathsf{F}_2(\eta^{(1)}), F_2(\eta^{(1)})\right) \|F_2(\eta^{(1)})\|_{P,2} \\
&+ \frac{\|M\|_{P,2} J^2\left(\delta_n, \mathsf{F}_2(\eta^{(1)}), F_2(\eta^{(1)})\right)}{\delta_n^2\sqrt{n}}
\end{aligned}
$$

with $\delta_n = \sigma_n/\|F_2(\eta^{(1)})\|_{P,2}$, $M = \max_{1\leq i\leq n} F_2(\eta^{(1)})(W_i)$, where $\sup_{f\in\mathsf{F}_2(\eta^{(1)})} \mathsf{E}[f^2] \leq \sigma_n^2 \leq \|F_2(\eta^{(1)})\|_{P,2}^2$. Under Assumption 12(i) and using $\|M\|_{P,q} \leq n^{1/q}\|F_2(\eta^{(1)})\|_{P,q}$ this implies

$$
\mathsf{E}\left[\|\mathsf{G}_{n,k}\|_{\mathsf{F}_2(\eta^{(1)})}\right] \lesssim u_n \|F_2(\eta^{(1)})\|_{P,2} + n^{1/q-1/2} V_n u_n^2 \frac{\|F_2(\eta^{(1)})\|_{P,2}^2}{\sigma_n^2}.
$$

This enables the use of Theorem 5.1 of Chernozhukov et al. (2014) with $t = \log(n)$ such

that for any fixed $\eta^{(1)} \in \widetilde{\mathcal{T}}_N^{(1)}$

$$\begin{aligned}
\|\mathsf{G}_{n,k}\|_{\mathsf{F}_2(\eta^{(1)})} \leq\ & (1+\alpha)\mathsf{E}\left[\|\mathsf{G}_{n,k}\|_{\mathsf{F}_2(\eta^{(1)})}\right] \\
& + C(q)\left[\left(\sigma_n + \frac{\|M\|_{P,q}}{\sqrt{n}}\right)\sqrt{\log(n)} + \frac{\|M\|_{P,2}\log(n)}{\alpha\sqrt{n}}\right] \\
\lesssim\ & (1+\alpha)\left(u_n\|F_2(\eta^{(1)})\|_{P,2} + n^{1/q-1/2}V_n u_n^2\frac{\|F_2(\eta^{(1)})\|_{P,2}^2}{\sigma_n^2}\right) \\
& + C(q)\left((\sigma_n + n^{1/q-1/2}V_n)\sqrt{\log(n)} + n^{1/q-1/2}V_n\frac{\log(n)}{\alpha}\right)
\end{aligned}$$

with probability $> 1-\log(n)^{-q/2}$ for all $\alpha > 0$ and $C(q) > 0$ is a constant only depending on q. Consequently

$$\begin{aligned}
\|\mathsf{G}_{n,k}\|_{\mathsf{F}_2(\eta^{(1)})} &\lesssim u_n\|F_2(\eta^{(1)})\|_{P,2} + \sigma_n\sqrt{\log(n)} \\
&\quad + n^{1/q-1/2}V_n\left(u_n^2\frac{\|F_2(\eta^{(1)})\|_{P,2}^2}{\sigma_n^2} \vee \log(n)\right) \\
&\leq \tilde{r}_N
\end{aligned}$$

with probability $> 1 - c\log(n)^{-1}$, where $\tilde{r}_N$ does not depend on $\eta^{(1)}$ due to growth condition 12 (iii). Hence, by Lemma 6.1 in Chernozhukov et al. (2018)

$$\mathcal{I}_{3,k}^{(1)} = O_{P_N}(\tilde{r}_N).$$

This implies that

$$|\tilde{R}_{N,2}| = O_{P_N}(N^{-1/2}(r'_N + \tilde{r}_N) + \lambda_N + \lambda'_N)$$

Next, we show $|\tilde{R}_{N,1}| = |\hat{J}_0 - J_0| = O_{P_N}\left(N^{-1/2} + r_N + \tilde{r}_N^a\right)$. As in Chernozhukov et al. (2018) is suffices to show that for any $k \in [K]$,

$$\begin{aligned}
&\left|\mathsf{E}_{n,k}\left[\psi^a\left(W;(\hat{\eta}_{0,k}^{(1)},\tilde{\eta}_{0,k}^{(2)})\right)\right] - \mathsf{E}\left[\psi^a\left(W;(\eta_0^{(1)},\eta_{0,k}^{(2)})\right)\right]\right| \\
=&O_{P_N}(N^{-1/2} + r_N + \tilde{r}_N).
\end{aligned}$$

It holds

$$\begin{aligned}
&\left|\mathsf{E}_{n,k}\left[\psi^a\left(W;(\hat{\eta}_{0,k}^{(1)},\tilde{\eta}_{0,k}^{(2)})\right)\right] - \mathsf{E}\left[\psi^a\left(W;(\eta_0^{(1)},\eta_{0,k}^{(2)})\right)\right]\right| \\
\leq&\left|\mathsf{E}_{n,k}\left[\psi^a\left(W;(\hat{\eta}_{0,k}^{(1)},\tilde{\eta}_{0,k}^{(2)})\right)\right] - \mathsf{E}_{n,k}\left[\psi^a\left(W;(\hat{\eta}_{0,k}^{(1)},\eta_0^{(2)})\right)\right]\right| \\
&+\left|\mathsf{E}_{n,k}\left[\psi^a\left(W;(\hat{\eta}_{0,k}^{(1)},\eta_0^{(2)})\right)\right] - \mathsf{E}\left[\psi^a\left(W;(\hat{\eta}_{0,k}^{(1)},\eta_0^{(2)})\right)|(W_i)_{i\in I_k^c}\right]\right| \\
&+\left|\mathsf{E}\left[\psi^a\left(W;(\hat{\eta}_{0,k}^{(1)},\eta_0^{(2)})\right)|(W_i)_{i\in I_k^c}\right] - \mathsf{E}\left[\psi^a\left(W;(\eta_0^{(1)},\eta_{0,k}^{(2)})\right)\right]\right| \\
=&\mathcal{I}_{1,k}^{(1)} + \mathcal{I}_{1,k}^{(2)} + \mathcal{I}_{2,k},
\end{aligned}$$

where $\mathcal{I}_{1,k}^{(2)} = O_{P_N}(N^{-1/2})$ and $\mathcal{I}_{2,k} = O_{P_N}(r_N)$ by the same terms as in Step 2 of the proof of Theorem 3.1 in Chernozhukov et al. (2018). Next, we show $\mathcal{I}_{1,k}^{(1)} = O_{P_N}(r_N + \tilde{r}_N)$.

$$\begin{aligned}
\mathcal{I}_{1,k}^{(1)} &= \left| \mathsf{E}_{n,k} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)}) \right) \right] - \mathsf{E}_{n,k} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \eta_0^{(2)}) \right) \right] \right| \\
&\leq \left| \mathsf{G}_{n,k} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)}) \right) \right] - \mathsf{G}_{n,k} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \eta_0^{(2)}) \right) \right] \right| \\
&\quad + \left| \mathsf{E} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)}) \right) \right] - \mathsf{E} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \eta_0^{(2)}) \right) \right] \right|
\end{aligned}$$

At first, remark

$$\begin{aligned}
&\left| \mathsf{E} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)}) \right) \right] - \mathsf{E} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \eta_0^{(2)}) \right) \right] \right| \\
\leq &\left| \mathsf{E} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)}) \right) \right] - \mathsf{E} \left[\psi^a \left(W; (\eta_0^{(1)}, \eta_0^{(2)}) \right) \right] \right| \\
&+ \left| \mathsf{E} \left[\psi^a \left(W; (\eta_0^{(1)}, \eta_0^{(2)}) \right) \right] - \mathsf{E} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \eta_0^{(2)}) \right) \right] \right| \\
&\lesssim r_N + \tilde{r}_N^a
\end{aligned}$$

by Assumption 12 (ii). Further, with the same argument as above

$$\begin{aligned}
&\mathsf{E} \left[\left| \mathsf{G}_{n,k} \left[\psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)}) \right) - \psi^a \left(W; (\hat{\eta}_{0,k}^{(1)}, \eta_0^{(2)}) \right) \right] \right| \middle| (W_i)_{i \in I_k^c} \right] \\
\leq &\sup_{\eta^{(1)} \in \tilde{\mathcal{T}_N}^{(1)}} \|\mathsf{G}_{n,k}\|_{\mathsf{F}_2^a(\eta^{(1)})}
\end{aligned}$$

with

$$\mathsf{F}_2^a(\eta^{(1)}) := \left\{ \psi^a \left(\cdot; (\eta^{(1)}, \eta^{(2)}) \right) - \psi^a \left(\cdot; (\eta^{(1)}, \eta_0^{(2)}) \right) \middle| (\eta^{(1)}, \eta^{(2)}) \in \widetilde{\mathcal{T}}_N \right\}.$$

Following the same arguments as above, combined with Assumption 12 (ii), we obtain

$$\begin{aligned}
\|\mathsf{G}_{n,k}\|_{\mathsf{F}_2^a(\eta^{(1)})} &\lesssim u_{n,a} \|F_2^a(\eta^{(1)})\|_{P,2} + \sigma_{n,a} \sqrt{\log(n)} \\
&\quad + n^{1/q - 1/2} V_{n,a} \left(u_{n,a}^2 \frac{\|F_2^a(\eta^{(1)})\|_{P,2}^2}{\sigma_{n,a}^2} \vee \log(n) \right) \\
&\leq \tilde{r}_N^a.
\end{aligned}$$

Again, by Lemma 6.1 in Chernozhukov et al. (2018)

$$\mathcal{I}_{1,k}^{(1)} = O_{P_N}(r_N + \tilde{r}_N^a)$$

and therefore

$$|\tilde{R}_{N,1}| = O_{P_N} \left(N^{-1/2} + r_N + \tilde{r}_N^a \right).$$

□

Proof of Theorem 2.

The first part is an application of Theorem 5.1 Chernozhukov et al. (2018) with nuisance realization set $\widetilde{\mathcal{T}}_N$.
The second part follows by verifying Assumption 12. At first remark that Assumption 12(ii) holds since $\psi^a(W;\eta) = -1$ for all $\eta \in T$. Further, it holds

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi\left(W;\theta_0,(\eta^{(1)},\eta^{(2)})\right)-\psi\left(W;\theta_0,(\eta^{(1)},\eta_0^{(2)})\right)\right)^2\right]^{1/2}\\
=&\left\|\psi\left(W;\theta_0,(\eta^{(1)},\eta^{(2)})\right)-\psi\left(W;\theta_0,(\eta^{(1)},\eta_0^{(2)})\right)\right\|_{P,2}\\
\leq&\mathcal{I}_1+\mathcal{I}_2
\end{aligned}$$

with

$$\begin{aligned}
\mathcal{I}_1 &:= \left\|\left(\tilde{m}(X)^{-1}-m_0(X)^{-1}\right)D(Y-g(1,X))\right\|_{P,2}\\
\mathcal{I}_2 &:= \left\|\left((1-\tilde{m}(X))^{-1}-(1-m_0(X))^{-1}\right)(1-D)(Y-g(0,X))\right\|_{P,2}.
\end{aligned}$$

Remark that $\epsilon \leq m_0(X) \leq 1-\epsilon$ and $\epsilon \leq \tilde{m}(X) \leq 1-\epsilon$, such that

$$\begin{aligned}
\mathcal{I}_1 &\leq \epsilon^{-2}\left\|(m_0(X)-\tilde{m}(X))D(Y-g(1,X))\right\|_{P,2}\\
&\leq \epsilon^{-2}\left\|(m_0(X)-\tilde{m}(X))(g_0(1,X)+U-g(1,X))\right\|_{P,2}\\
&\leq \epsilon^{-2}\left\|(m_0(X)-\tilde{m}(X))(g_0(1,X)-g(1,X))\right\|_{P,2}+\epsilon^{-2}\left\|(m_0(X)-\tilde{m}(X))U\right\|_{P,2}\\
&\lesssim \left\|m_0(X)-\tilde{m}(X)\right\|_{P,2}
\end{aligned}$$

since $\mathsf{E}[U^2|X] \leq C$ and $\|g-g_0\|_{P,\infty} \leq C$. With a analogous argument it holds

$$\mathcal{I}_2 \lesssim \left\|m_0(X)-\tilde{m}(X)\right\|_{P,2}.$$

This implies

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_2(\eta^{(1)})}\|f\|_{P,2}^2\\
\leq&\sup_{(\eta^{(1)},\eta^{(2)})\in\widetilde{\mathcal{T}}_N}\mathbb{E}\left[\left(\psi\left(W;\theta_0,(\eta^{(1)},\eta^{(2)})\right)-\psi\left(W;\theta_0,(\eta^{(1)},\eta_0^{(2)})\right)\right)^2\right]\\
\leq& C\left\|m_0(X)-\tilde{m}(X)\right\|_{P,2}^2\\
\lesssim&\epsilon_N^2.
\end{aligned}$$

Further, remark that using the same argument as above

$$\begin{aligned}
&\left|\psi\left(W;\theta_0,(\eta^{(1)},\eta^{(2)})\right)-\psi\left(W;\theta_0,(\eta^{(1)},\eta_0^{(2)})\right)\right|\\
\lesssim&|m_0(X)-\tilde{m}(X)||D(Y-g(1,X))+(1-D)(Y-g(0,X))|\\
\leq&L(W)|m_0(X)-\tilde{m}(X)|
\end{aligned}$$

where

$$\sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \|L(W)\|_{P,q} \lesssim \|U\|_{P,q} + \sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \|g_0(D,X) - g(D,X)\|_{P,q} \lesssim 1.$$

Applying Lemma O.1 from Belloni et al. (2018) implies

$$\begin{aligned}
&\log \sup_Q N(\epsilon \|L(W)\|_{Q,2}, \mathcal{F}_2(\eta^{(1)}), \|\cdot\|_{Q,2}) \\
\leq &\log \sup_Q N(\epsilon, \mathcal{M}, \|\cdot\|_{Q,2}) \\
\leq &C\epsilon^{-1}
\end{aligned}$$

by Assumption 14 (ii) , where we can use 1 as an envelope for $\mathcal{M}$. According to the discussion of Assumption 12, it holds

$$u_n \lesssim \sigma_n^{1/2} \|F_2(\eta^{(1)})\|_{P,2}^{-1/2}.$$

Further,

$$\begin{aligned}
\|F_2(\eta^{(1)})\|_{P,q} &\leq \left\| \sup_{\eta^{(2)}\in\widetilde{\mathcal{T}}_N^{(2)}} \left(\psi\left(W;\theta_0,(\eta^{(1)},\eta^{(2)})\right) - \psi\left(W;\theta_0,(\eta^{(1)},\eta_0^{(2)})\right)\right)\right\|_{P,q} \\
&= \left\| L(W) \sup_{\eta^{(2)}\in\widetilde{\mathcal{T}}_N^{(2)}} |\eta^{(2)}(X) - \eta_0^{(2)}(X)| \right\|_{P,q} \lesssim 1.
\end{aligned}$$

Choose $\sigma_n = \epsilon_N \vee \log^{1/2}(n) n^{1/q-1/2}$, such that

$$\sup_{f\in \mathsf{F}_2(\eta^{(1)})} \|f\|_{P,2}^2 \leq \sigma_n^2 \leq \|F_2(\eta^{(1)})\|_{P,2}^2$$

for each $\eta^{(1)}$ and n large enough. Consequently,

$$\begin{aligned}
&\sigma_n^{1/2} \sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \|F_2(\eta^{(1)})\|_{P,2}^{1/2} + \sigma_n\sqrt{\log(n)} \\
&\quad + n^{1/q-1/2} V_n \left(\frac{\sup_{\eta^{(1)}\in\widetilde{\mathcal{T}}_N^{(1)}} \|F_2(\eta^{(1)})\|_{P,2}}{\sigma_n} \vee \log(n) \right) \\
&\leq C\sigma_n + \sigma_n\sqrt{\log(n)} + Cn^{1/q-1/2}\left(\sigma_n^{-1} \vee \log(n)\right) = o(1).
\end{aligned}$$

□

3.6.3 Details on Calibration

Details on Calibration Methods

Venn-Abers Calibration

Venn-Abers predictors (VAPs) (Vovk et al. 2004) provide calibrated probability estimates by considering both possible labels for a test instance and fitting separate isotonic regressions for each case. For binary outcomes (Vovk and Petej 2014) $D \in \{0, 1\}$, VAPs assign each test unit X_{l+1} two probabilities: $\tilde{m}_0(W)$ and $\tilde{m}_1(W)$, derived from isotonic regression under the assumed labels $D_{l+1} = 0$ and $D_{l+1} = 1$. These estimates are oracle calibrated—the true label's corresponding probability is valid. While aligning one probability distribution with the true outcome ensures accurate calibration, the "oracle" selector $S = D$ is not known in practice. This means we often need to use heuristic methods, like averaging, to combine information, even though these methods do not come with formal guarantees. Computational expense also arises from refitting isotonic models per test instance.

Inductive Venn-Abers predictors (IVAPs) (Lambrou et al. 2012, Nouretdinov et al. 2018) address these issues by splitting data into a proper training set (to fit a propensity model $\hat{m}(X)$) and a calibration set (to fit isotonic maps $\tilde{m}_0, \tilde{m}_1$). For square loss, IVAPs combine $p_0 = \tilde{m}_0(\hat{m}(X))$ and $p_1 = \tilde{m}_1(\hat{m}(X))$ into a single probability: $p = p_1 + \frac{p_0^2}{2} - \frac{p_1^2}{2}$. As Vovk and Petej (2014) point out, this can be rewritten as: $p = \bar{p} + (p_1 - p_0)\left(\frac{1}{2} - \bar{p}\right), \quad \bar{p} = \frac{p_0+p_1}{2}$. Thus, p is a regularized version of $\bar{p}$ moving the prediction towards $\frac{1}{2}$. This characteristic is particularly advantageous in mitigating the risk of inflated weights during treatment effect estimation.

Platt Scaling

Platt scaling (Platt 1999) leverages the robust calibration properties of log-loss. It applies logistic regression, expressed by the function $f(\hat{m}(X)) := \frac{1}{1+\exp(A\hat{m}(X)+B)}$, to the scores produced by an estimator, using treatment assignment labels as targets for calibration, where $A < 0$ and B are parameters. The parameters A and B are estimated using a maximum likelihood method on the same training set as the original classifier $m(x)$. To avoid overfitting, a held-out calibration set or cross-validation can be used. Platt also recommends clipping the outputs d to target probabilities in the range $\left(\frac{1}{N_0+2}, \frac{N_1+1}{N_1+2}\right)$, where N_0 is the number of control units and N_1 is the number of treated units in the calibration set. This motivates the testing of additional clipping in small sample sizes, a setting where the parametric form assumption is especially appropriate. Unlike isotonic regression, for which van der Laan et al. (2023) established distribution-free calibration guarantees, Platt scaling lacks universal theoretical guarantees. This limitation is critical in propensity calibration, where treatment assignments can be unbalanced, or propensity scores deviate from logistic normality (e.g., highly skewed or multimodal distributions). Empirical studies further challenge its reliability: Kumar et al. (2019) show that Platt scaling's apparent calibration is often inflated due to the systematic underestimation of errors in continuous output spaces, where true calibration cannot be verified without uncheckable smoothness assumptions. Recent work by Li and Sur (2025) identifies specific conditions - notably Gaussian-like or light-tailed feature distributions - under which Platt scaling achieves

Bregman optimality, minimizing divergences such as log loss or squared error, even in high-dimensional settings.

Alternative Calibration Metrics

Commonly employed are the L1-Norm and L2-Norm. (Nixon et al. 2020) suggests using either the L2-norm or adaptive intervals, as the L1-norm is highly susceptible to the decision on the number of intervals. To assess subpopulations with extreme propensity scores, the maximum calibration error (MCE) (Naeini et al. 2015) can be used. MCE is defined as the maximum deviation across all bins, given by: $\max_{i\in\{1,\dots,M\}} |\mathrm{acc}_i(B_i) - \mathrm{conf}_i(B_i)|$. This error reflects the worst-case deviation between predicted and actual values across all intervals. All calibration metrics are visualized with respect to the observation size and the underlying propensity learner in Supplement 3.6.6.

3.6.4 Details on Calibration for Double Machine Learning

Details on Calibration for Partially Linear Regression Models

Consider the partially linear regression model as in Chernozhukov et al. (2018). Let $D \in \{0,1\}$ be a binary treatment variable and $W = (Y, D, X)$, where

$$Y = \theta_0 D + g_0(X) + U,\ \mathsf{E}[U|D,X] = 0, \quad D = m_0(X) + V,\ \mathsf{E}[V|X] = 0, \tag{3.10}$$

with θ_0 being the parameter of interest. Let

$$\psi(W;\theta,\eta) := (Y - l(X) - \theta(D - m(X))(D - m(X))$$

be the "partialling-out" score function, where $\eta = (l, m)$ denotes the nuisance functions for the outcome regression $l_0(X) = \mathbb{E}[Y|X]$ and propensity score $m_0(X) = \mathbb{E}[D|X]$.

Assumption 15 (cf. Assumption 4.1 in Chernozhukov et al. (2018))**.** *Let $(\delta_N)_{n=1}^{\infty}$ and $(\Delta_N)_{n=1}^{\infty}$ be sequences of positive constants approaching 0 as before. Also, let c, C and q be fixed strictly positive constants such that $q > 4$, and let $K \geq 2$ be a fixed integer. Moreover, for any $\eta = (\eta_1, \eta_2)$, where η_1 and η_2 are functions mapping the support of X to $\mathbb{R}$, denote $\|\eta\|_{p,q} = \|\eta_1\|_{p,q} \vee \|\eta_2\|_{p,q}$. For simplicity, assume that N/K is an integer. For all probability laws $P \in \mathcal{P}$ for the triple (Y, D, X), the following conditions hold:*

(a) Equations (3.10) *hold*

(b) $\|Y\|_{P,q} + \|D\|_{P,q} \leq C$

(c) $\|UV\|_{P,2} \geq c^2$ *and* $\mathsf{E}_P[V^2] \geq c$

(d) $\|\mathbb{E}_P[U^2 \mid X]\|_{P,\infty} \leq C$ *and* $\|\mathbb{E}_P[V^2 \mid X]\|_{P,\infty} \leq C$

(e) Given a random subset I of $[N]$ of size $n = N/K$, the nuisance parameter estimator $\hat{\eta}_0 = ((W_i)_{i\in I^c})$ obeys the following conditions for all $N \geq 1$. With P-probability no less than $1 - \Delta_N$,

(i) $\|\hat{\eta}_0 - \eta_0\|_{P,q} \leq C$ *and* $\|\hat{\eta}_0 - \eta_0\|_{P,2} \leq \delta_N$

(ii) $\|\hat{m}_0 - m_0\|_{P,2} \times \left(\|\hat{m}_0 - m_0\|_{P,2} + \|\hat{\ell}_0 - \ell_0\|_{P,2}\right) \leq \delta_N N^{-1/2}$.

Under Assumption 15, Remark 2 holds (cf. Theorem 4.1 in Chernozhukov et al. (2018)). To apply Theorem 1 using a calibrated propensity score estimator, Assumption 4 has to be satisfied. To this end, we introduce the following Assumption 16.

Assumption 16 (Calibration rate and complexity)**.** *Let* $\tilde{m}(X)$ *be an estimator of* $\tilde{m}_0(X) := \mathsf{E}[D|\hat{m}(X)]$. *We assume*

(i) The following convergence rates hold with P*-probability no less than* $1 - \Delta_N$*,*

$$\|\tilde{m}(X) - \tilde{m}_0(X)\|_{P,2} \lesssim \varepsilon_N,$$

such that

$$\varepsilon_N \cdot \left(\varepsilon_N + \|\hat{\ell}_0 - \ell_0\|_{P,2}\right) \leq \delta_N N^{-1/2}.$$

(ii) Let $\tilde{m}(\cdot) \in \mathcal{M}$*, such that the covering numbers obey*

$$\sup_Q N(\epsilon, \mathcal{M}, L_2(Q)) \leq C\epsilon^{-1}.$$

It is worth noting that this calibration assumption does not necessarily require D to be binary.

Theorem 3. *Under Assumptions 15 and 16(i) Remark 3 is valid. If additionally Assumption 16(ii) is satisfied, Theorem 1 holds.*

The proof of Theorem 3 is similar to the proof of Theorem 2 and is therefore left out. It just needs an additional complexity argument to verify Assumption 4 (ii) based on Assumption 16(ii).

Algorithms of Section 3.3

In this section, we provide the pseudo-codes for additional algorithms tested in our simulation study.

Algorithm 2 (nested K-fold cross-fitting calibration) DML 2 Algorithm

Input: Data $(W_i)_{i=1}^N$. A K-fold random partition $(I_k)_{k=1}^K$ of $[N] = \{1, \ldots, N\}$ such that each fold I_k is of size $n = N/K$. For each $k \in [K] = \{1, \ldots, K\}$ define $I_k^c := \{1, \ldots, N\} \setminus I_k$.

2: For each $k \in [K]$, fit a machine learning estimator

$$\hat{\eta}_{0,k}^{(1)} = \hat{\eta}_0^{(1)}\big((W_i)_{i \in I_k^c}\big)$$

of $\eta_0^{(1)}$, where $\hat{\eta}_{0,k}^{(1)}$ is a random element in $T^{(1)}$, where the randomness only depends on the $(W_i)_{i \in I_k^c}$. For each $k \in [K]$, split the training partition I_k^c into two disjoint samples $I_{k,1}^c$ and $I_{k,2}^c$.

4: Use the first subset $I_{k,1}^c$ to fit a machine learning estimator

$$\hat{\eta}_{0,k}^{(2)} = \hat{\eta}_0^{(2)}\big((W_i)_{i \in I_{k,1}^c}\big)$$

of $\eta_0^{(2)}$, where $\hat{\eta}_{0,k}^{(2)}$ is a random element in $T^{(2)}$, where the randomness only depends on the $(W_i)_{i \in I_{k,1}^c}$. Use the second subset $I_{k,2}^c$ and the estimated nuisance element $\hat{\eta}_{0,k}^{(2)}$ to fit a re-estimation procedure

$$\tilde{\eta}_{0,k}^{(2)} = \tilde{\eta}_0^{(2)}\big((W_i)_{i \in I_{k,2}^c}, \hat{\eta}_{0,k}^{(2)}\big)$$

of $\eta_0^{(2)}$, where $\tilde{\eta}_{0,k}^{(2)}$ is a random element in $T^{(2)}$, where the randomness only depends on the $(W_i)_{i \in I_k^c}$.

6: Construct the estimator $\tilde{\theta}_0$ as the solution to

$$\frac{1}{K} \sum_{k=1}^{K} \mathsf{E}_{n,k}\big[\psi\big(W; \tilde{\theta}_0, (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,k}^{(2)})\big)\big] = 0,$$

where $\mathsf{E}_{n,k}[\psi(W)] = n^{-1} \sum_{i \in I_k} \psi(W_i)$ is the empirical expectation over $(W_i)_{i \in I_k}$.

Algorithm 4 (single split cross-fitting calibration) DML 2 Algorithm

1: **Input:** Data $(W_i)_{i=1}^N$. A K-fold random partition $(I_k)_{k=1}^K$ of $[N] = \{1, \dots, N\}$ such that each fold I_k is of size $n = N/K$. A 2-fold random partition $\tilde{I}_1 \cup \tilde{I}_2 = \{1, \dots, N\}$ and $\tilde{I}_1 \cap \tilde{I}_2 = \emptyset$. For each $k \in [K] = \{1, \dots, K\}$ define $I_k^c := \{1, \dots, N\} \setminus I_k$ and for $j \in [2] = \{1, 2\}$ define $\tilde{I}_j^c := \{1, \dots, N\} \setminus \tilde{I}_j$.

2: For each $k \in [K]$, fit a machine learning estimator

$$\hat{\eta}_{0,k}^{(1)} = \hat{\eta}_0^{(1)}\big((W_i)_{i \in I_k^c}\big)$$

of $\eta_0^{(1)}$, where $\hat{\eta}_{0,k}^{(1)}$ is a random element in $T^{(1)}$, where the randomness only depends on the $(W_i)_{i \in I_k^c}$.

3: For each $j \in [2]$, fit a machine learning estimator

$$\hat{\eta}_{0,j}^{(2)} = \hat{\eta}_0^{(2)}\big((W_i)_{i \in \tilde{I}_j^c}\big)$$

of $\eta_0^{(2)}$, where $\hat{\eta}_{0,j}^{(2)}$ is a random element in $T^{(2)}$, where the randomness only depends on the $(W_i)_{i \in \tilde{I}_j^c}$.

4: For each $j \in [2]$, rely on estimated nuisance element $\hat{\eta}_{0,j}^{(2)}$ to fit a re-estimation procedure

$$\tilde{\eta}_{0,j}^{(2)} = \tilde{\eta}_0^{(2)}\big((W_i)_{i \in \tilde{I}_j}, \hat{\eta}_{0,j}^{(2)}\big)$$

of $\eta_0^{(2)}$, where $\tilde{\eta}_{0,j}^{(2)}$ is a random element in $T^{(2)}$.

5: Construct the estimator $\tilde{\theta}_0$ as the solution to

$$\frac{1}{2K} \sum_{k=1}^{K} \sum_{j=1}^{2} \mathsf{E}_{n,k,j}\big[\psi\big(W; \tilde{\theta}_0, (\hat{\eta}_{0,k}^{(1)}, \tilde{\eta}_{0,j}^{(2)})\big)\big] = 0,$$

where $\mathsf{E}_{n,k,j}[\psi(W)] = n^{-1} \sum_{i \in I_k \cap \tilde{I}_j} \psi(W_i)$ is the empirical expectation over $(W_i)_{i \in I_k \cap \tilde{I}_j}$.

3.6.5 Details on the DGPs

Table 3.3: Overview of the DGPs

DGP	Covariates	Description of $D \mid X$	Description of $g_0(D, X)$
1	$X_i \sim N(0, \Sigma)$ with $\Sigma_{kj} = (0.5)^{\lvert j-k \rvert}$	Sigmoid function combined with Uniform indicator. D is exogenous conditional on X.	Heterogeneous treatment effect, linear with simple interaction, $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$
2	$X_1 \sim$ Bin$(1, 0.5), X_2 \mid X_1 \sim \Gamma, X_3 \mid X_2 \sim$ Beta	Highly nonlinear, tree-based via conditions, $D \sim$ Bin$(1, m_0(X))$	Homogeneous treatment effect, Poisson with simple linear combination of covariates
3	$X \sim$ Unif$[-1, 1]^4$	Expit function combined with $D \sim$ Bin$(1, m_0(X))$	Heterogeneous treatment effect, nonlinear transformations of the covariates, $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$
4	$X \sim$ Unif$[0, 1]^{20}$	Unbalanced and nonlinear $m_0(X) = \alpha\,(1 + \beta_{2,4}(\min(X_1, X_2)))$, $D \sim$ Bin$(1, m_0(X))$	Heterogeneous treatment effect, scaled Friedman function for $g_0(X)$ and simple interaction of D and X, $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$

A detailed description of DGP 2-4 is given in the Supplementary Material.

In our first DGP, data units were generated following Belloni et al. (2017)[9].

DGP 2 is adapted from Deshpande and Kuleshov (2023). Three covariates are simulated: gender (X_1), age (X_2), and disease severity (X_3), while treatment (D) corresponds to the administration of a drug. The outcome Y is the time taken for the recovery of a patient. In addition to the original setting, we introduced non-deterministic treatment assignments to ensure positivity. The function for the treatment assignment resembles the approach of tree-based models, favoring the IRM and PLR in combination with LGBM and random forest.

For a complicated nonlinear outcome regression, we implemented DGP 3 (van der Laan et al. 2023). The included transformations make it difficult for linear models as well as tree-based classifiers through the local linearity. Given the nonlinear and heterogeneous influence of the treatment on the outcome, we expect the IRM model to outperform the PLR. However, the treatment assignment itself can be modeled consistently, favoring the IPW.

The unbalanced treatment assignment DGP 4 follows Ballinari (2024) and is adapted from Nie and Wager (2020). Three settings of the share of treated $\mathsf{E}[D]$ are tested: $\mathsf{E}[D] \in \{0.05, 0.1, 0.2\}$. To address positivity violations, we tested three clipping thresholds for the propensity scores: $1e-12$, 0.01, and 0.1. Given the majority non-treatment class, a lower threshold is sufficient to limit the contribution of a single observation.

DGP 1 IRM, Detailed Description

$$
\begin{aligned}
d_i &= 1\left\{\frac{\exp(c_D X_i'\beta)}{1+\exp(c_D X_i'\beta)} > V_i\right\}, && V_i \sim \mathcal{U}(0,1),\\
y_i &= \theta D_i + c_Y X_i'\beta D_i + \zeta_i, && \zeta_i \sim \mathcal{N}(0,1),
\end{aligned}
$$

where $v_i \sim U(0,1)$, $\zeta_i \sim N(0,1)$, V_i and ζ_i are independent, $p = \dim(X_i)$, the covariates $X_i \sim N(0,\Sigma)$ with $\Sigma_{kj} = (0.5)^{|j-k|}$. β is a $p \times 1$ vector with elements set as $\beta_{1,j} = (1/j)^2$ for $j = 1, ..., p$. c_D and c_Y are scalars given by

$$
c_Y = \sqrt{\frac{R_Y^2}{(1-R_Y^2)\beta'\Sigma\beta}}, \qquad c_D = \sqrt{\frac{(\pi^2/3)R_D^2}{(1-R_D^2)\beta'\Sigma\beta}}.
$$

that control the strength of the relationship between the controls, the outcome, and the treatment variable. The underlying treatment effect is heterogeneous. Hence, to accurately model the interaction effects between treatment and covariates, the IRM should be considered. Simple propensity score weighting in the IPW, as well as the additive structural form assumptions for the treatment effect of the PLR, should be inconsistent for this DGP.

DGP 2 Drug Effectiveness, Detailed Description

In our second DGP, data units were generated as follows:

[9]This DGP is available at DoubleMLIRMData.

- Gender indicator X_1 and age X_2 are simulated as:

$$X_1 \sim \text{Bin}(0, 0.5), \quad X_2 \sim \Gamma(\kappa_{\text{age}}, \psi_{\text{age}}).$$

The distribution of age is gender dependent:

$$\kappa_{\text{age}} = \left(\frac{\mu_{\text{age}}}{\sigma_{\text{age}}}\right)^2, \quad \psi_{\text{age}} = \frac{\sigma^2_{\text{age}}}{\mu_{\text{age}}}, \text{ and}$$

$$\mu_{\text{age}} = \begin{cases} 49 & \text{if } X_1 = 1, \\ 51 & \text{if } X_1 = 0, \end{cases} \quad \sigma_{\text{age}} = \begin{cases} 7 & \text{if } X_1 = 1, \\ 8 & \text{if } X_1 = 0. \end{cases}.$$

- Disease severity X_3 is simulated based on age:

$$\mu_{\text{disease}} = \text{clip}\left(\frac{X_2 - 20}{20}, 0.1, 5\right), \quad X_3 \sim \text{Beta}(\mu_{\text{disease}}, 2).$$

- Propensity scores are generated using linear predictors, incorporating adjustable levels of overlap through the parameter β:

$$\text{coefficients} = \begin{bmatrix} (-0.4, 0.2, 0.8), \\ (-0.4 + 2(1-\beta), 0.2, 0.8), \\ (-0.4 + 2(1-\beta), 0.3, 1), \\ (-0.4 + 2(1-\beta), 0.1, 1.2), \\ (-0.4 + 2(1-\beta), 0.1, 1.2) \end{bmatrix},$$

$$\begin{aligned} \text{linear\_predictors} = \Big(& b_0 + b_1 \cdot \text{normalize}(X_2) + b_2 \cdot \text{normalize}(X_3) \\ & + \epsilon \mid (b_0, b_1, b_2) \in \text{coefficients}\Big), \\ & \epsilon \sim \mathcal{N}(0, 0.5). \end{aligned}$$

Propensity scores are computed via logistic transformation:

$$\text{prop\_scores} = \left[\frac{1}{1 + \exp(-\text{lp})} \mid \text{lp} \in \text{linear\_predictors}\right].$$

- Treatment assignment m_0 is initialized with prop_scores[0] and updated based on conditions:

$$\text{conditions} = \begin{cases} ((X_1 = 0) \text{ and } (X_2 > 55) \text{ and } (X_3 \leq 0.55), \text{prop\_scores}[1]), \\ ((X_1 = 1) \text{ and } (X_2 > 55) \text{ and } (X_3 \leq 0.55), \text{prop\_scores}[2]), \\ ((X_1 = 0) \text{ and } (X_3 > 0.55), \text{prop\_scores}[3]), \\ ((X_1 = 1) \text{ and } (X_3 > 0.55), \text{prop\_scores}[4]) \end{cases}$$

$$\text{for } (\text{condition}, \text{score}) \in \text{conditions} : m_0[\text{condition}] = \text{score}[\text{condition}].$$

Final treatment assignment D_i is:

$$D_i \sim \text{Bin}(1, m_0).$$

- The outcome Y_i is simulated as:

$$Y_{di} = \text{Pois}(2 + 0.5X_1 + 0.03X_2 + 2X_3 - d), \qquad Y_i = D_i \cdot Y_{1i} + (1 - D_i) \cdot Y_{0i}.$$

This DGP captures complex relationships between gender, age, disease severity, and treatment assignment, providing a realistic framework for evaluating causal inference methods. The treatment assignments are highly nonlinear and resemble decision trees with underlying local linearities.

DGP 3 Nonlinear, Detailed Description

For the nonlinear outcome regression, the detailed data generation process is as follows:

$$X \sim \text{Unif}[-1, 1]^4, \quad D|X \sim \text{Bin}(1, p(X)), \quad Y \sim \text{Bin}(1, \mu(X, D))$$

with the treatment assignment probability given by:

$$p(X) = \text{expit}\{-0.25 - X_1 + 0.5X_2 - X_3 + 0.5X_4\}.$$

The outcome function is based on nonlinear transformations of the covariates, with added treatment effect interactions.

$$\begin{aligned}\mu(X, D) &= 1.5 + 1.5D + 2D|X_1||X_2| - 2.5(1 - D)|X_2|X_3 + 2.5X_3 \\ &+ 2.5(1 - D)\sqrt{|X_4|} - 1.5D \cdot I(X_2 < 0.5) + 1.5(1 - D)I(X_4 < 0).\end{aligned}$$

The outcome is then generated by:

$$Y_d \sim \text{N}(\mu(X, D), 1), \quad Y = DY_1 + (1 - D)Y_0.$$

DGP 4 Unbalanced, Detailed Description

The unbalanced treatment assignment setting follows Ballinari (2024) and is adapted from Nie and Wager (2020).

$$\begin{aligned}&X_i \sim \text{Unif}[0, 1]^{20}, \quad D|X \sim \text{Bin}(1, p(X)), \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \\ &Y = b(X) + (D - 0.5)(X_1 + X_2) + \epsilon_i.\end{aligned}$$

The baseline main effect is the scaled Friedman (1991) function:

$$b(X_i) = \sin(\pi X_1 X_2) + 2(X_3 - 0.5)^2 + X_4 + 0.5X_5.$$

For the propensity score, we follow Künzel et al. (2019) and set:

$$m_0(X) = \alpha \left(1 + \beta_{2,4}(\min(X_1, X_2))\right),$$

where $\beta_{2,4}(\cdot)$ is the beta cumulative distribution function with shape parameters 2 and 4. This unbalanced setting entails difficult nuisance components and an easy treatment effect function, favoring doubly robust causal models compared to IPW.

The share of treated is $\mathsf{E}[D] = (31/21)\alpha$. Three settings of the share of treated $\mathsf{E}[D]$ are tested: $\mathsf{E}[D] \in \{0.05, 0.1, 0.2\}$. For unbalanced applications, several adjustments for ATE estimators exist to address positivity violations. The most commonly used method is clipping the propensity scores to preset levels. Throughout our simulations, we tested three different levels of clipping thresholds.

As Imbens (2004) points out, for the IPTW with $n = 4000$ observations, choosing a clipping cut-off at 0.01[10] limits the contribution of a single observation to 0.025. We also tested more severe clipping as employed in Nie and Wager (2020), with a clipping level at 0.1, and alternatively, without clipping [11]. Particularly, for the setting with only 5% treated, we expect that a clipping threshold of 0.1 introduces a strong bias.

[10] Given the majority non-treatment class, a lower threshold is sufficient.

[11] To ensure computational stability we set the threshold at $1e-12$.

3.6.6 Detailed Simulation Results

3.6.7 Extended Results Overview

Table 3.4: IPW Results

DGP	Method	m = Logit			m = Random Forest			m = LGBM		
		MAE	RMSE	Std. dev.	MAE	RMSE	Std. dev.	MAE	RMSE	Std. dev.
1	Alg-1-Clipped	0.08	0.11	0.11	0.18	0.19	0.06	0.94	0.98	0.31
	Alg-1-Uncalib	0.10	0.16	0.16	1.09e+06	1.09e+07	1.09e+07	1.57	1.77	0.83
	Alg-2-nested-cf-IVAP	0.10	0.12	0.09	0.14	0.15	0.07	0.17	0.19	0.08
	Alg-2-nested-cf-Iso	0.20	0.25	0.19	0.13	0.17	0.16	0.13	**0.16**	0.16
	Alg-2-nested-cf-Platt	0.10	0.17	0.17	0.14	0.15	0.06	0.22	0.23	0.06
	Alg-3-cf-IVAP	0.11	0.13	0.07	0.14	0.16	0.07	0.17	0.18	0.07
	Alg-3-cf-Iso	0.09	0.11	0.07	0.13	0.15	0.06	0.15	0.17	0.06
	Alg-3-cf-Platt	0.13	0.14	0.06	0.35	0.36	0.05	0.31	0.31	0.05
	Alg-4-single-split-IVAP	0.11	0.12	0.07	0.31	0.31	0.03	0.33	0.33	0.04
	Alg-4-single-split-Iso	0.13	0.18	0.16	0.30	0.30	0.03	0.32	0.32	0.04
	Alg-4-single-split-Platt	0.10	0.12	0.09	0.16	0.17	0.06	0.23	0.24	0.06
	Alg-5-full-sample-IVAP	0.08	0.10	0.08	0.12	**0.14**	0.07	0.14	**0.16**	0.07
	Alg-5-full-sample-Iso	0.07	**0.09**	0.08	0.12	**0.14**	0.06	0.14	**0.16**	0.07
	Alg-5-full-sample-Platt	0.09	0.11	0.10	0.14	0.16	0.06	0.22	0.22	0.05
2	Alg-1-Clipped	0.09	0.12	0.11	4.45	4.59	1.11	2.67	2.74	0.61
	Alg-1-Uncalib	0.09	0.12	0.11	1.17e+10	1.37e+10	7.23e+09	2.86	2.98	0.81
	Alg-2-nested-cf-IVAP	0.21	0.26	0.26	0.28	0.33	0.26	0.29	0.33	0.25
	Alg-2-nested-cf-Iso	0.60	0.81	0.66	0.47	0.61	0.61	0.48	0.64	0.62
	Alg-2-nested-cf-Platt	0.20	0.25	0.25	0.29	0.34	0.24	0.29	0.34	0.23
	Alg-3-cf-IVAP	0.10	0.12	0.12	0.21	0.24	0.12	0.20	0.23	0.11
	Alg-3-cf-Iso	0.09	**0.11**	0.11	0.21	0.24	0.11	0.20	0.23	0.11
	Alg-3-cf-Platt	0.13	0.16	0.11	0.44	0.45	0.11	0.50	0.51	0.10
	Alg-4-single-split-IVAP	0.11	0.14	0.14	0.96	0.97	0.14	0.90	0.91	0.11
	Alg-4-single-split-Iso	0.61	0.77	0.53	0.95	0.96	0.14	0.87	0.88	0.11
	Alg-4-single-split-Platt	0.10	0.12	0.11	0.26	0.28	0.12	0.27	0.29	0.11
	Alg-5-full-sample-IVAP	0.09	**0.11**	0.11	0.20	**0.22**	0.11	0.18	**0.20**	0.11
	Alg-5-full-sample-Iso	0.09	**0.11**	0.11	0.20	0.23	0.10	0.18	**0.20**	0.10
	Alg-5-full-sample-Platt	0.10	0.12	0.11	0.27	0.29	0.10	0.27	0.29	0.10
3	Alg-1-Clipped	0.06	**0.08**	0.08	0.54	0.58	0.21	2.00	2.02	0.31
	Alg-1-Uncalib	0.06	**0.08**	0.08	3.07e+08	8.49e+08	7.91e+08	2.05	2.07	0.34
	Alg-2-nested-cf-IVAP	0.12	0.15	0.14	0.36	0.38	0.11	0.46	0.48	0.11
	Alg-2-nested-cf-Iso	0.94	1.03	0.41	0.26	**0.32**	0.31	0.27	**0.34**	0.34
	Alg-2-nested-cf-Platt	0.11	0.14	0.14	0.40	0.42	0.10	0.53	0.54	0.10
	Alg-3-cf-IVAP	0.07	**0.08**	0.08	0.32	0.33	0.08	0.38	0.38	0.07
	Alg-3-cf-Iso	0.07	0.09	0.08	0.36	0.37	0.07	0.42	0.42	0.07
	Alg-3-cf-Platt	0.12	0.14	0.07	0.87	0.87	0.08	0.82	0.82	0.08
	Alg-4-single-split-IVAP	0.07	0.09	0.09	1.13	1.13	0.07	1.03	1.03	0.08
	Alg-4-single-split-Iso	0.73	0.80	0.34	1.13	1.13	0.07	1.03	1.03	0.08
	Alg-4-single-split-Platt	0.08	0.09	0.08	0.44	0.45	0.08	0.57	0.57	0.08
	Alg-5-full-sample-IVAP	0.06	**0.08**	0.08	0.35	0.36	0.07	0.40	0.41	0.07
	Alg-5-full-sample-Iso	0.06	**0.08**	0.08	0.37	0.37	0.07	0.42	0.43	0.07
	Alg-5-full-sample-Platt	0.07	0.09	0.08	0.42	0.43	0.07	0.52	0.53	0.07
4	Alg-1-Clipped	0.13	0.15	0.06	0.54	0.55	0.13	6.14	6.17	0.52
	Alg-1-Uncalib	0.13	0.15	0.06	4.64e+08	6.89e+08	5.11e+08	8.37	8.45	1.21
	Alg-2-nested-cf-IVAP	0.10	0.13	0.11	0.12	0.15	0.11	0.11	0.14	0.11
	Alg-2-nested-cf-Iso	0.21	0.24	0.13	0.19	0.23	0.13	0.25	0.29	0.15
	Alg-2-nested-cf-Platt	0.13	0.16	0.11	0.15	0.18	0.11	0.15	0.18	0.12
	Alg-3-cf-IVAP	0.05	**0.05**	0.05	0.06	**0.08**	0.05	0.06	**0.07**	0.06
	Alg-3-cf-Iso	0.05	0.06	0.05	0.09	0.10	0.05	0.07	0.09	0.06
	Alg-3-cf-Platt	0.07	0.09	0.06	0.13	0.16	0.10	0.18	0.23	0.16
	Alg-4-single-split-IVAP	0.05	0.06	0.06	1.40	1.41	0.09	1.43	1.43	0.07
	Alg-4-single-split-Iso	0.07	0.09	0.07	0.88	0.98	0.43	0.96	1.02	0.37
	Alg-4-single-split-Platt	0.10	0.12	0.06	0.12	0.13	0.06	0.12	0.13	0.05
	Alg-5-full-sample-IVAP	0.06	0.08	0.06	0.10	0.11	0.05	0.09	0.10	0.06
	Alg-5-full-sample-Iso	0.07	0.08	0.05	0.11	0.12	0.05	0.10	0.11	0.06
	Alg-5-full-sample-Platt	0.09	0.11	0.06	0.12	0.13	0.06	0.12	0.13	0.06

DGP 1: n = 2000, p = 20, R2_d = 0.5; DGP 2: n = 2000, p = 3, overlap = 0.5;
DGP 3: n = 2000, p = 4; DGP 4: n = 4000, p = 20, share treated = 0.1; g = LGBM,
lowest RMSEs per DGP and propensity learner are highlighted

Table 3.5: IRM Results

DGP	Method	m = Logit			m = Random Forest			m = LGBM		
		MAE	RMSE	Std. dev.	MAE	RMSE	Std. dev.	MAE	RMSE	Std. dev.
1	Alg-1-Clipped	0.07	0.10	0.10	0.06	0.08	0.07	0.22	0.27	0.24
	Alg-1-Uncalib	0.08	0.13	0.13	1.85e+06	1.85e+07	1.84e+07	0.48	0.60	0.54
	Alg-2-nested-cf-IVAP	0.07	0.09	0.08	0.06	**0.08**	0.07	0.07	**0.08**	0.08
	Alg-2-nested-cf-Iso	0.12	0.15	0.14	0.10	0.12	0.12	0.10	0.12	0.12
	Alg-2-nested-cf-Platt	0.08	0.14	0.14	0.06	**0.08**	0.07	0.07	**0.08**	0.07
	Alg-3-cf-IVAP	0.06	0.08	0.07	0.06	**0.08**	0.08	0.07	**0.08**	0.07
	Alg-3-cf-Iso	0.06	0.08	0.07	0.06	**0.08**	0.08	0.06	**0.08**	0.07
	Alg-3-cf-Platt	0.06	**0.07**	0.07	0.07	0.08	0.06	0.07	**0.08**	0.07
	Alg-4-single-split-IVAP	0.06	0.08	0.07	0.07	**0.08**	0.06	0.07	0.09	0.06
	Alg-4-single-split-Iso	0.10	0.13	0.12	0.07	**0.08**	0.06	0.07	0.09	0.06
	Alg-4-single-split-Platt	0.07	0.08	0.08	0.06	**0.08**	0.07	0.07	**0.08**	0.07
	Alg-5-full-sample-IVAP	0.07	0.08	0.08	0.07	0.09	0.08	0.07	**0.08**	0.08
	Alg-5-full-sample-Iso	0.06	0.08	0.08	0.06	**0.08**	0.07	0.07	**0.08**	0.08
	Alg-5-full-sample-Platt	0.07	0.09	0.09	0.06	**0.08**	0.07	0.07	**0.08**	0.07
2	Alg-1-Clipped	0.09	**0.11**	0.11	0.31	0.39	0.39	0.20	0.26	0.26
	Alg-1-Uncalib	0.09	**0.11**	0.11	2.22e+09	2.89e+09	2.86e+09	0.24	0.32	0.32
	Alg-2-nested-cf-IVAP	0.09	**0.11**	0.11	0.09	**0.11**	0.11	0.09	0.12	0.12
	Alg-2-nested-cf-Iso	0.18	0.23	0.22	0.16	0.22	0.22	0.19	0.24	0.24
	Alg-2-nested-cf-Platt	0.09	**0.11**	0.11	0.09	**0.11**	0.11	0.09	**0.11**	0.11
	Alg-3-cf-IVAP	0.09	**0.11**	0.11	0.10	0.12	0.12	0.10	0.12	0.12
	Alg-3-cf-Iso	0.09	**0.11**	0.11	0.09	0.12	0.12	0.10	0.12	0.12
	Alg-3-cf-Platt	0.09	**0.11**	0.11	0.09	**0.11**	0.11	0.09	**0.11**	0.11
	Alg-4-single-split-IVAP	0.09	**0.11**	0.11	0.09	**0.11**	0.11	0.09	**0.11**	0.11
	Alg-4-single-split-Iso	0.15	0.20	0.20	0.09	**0.11**	0.11	0.09	**0.11**	0.11
	Alg-4-single-split-Platt	0.09	**0.11**	0.11	0.09	**0.11**	0.11	0.09	**0.11**	0.11
	Alg-5-full-sample-IVAP	0.09	**0.11**	0.11	0.09	**0.11**	0.11	0.09	0.12	0.12
	Alg-5-full-sample-Iso	0.09	**0.11**	0.11	0.09	**0.11**	0.11	0.09	0.12	0.12
	Alg-5-full-sample-Platt	0.09	**0.11**	0.11	0.09	**0.11**	0.11	0.09	**0.11**	0.11
3	Alg-1-Clipped	0.05	0.07	0.07	0.08	0.10	0.10	0.11	0.13	0.13
	Alg-1-Uncalib	0.05	0.07	0.07	9.97e+07	2.74e+08	2.73e+08	0.11	0.14	0.14
	Alg-2-nested-cf-IVAP	0.05	0.07	0.07	0.05	**0.07**	0.06	0.05	**0.07**	0.07
	Alg-2-nested-cf-Iso	0.10	0.13	0.12	0.08	0.10	0.10	0.10	0.13	0.13
	Alg-2-nested-cf-Platt	0.05	0.07	0.07	0.05	**0.07**	0.06	0.05	**0.07**	0.06
	Alg-3-cf-IVAP	0.05	0.07	0.06	0.06	**0.07**	0.07	0.05	**0.07**	0.07
	Alg-3-cf-Iso	0.05	0.07	0.07	0.06	**0.07**	0.07	0.05	**0.07**	0.06
	Alg-3-cf-Platt	0.05	**0.06**	0.06	0.05	**0.07**	0.06	0.05	**0.07**	0.06
	Alg-4-single-split-IVAP	0.05	0.07	0.07	0.05	**0.07**	0.06	0.05	**0.07**	0.06
	Alg-4-single-split-Iso	0.10	0.12	0.11	0.05	**0.07**	0.06	0.05	**0.07**	0.06
	Alg-4-single-split-Platt	0.05	0.07	0.06	0.06	**0.07**	0.07	0.05	**0.07**	0.06
	Alg-5-full-sample-IVAP	0.05	0.07	0.07	0.06	**0.07**	0.07	0.05	**0.07**	0.07
	Alg-5-full-sample-Iso	0.05	0.07	0.07	0.06	**0.07**	0.07	0.05	**0.07**	0.07
	Alg-5-full-sample-Platt	0.05	0.07	0.07	0.06	**0.07**	0.07	0.05	**0.07**	0.06
4	Alg-1-Clipped	0.04	0.06	0.06	0.07	0.09	0.09	0.17	0.21	0.20
	Alg-1-Uncalib	0.04	0.06	0.06	1.74e+08	2.70e+08	2.68e+08	0.36	0.46	0.43
	Alg-2-nested-cf-IVAP	0.04	**0.05**	0.05	0.05	**0.06**	0.06	0.04	**0.06**	0.06
	Alg-2-nested-cf-Iso	0.05	0.07	0.07	0.05	0.07	0.07	0.06	0.07	0.07
	Alg-2-nested-cf-Platt	0.04	0.06	0.05	0.05	**0.06**	0.06	0.04	**0.06**	0.06
	Alg-3-cf-IVAP	0.04	0.06	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-3-cf-Iso	0.04	0.06	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-3-cf-Platt	0.04	0.06	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-4-single-split-IVAP	0.04	0.06	0.06	0.04	**0.06**	0.06	0.04	**0.06**	0.06
	Alg-4-single-split-Iso	0.05	0.06	0.06	0.05	**0.06**	0.06	0.04	**0.06**	0.06
	Alg-4-single-split-Platt	0.04	0.06	0.06	0.05	**0.06**	0.06	0.04	**0.06**	0.06
	Alg-5-full-sample-IVAP	0.04	0.06	0.06	0.05	**0.06**	0.06	0.04	**0.06**	0.06
	Alg-5-full-sample-Iso	0.04	0.06	0.06	0.05	**0.06**	0.06	0.04	**0.06**	0.06
	Alg-5-full-sample-Platt	0.04	0.06	0.06	0.05	**0.06**	0.06	0.04	**0.06**	0.05

DGP 1: n = 2000, p = 20, R2_d = 0.5; DGP 2: n = 2000, p = 3, overlap = 0.5;
DGP 3: n = 2000, p = 4; DGP 4: n = 4000, p = 20, share treated = 0.1; g = LGBM,
lowest RMSEs per DGP and propensity learner are highlighted

Table 3.6: PLR Results

DGP	Method	m = Logit			m = Random Forest			m = LGBM		
		MAE	RMSE	Std. dev.	MAE	RMSE	Std. dev.	MAE	RMSE	Std. dev.
1	Alg-1-Clipped	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-1-Uncalib	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-2-nested-cf-IVAP	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-2-nested-cf-Iso	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-2-nested-cf-Platt	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-3-cf-IVAP	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-3-cf-Iso	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	0.07	0.07
	Alg-3-cf-Platt	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	0.07	0.07
	Alg-4-single-split-IVAP	0.05	**0.06**	0.06	0.14	0.18	0.18	0.12	0.15	0.15
	Alg-4-single-split-Iso	0.05	**0.06**	0.06	0.15	0.20	0.20	0.13	0.16	0.16
	Alg-4-single-split-Platt	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-5-full-sample-IVAP	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
	Alg-5-full-sample-Iso	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	0.07	0.07
	Alg-5-full-sample-Platt	0.05	**0.06**	0.06	0.05	**0.06**	0.06	0.05	**0.06**	0.06
2	Alg-1-Clipped	0.09	**0.11**	0.10	0.12	0.14	0.10	0.10	0.11	0.10
	Alg-1-Uncalib	0.09	**0.11**	0.10	0.12	0.14	0.10	0.10	0.11	0.10
	Alg-2-nested-cf-IVAP	0.09	**0.11**	0.10	0.10	0.12	0.10	0.09	0.11	0.10
	Alg-2-nested-cf-Iso	0.09	**0.11**	0.10	0.10	0.12	0.10	0.09	0.11	0.10
	Alg-2-nested-cf-Platt	0.09	**0.11**	0.10	0.10	0.12	0.10	0.09	0.11	0.10
	Alg-3-cf-IVAP	0.09	**0.11**	0.10	0.10	0.12	0.10	0.09	0.11	0.10
	Alg-3-cf-Iso	0.09	**0.11**	0.11	0.09	**0.11**	0.10	0.09	**0.10**	0.10
	Alg-3-cf-Platt	0.09	**0.11**	0.10	0.11	0.13	0.10	0.09	0.11	0.10
	Alg-4-single-split-IVAP	0.09	**0.11**	0.10	0.23	0.26	0.14	0.13	0.16	0.12
	Alg-4-single-split-Iso	0.09	**0.11**	0.10	0.24	0.27	0.15	0.13	0.16	0.12
	Alg-4-single-split-Platt	0.09	**0.11**	0.10	0.09	**0.11**	0.10	0.09	0.11	0.10
	Alg-5-full-sample-IVAP	0.09	**0.11**	0.10	0.09	**0.11**	0.10	0.09	**0.10**	0.10
	Alg-5-full-sample-Iso	0.09	**0.11**	0.10	0.09	**0.11**	0.10	0.09	**0.10**	0.10
	Alg-5-full-sample-Platt	0.09	**0.11**	0.10	0.09	**0.11**	0.10	0.09	0.11	0.10
3	Alg-1-Clipped	0.06	**0.08**	0.08	0.06	**0.07**	0.07	0.08	0.10	0.07
	Alg-1-Uncalib	0.06	**0.08**	0.08	0.06	**0.07**	0.07	0.08	0.10	0.07
	Alg-2-nested-cf-IVAP	0.06	**0.08**	0.08	0.07	0.09	0.07	0.09	0.11	0.07
	Alg-2-nested-cf-Iso	0.06	**0.08**	0.08	0.07	0.09	0.07	0.09	0.11	0.07
	Alg-2-nested-cf-Platt	0.06	**0.08**	0.08	0.07	0.08	0.07	0.09	0.11	0.07
	Alg-3-cf-IVAP	0.06	**0.08**	0.08	0.06	0.08	0.07	0.06	0.08	0.07
	Alg-3-cf-Iso	0.07	0.09	0.08	0.06	0.08	0.08	0.06	0.08	0.08
	Alg-3-cf-Platt	0.07	**0.08**	0.08	0.10	0.12	0.08	0.07	0.09	0.08
	Alg-4-single-split-IVAP	0.06	**0.08**	0.08	1.01	1.02	0.16	0.43	0.44	0.11
	Alg-4-single-split-Iso	0.06	**0.08**	0.08	1.10	1.11	0.18	0.46	0.47	0.12
	Alg-4-single-split-Platt	0.06	**0.08**	0.08	0.06	0.08	0.08	0.08	0.09	0.07
	Alg-5-full-sample-IVAP	0.06	**0.08**	0.08	0.06	0.08	0.07	0.06	**0.08**	0.07
	Alg-5-full-sample-Iso	0.07	0.09	0.08	0.06	**0.07**	0.07	0.06	**0.07**	0.07
	Alg-5-full-sample-Platt	0.06	**0.08**	0.08	0.06	0.08	0.08	0.06	**0.07**	0.07
4	Alg-1-Clipped	0.08	**0.09**	0.06	0.05	**0.06**	0.05	0.05	**0.06**	0.05
	Alg-1-Uncalib	0.08	**0.09**	0.06	0.05	**0.06**	0.05	0.05	**0.06**	0.05
	Alg-2-nested-cf-IVAP	0.08	**0.09**	0.05	0.07	0.09	0.05	0.07	0.09	0.05
	Alg-2-nested-cf-Iso	0.08	**0.09**	0.06	0.07	0.09	0.05	0.07	0.09	0.05
	Alg-2-nested-cf-Platt	0.08	**0.09**	0.06	0.07	0.09	0.05	0.07	0.09	0.05
	Alg-3-cf-IVAP	0.07	**0.09**	0.06	0.07	0.08	0.05	0.07	0.09	0.05
	Alg-3-cf-Iso	0.08	0.10	0.06	0.08	0.09	0.05	0.08	0.10	0.05
	Alg-3-cf-Platt	0.08	**0.09**	0.05	0.07	0.09	0.05	0.08	0.09	0.06
	Alg-4-single-split-IVAP	0.08	0.10	0.06	0.56	0.59	0.19	0.56	0.58	0.14
	Alg-4-single-split-Iso	0.08	**0.09**	0.06	0.26	0.31	0.17	0.26	0.30	0.16
	Alg-4-single-split-Platt	0.08	**0.09**	0.06	0.07	0.09	0.05	0.08	0.09	0.05
	Alg-5-full-sample-IVAP	0.08	**0.09**	0.06	0.07	0.09	0.05	0.08	0.09	0.05
	Alg-5-full-sample-Iso	0.08	**0.09**	0.06	0.07	0.09	0.05	0.08	0.09	0.05
	Alg-5-full-sample-Platt	0.08	**0.09**	0.06	0.07	0.09	0.05	0.08	0.09	0.05

DGP 1: n = 2000, p = 20, R2_d = 0.5; DGP 2: n = 2000, p = 3, overlap = 0.5;
DGP 3: n = 2000, p = 4; DGP 4: n = 4000, p = 20, share treated = 0.1; g = LGBM,
lowest RMSEs per DGP and propensity learner are highlighted

3.6.8 Sensitivity Analysis

The following figures display the results over 100 repetitions for the ATE estimators. Additionally, *Oracle* estimates are provided, utilizing true propensity scores from the DGPs. Furthermore, the figures with respect to the sample size and number of covariates include the estimators with clipped propensity scores. The combination of clipping and calibration is also displayed in histogram plots. If no clipping is employed, the clipping threshold is set to 1^{-12} to ensure stability.

Table 3.7: Overview of Simulated Settings per DGP

DGP	Parameter	Values
1	n	[100, 200, 500, 1000, 2000]
	p	[5, 20, 50, 100, 200]
	R_d^2	[0.2, 0.5, 0.8]
	Clip	[1e-12, 0.01, 0.1]
2	n	[200, 500, 1000, 2000, 4000]
	p	3
	Overlap	[0.1, 0.5, 0.9]
	Clip	[1e-12, 0.01, 0.1]
3	n	[200, 500, 1000, 2000, 4000]
	p	4
	Clip	[1e-12, 0.01, 0.1]
4	n	[2000, 4000, 6000, 8000]
	p	20
	$\mathsf{E}[D]$	[0.05, 0.1, 0.2]
	Clip	[1e-12, 0.01, 0.1]

Propensity Scores by DGP

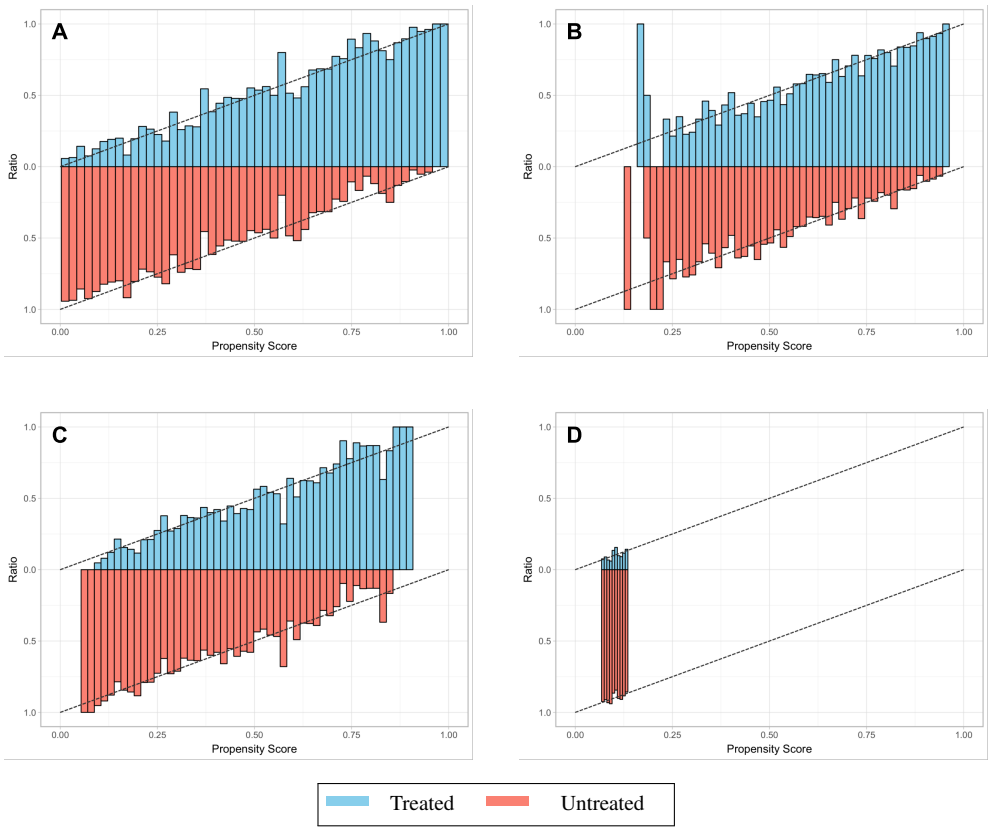

Figure 3.7: Underlying true propensity scores, divided by treatment allocation. The DGPs presented are: Panel A - DGP 1 (IRM), Panel B - DGP 2 (Drug), Panel C - DGP 3 (Nonlinear), and Panel D - DGP 4 (Unbalanced).

Propensity Learner

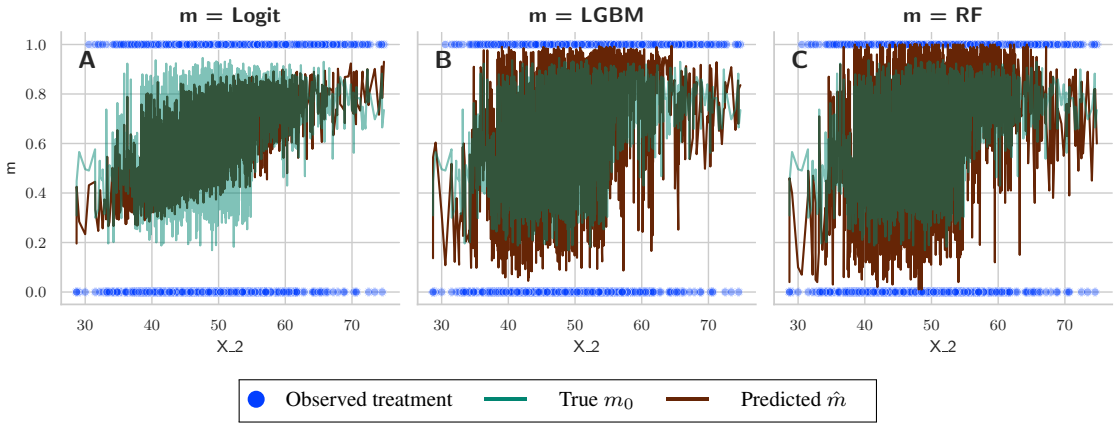

Figure 3.8: DGP 2 Drug, Overlap = 0.5, n = 2000, p = 3

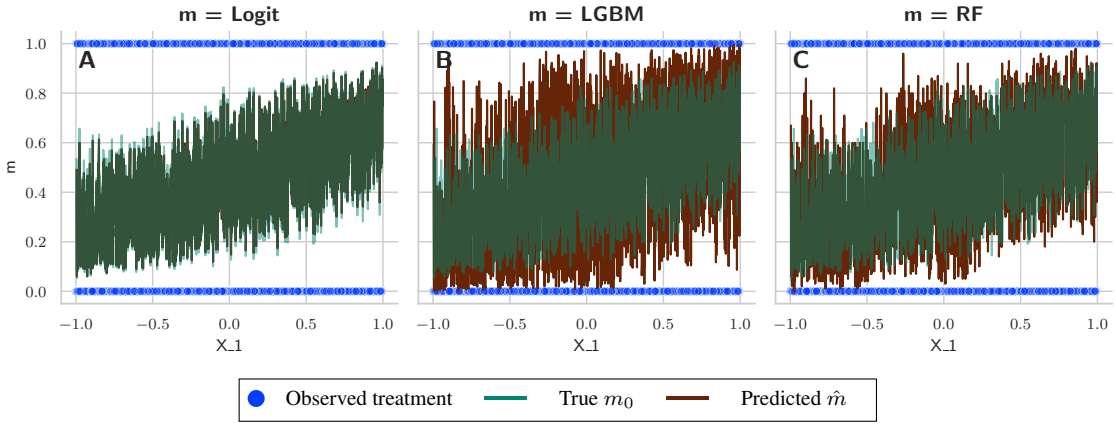

Figure 3.9: DGP 3 Nonlinear, n = 2000, p = 4

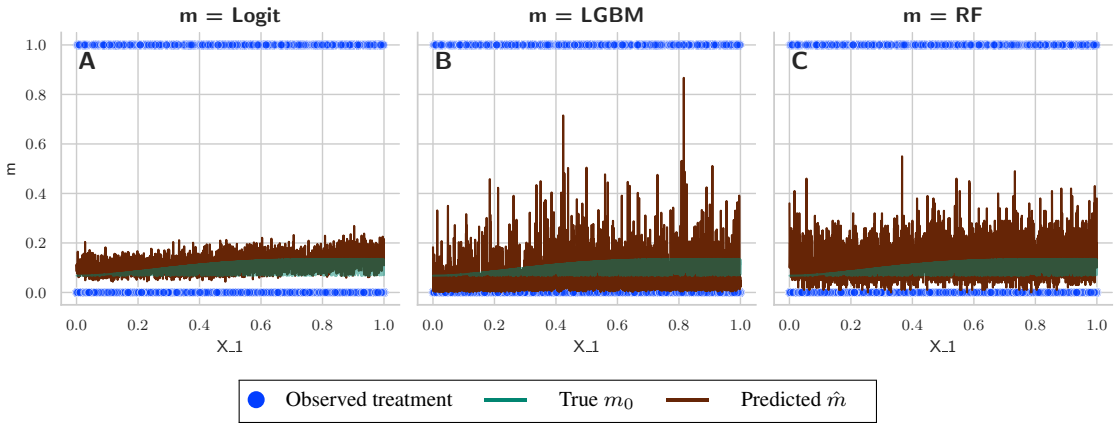

Figure 3.10: DGP 4 Unbalanced, α = 0.1, p = 20

Calibrated Propensity Scores

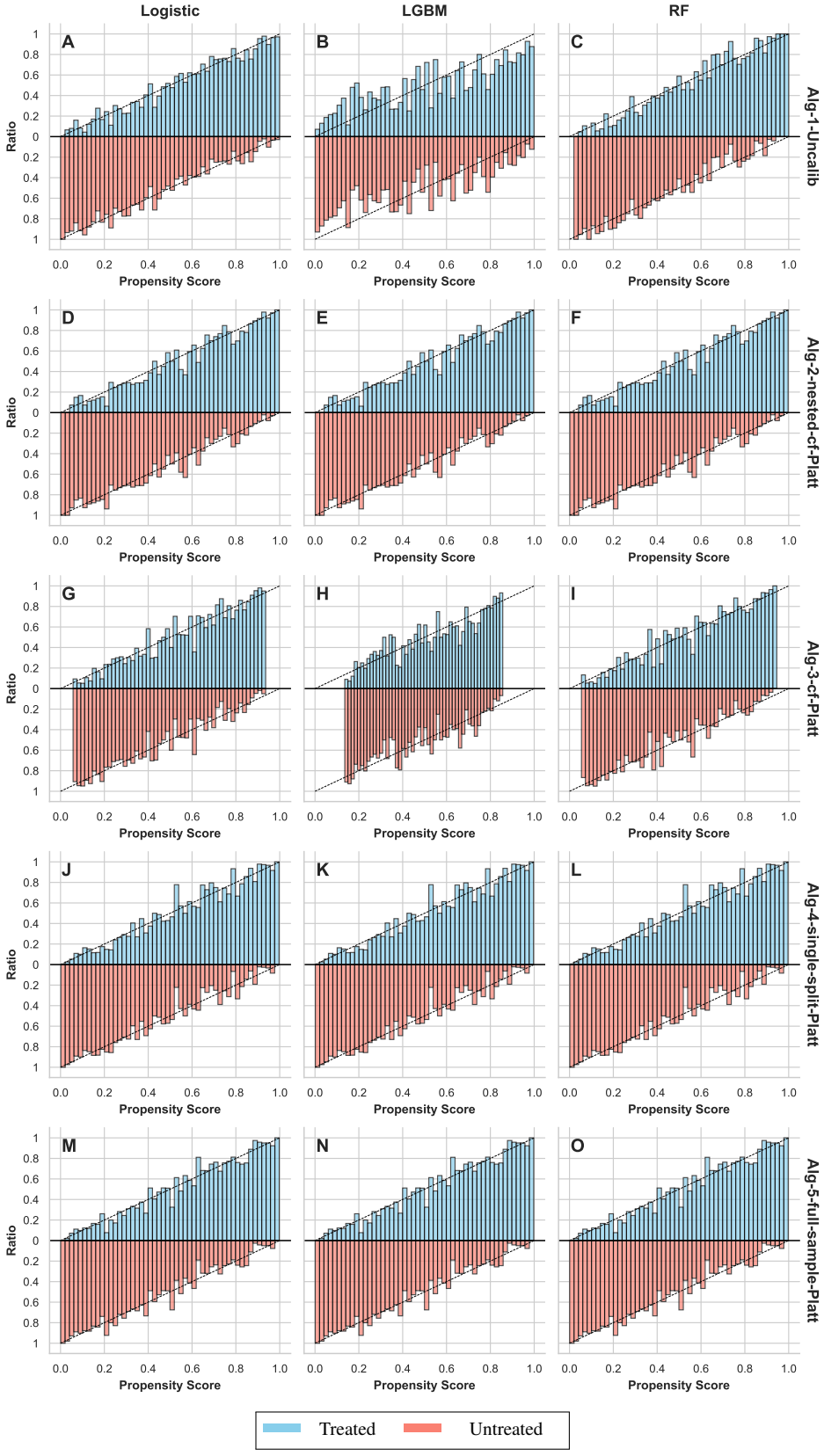

Figure 3.11: DGP 1 IRM, R2D = 0.5, m = LGBM, g = LGBM, n = 2000, p = 20, Calibration method for Algorithms 2-4: Platt Scaling

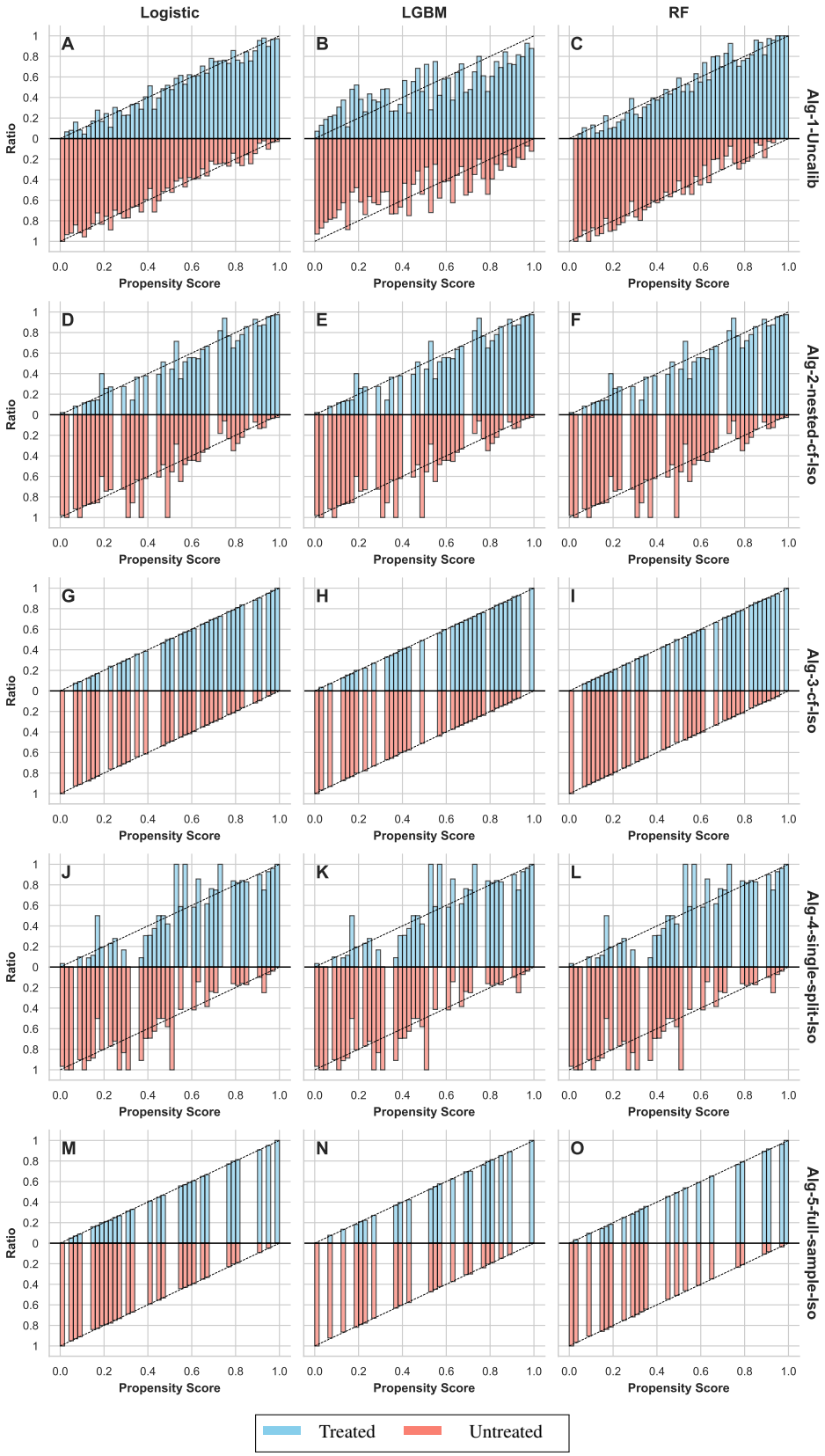

Figure 3.12: DGP 1 IRM, R2D = 0.5, m = LGBM, g = LGBM, n = 2000, p = 20, Calibration method for Algorithms 2-5: isotonic regression

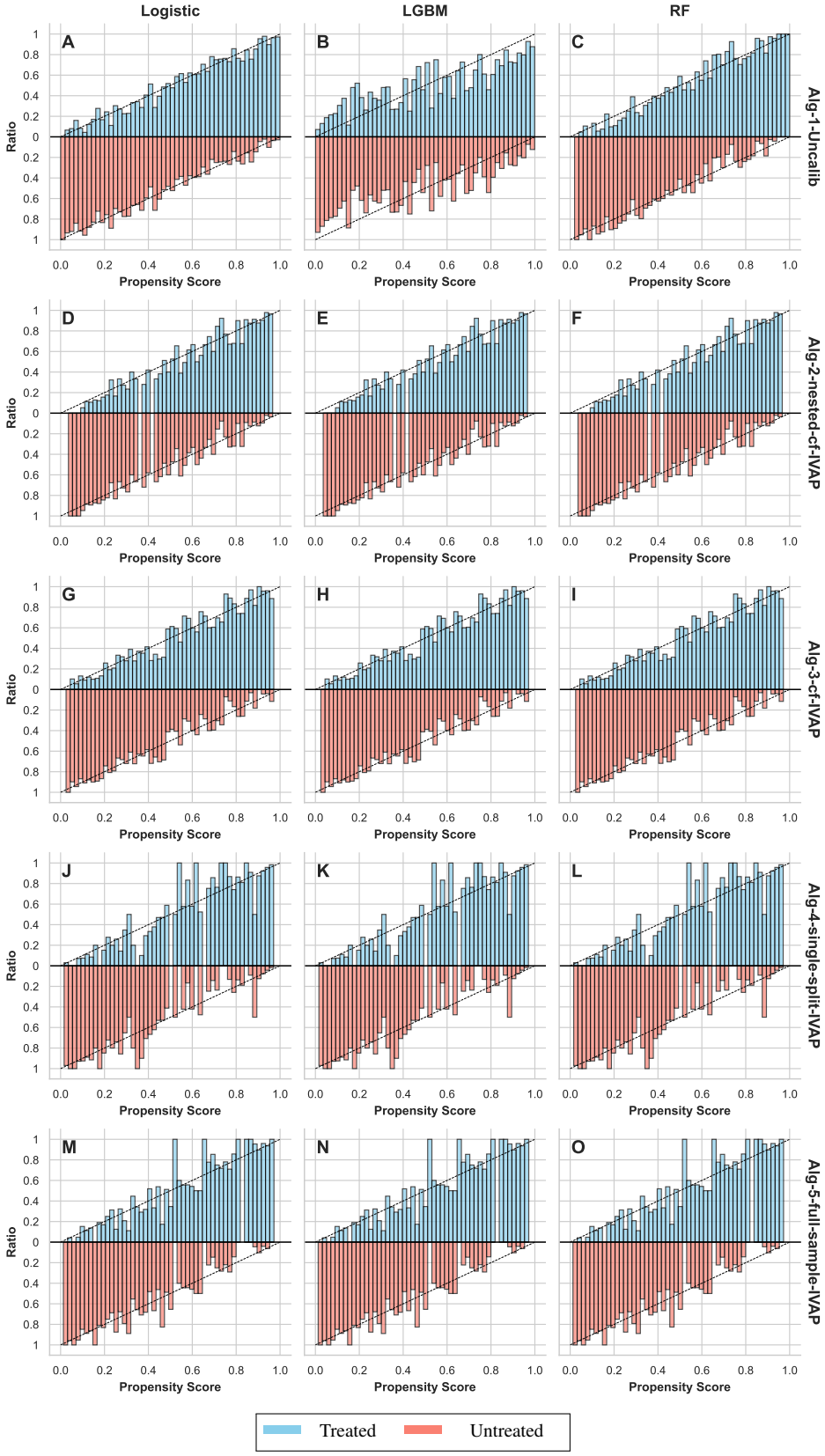

Figure 3.13: DGP 1 IRM, R2D = 0.5, m = LGBM, g = LGBM, n = 2000, p = 20, Calibration method for Algorithms 2-4: Venn-ABERS

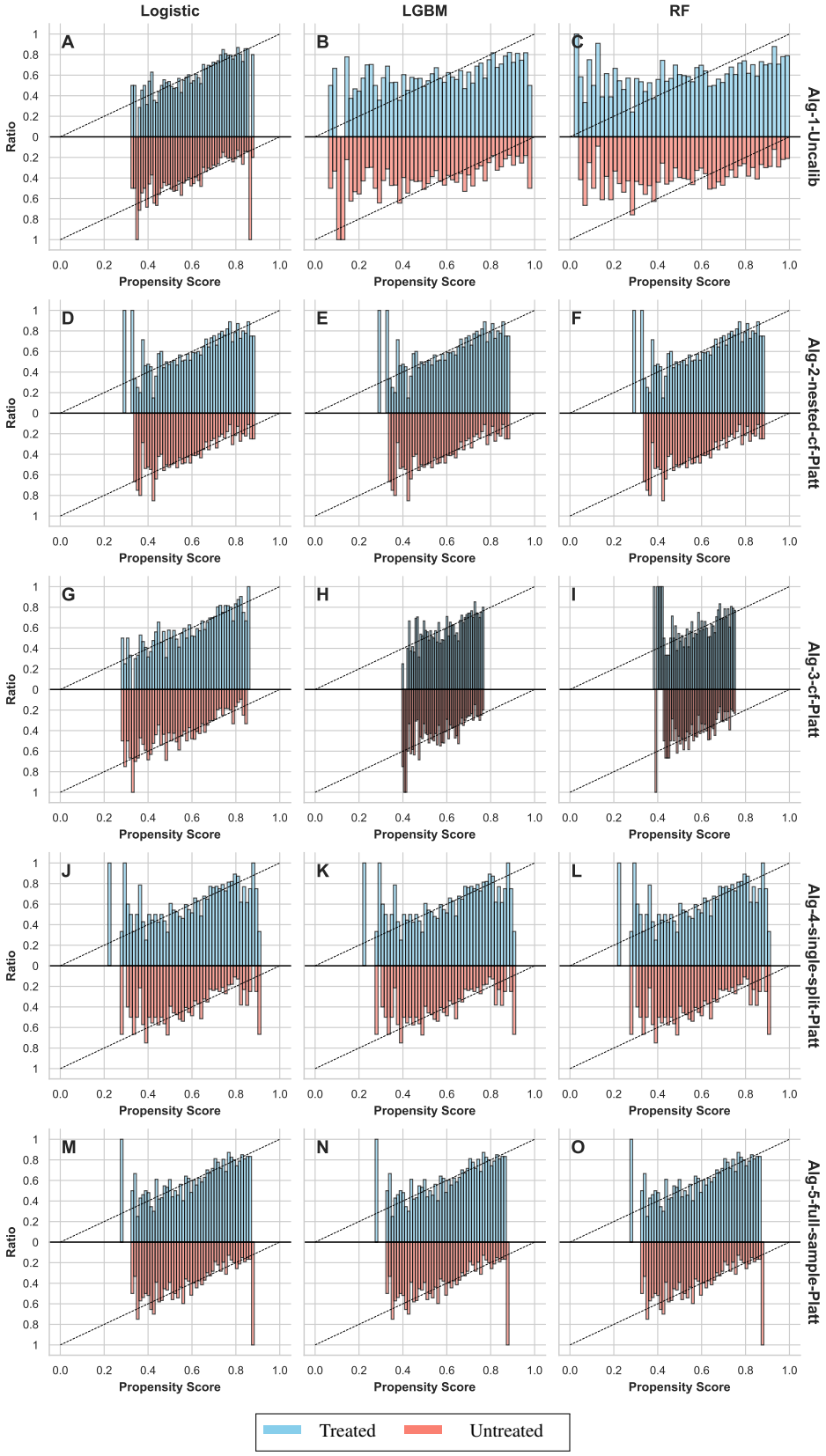

Figure 3.14: DGP 2 Drug, Overlap = 0.5, m = LGBM, g = LGBM, n = 2000, p = 3, Calibration method for Algorithms 2-4: Platt Scaling

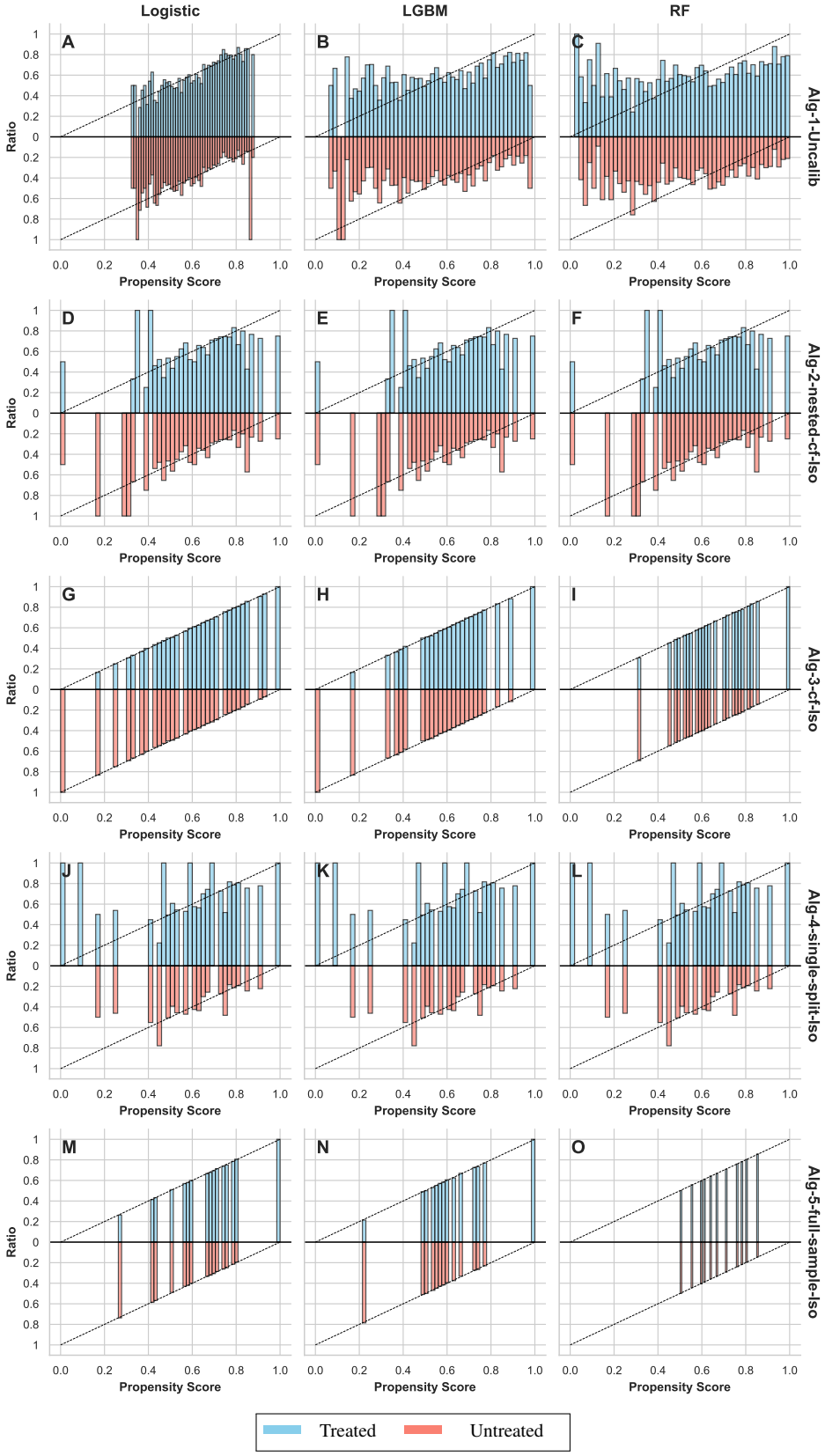

Figure 3.15: DGP 2 Drug, Overlap = 0.5, m = LGBM, g = LGBM, n = 2000, p = 3, Calibration method for Algorithms 2-5: isotonic regression

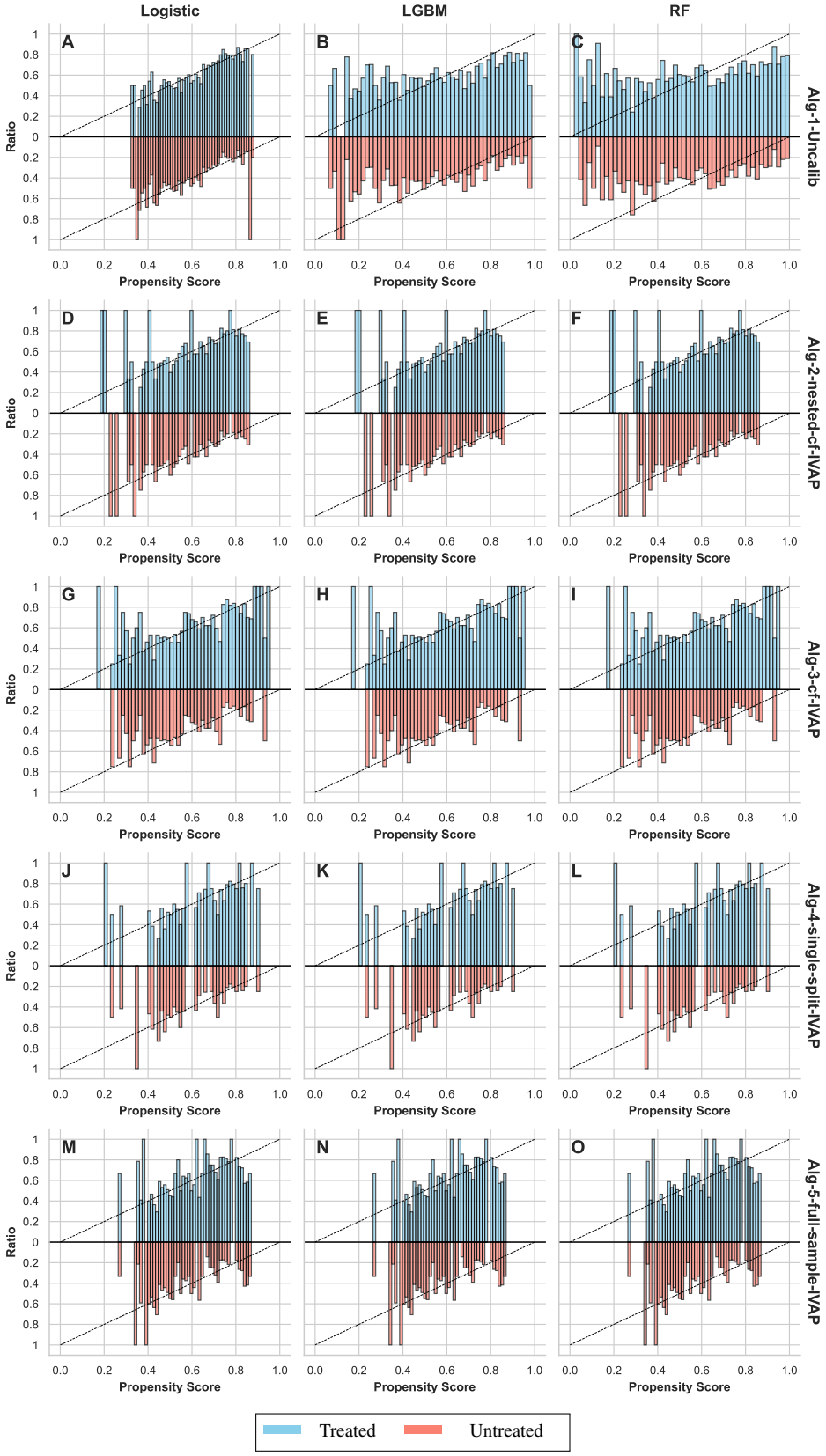

Figure 3.16: DGP 2 Drug, Overlap = 0.5, m = LGBM, g = LGBM, n = 2000, p = 3, Calibration method for Algorithms 2-4: Venn-ABERS

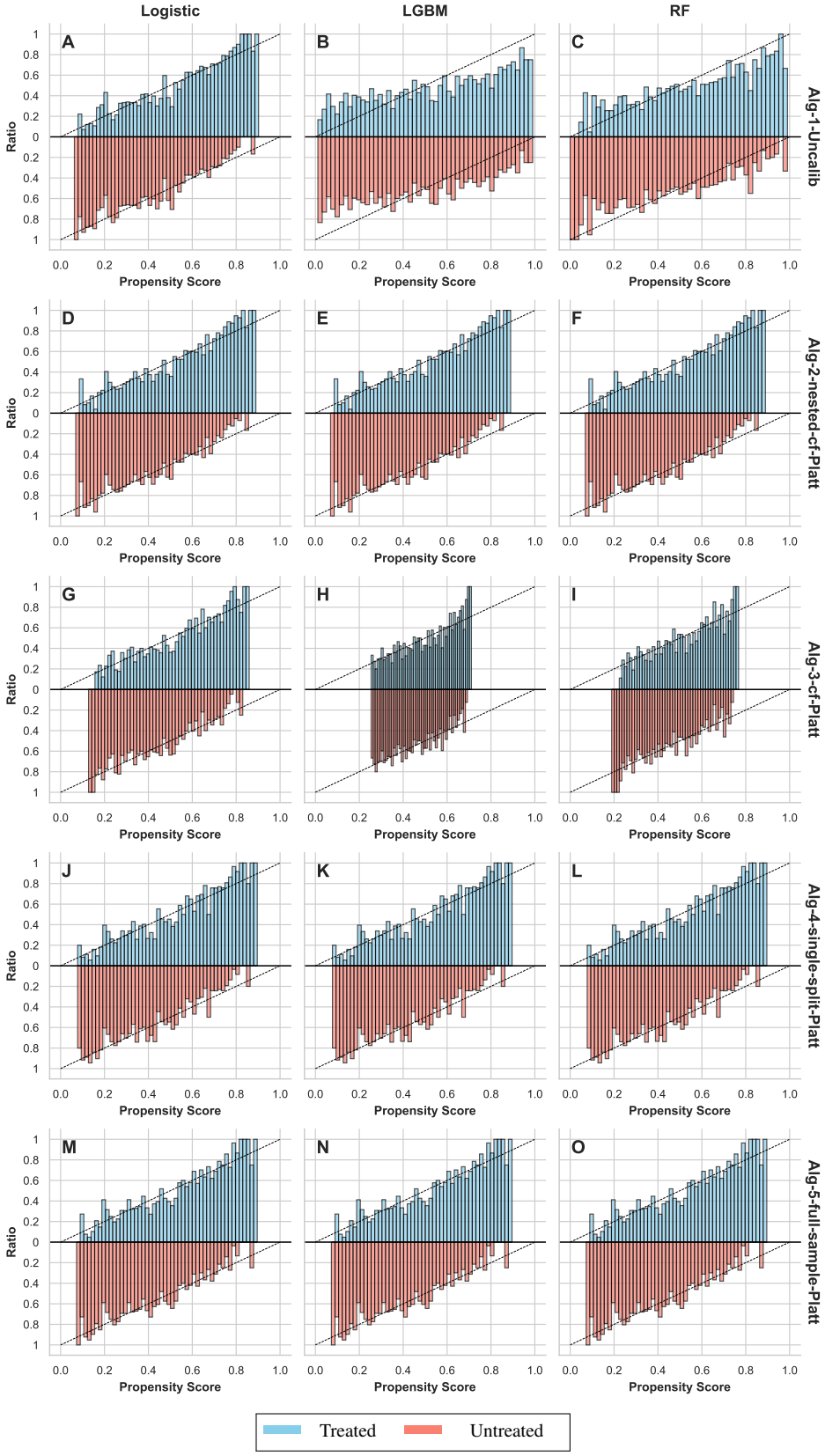

Figure 3.17: DGP 3 Nonlinear, m = LGBM, g = LGBM, n = 2000, p = 4, Calibration method for Algorithms 2-4: Platt Scaling

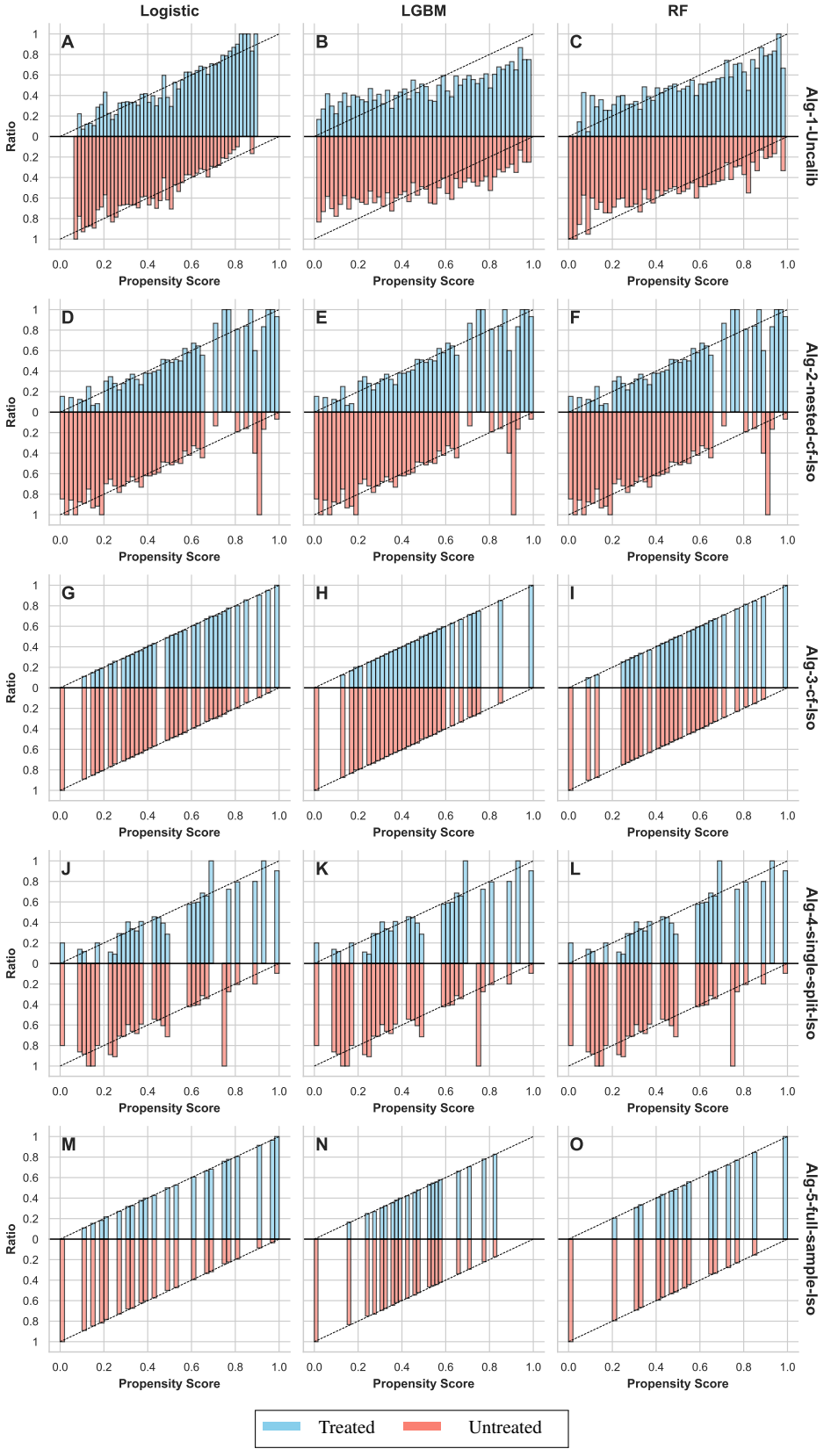

Figure 3.18: DGP 3 Nonlinear, m = LGBM, g = LGBM, n = 2000, p = 4, Calibration method for Algorithms 2-5: isotonic regression

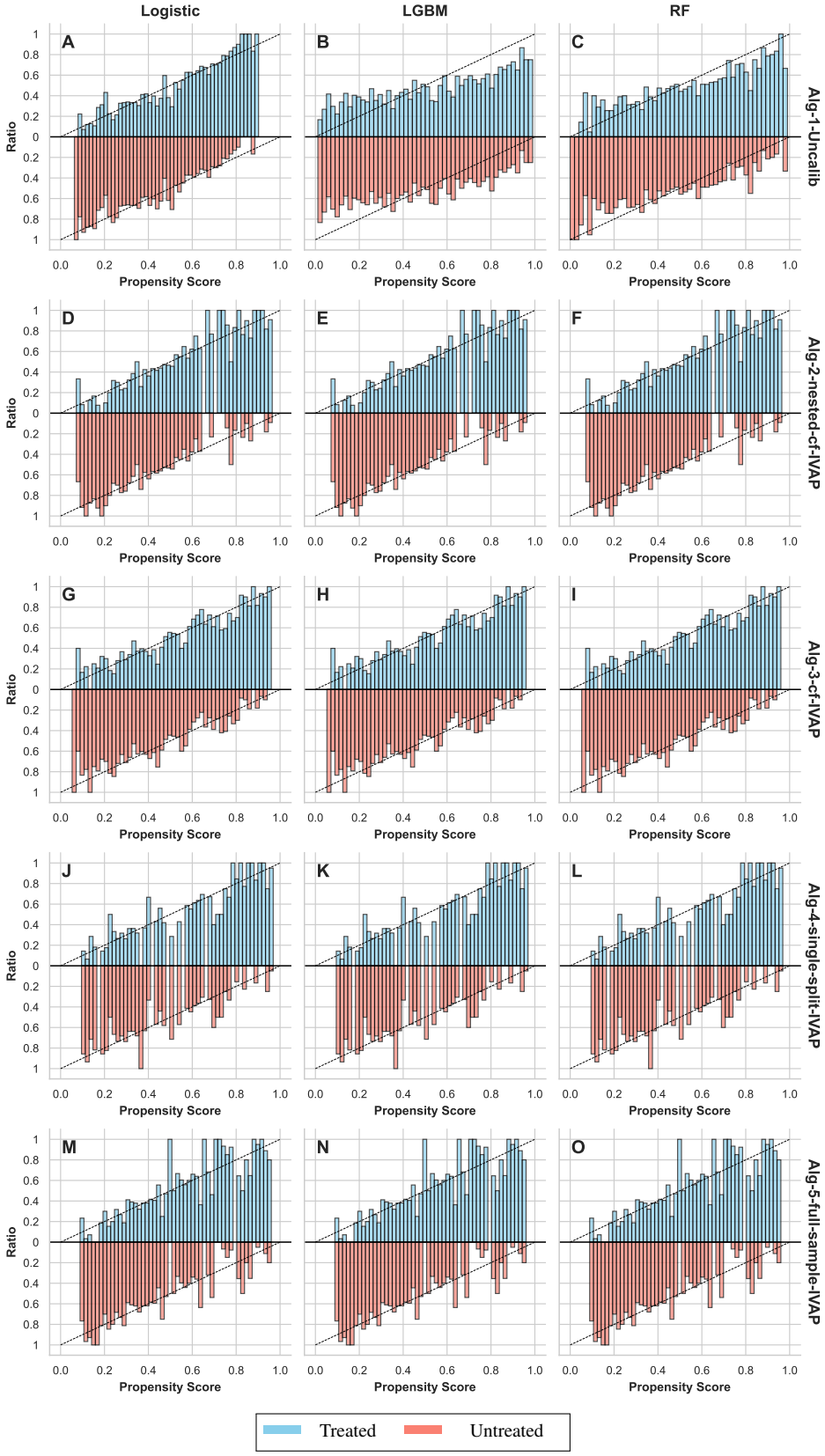

Figure 3.19: DGP 3 Nonlinear, m = LGBM, g = LGBM, n = 2000, p = 4, Calibration method for Algorithms 2-4: Venn-ABERS

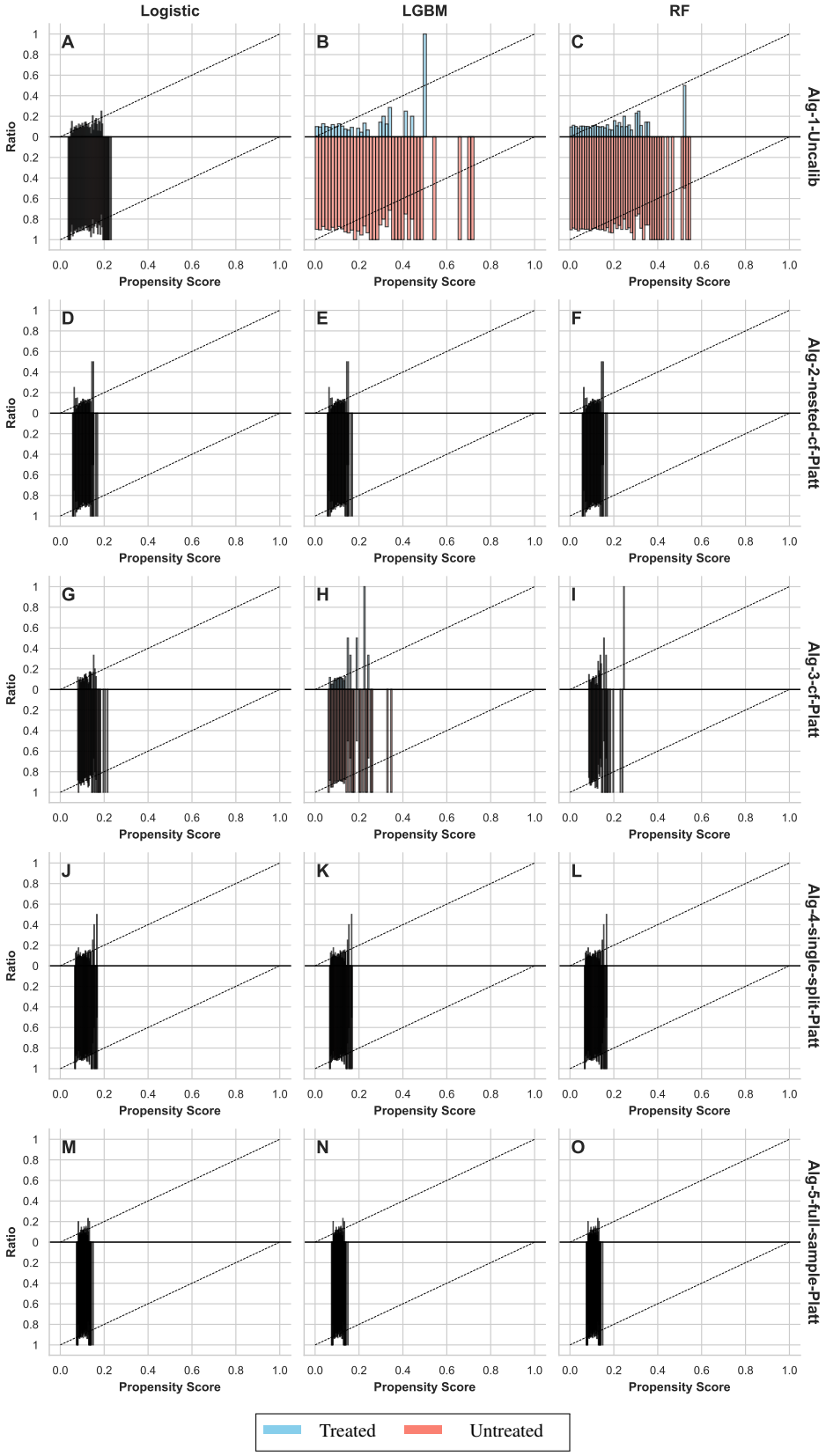

Figure 3.20: DGP 4 Unbalanced, $\alpha = 0.1$, m = LGBM, g = LGBM, n = 4000, p = 20, Calibration method for Algorithms 2-4: Platt Scaling

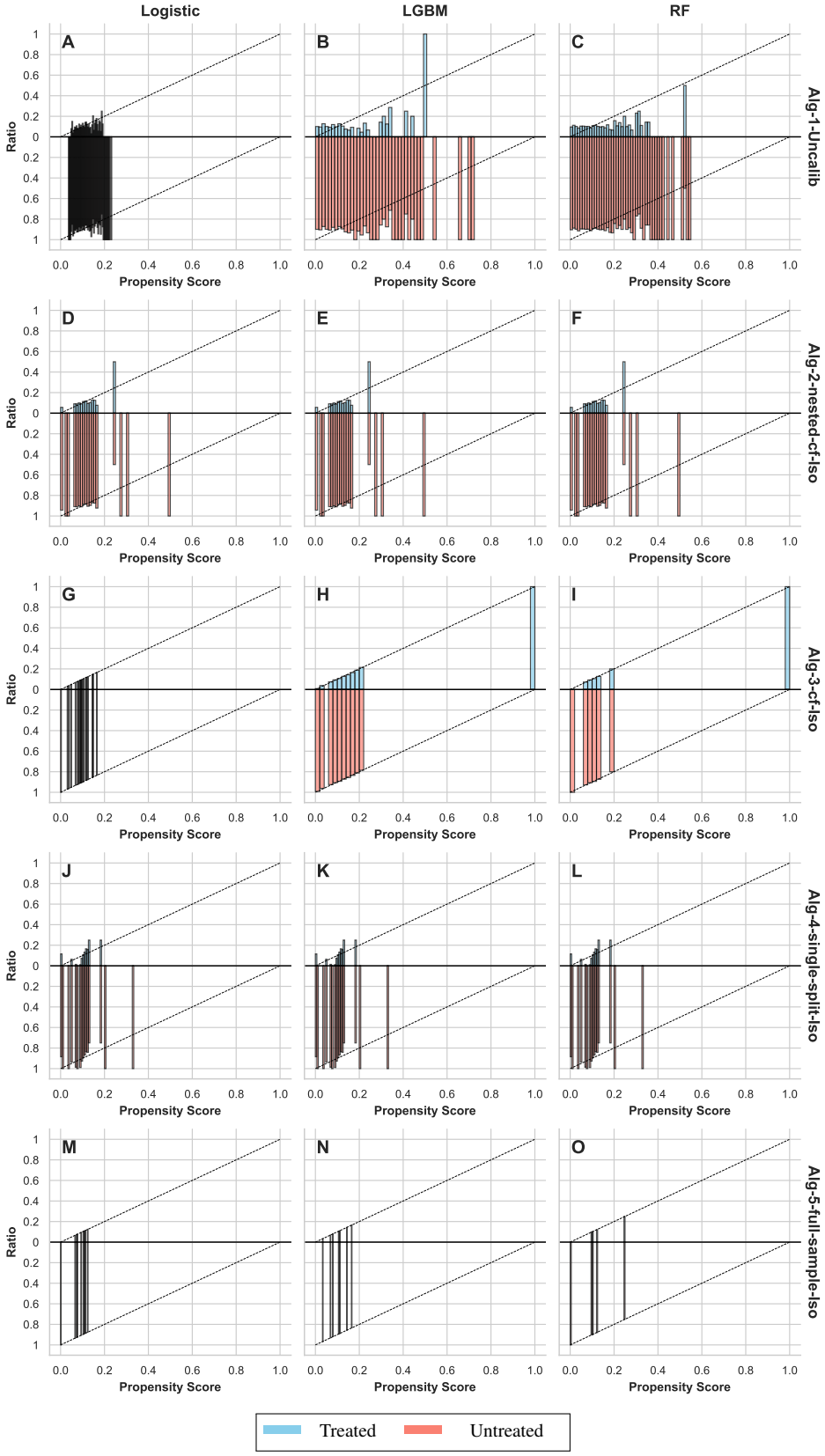

Figure 3.21: DGP 4 Unbalanced, α = 0.1, m = LGBM, g = LGBM, n = 4000, p = 20, Calibration method for Algorithms 2-5: isotonic regression

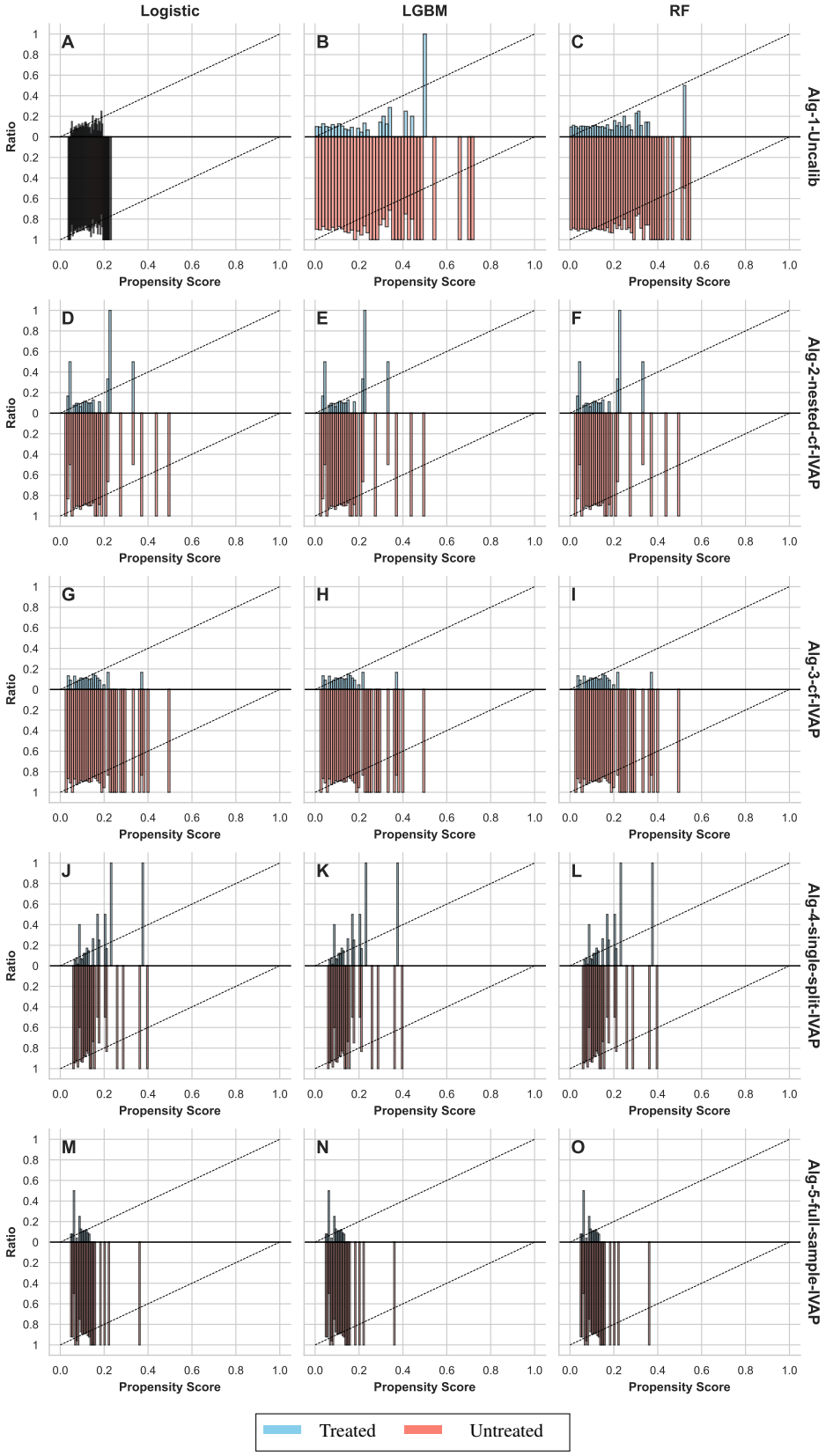

Figure 3.22: DGP 4 Unbalanced, α = 0.1, m = LGBM, g = LGBM, n = 4000, p = 20, Calibration method for Algorithms 2-4: Venn-ABERS

Calibration Errors

DGP 1 IRM

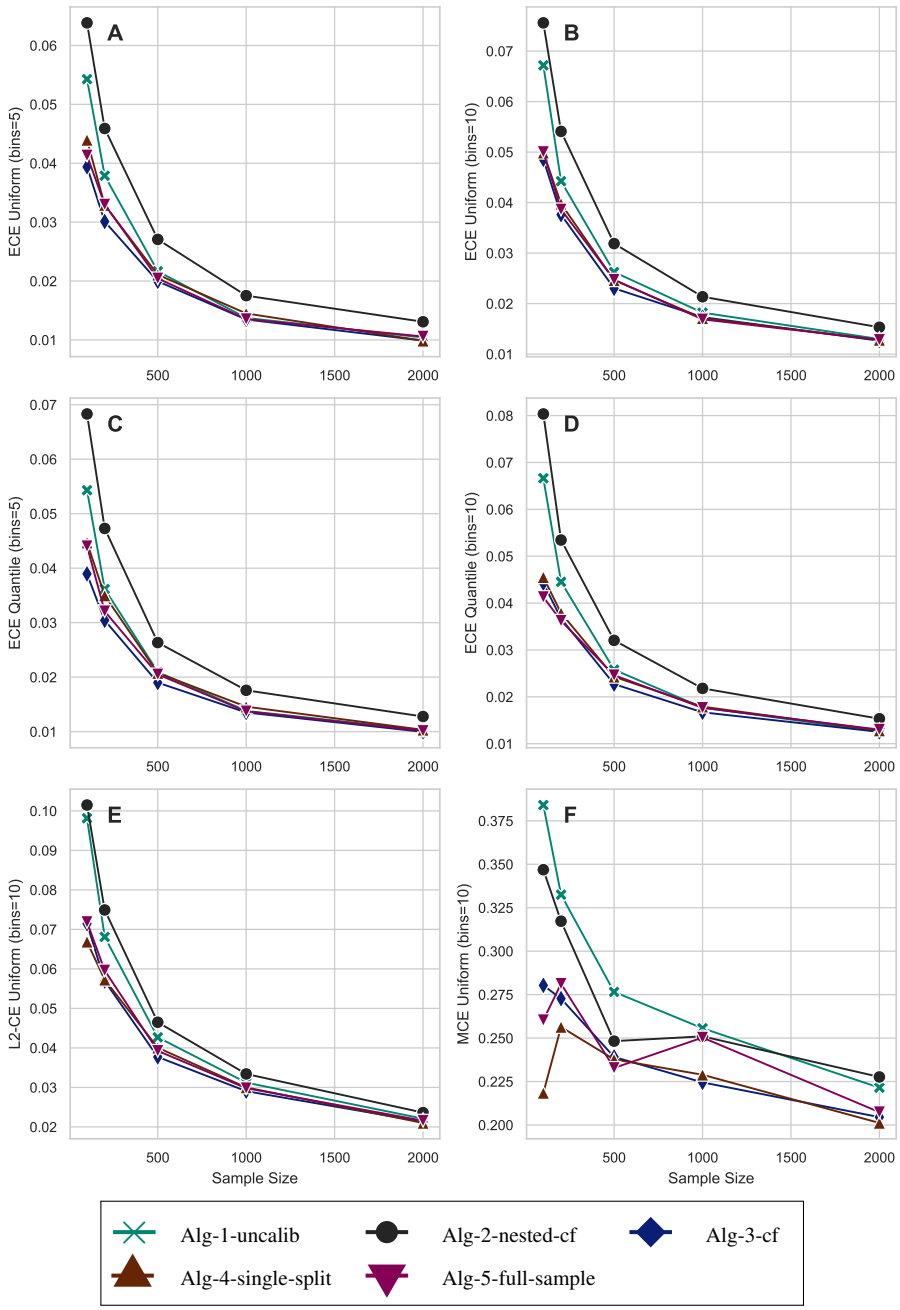

Figure 3.23: DGP 1 IRM, R2D = 0.5, m = Logit, p = 20, Clipping threshold for Algorithms 1,2,4 = 0.01, Calibration method for Algorithms 2-4: isotonic regression

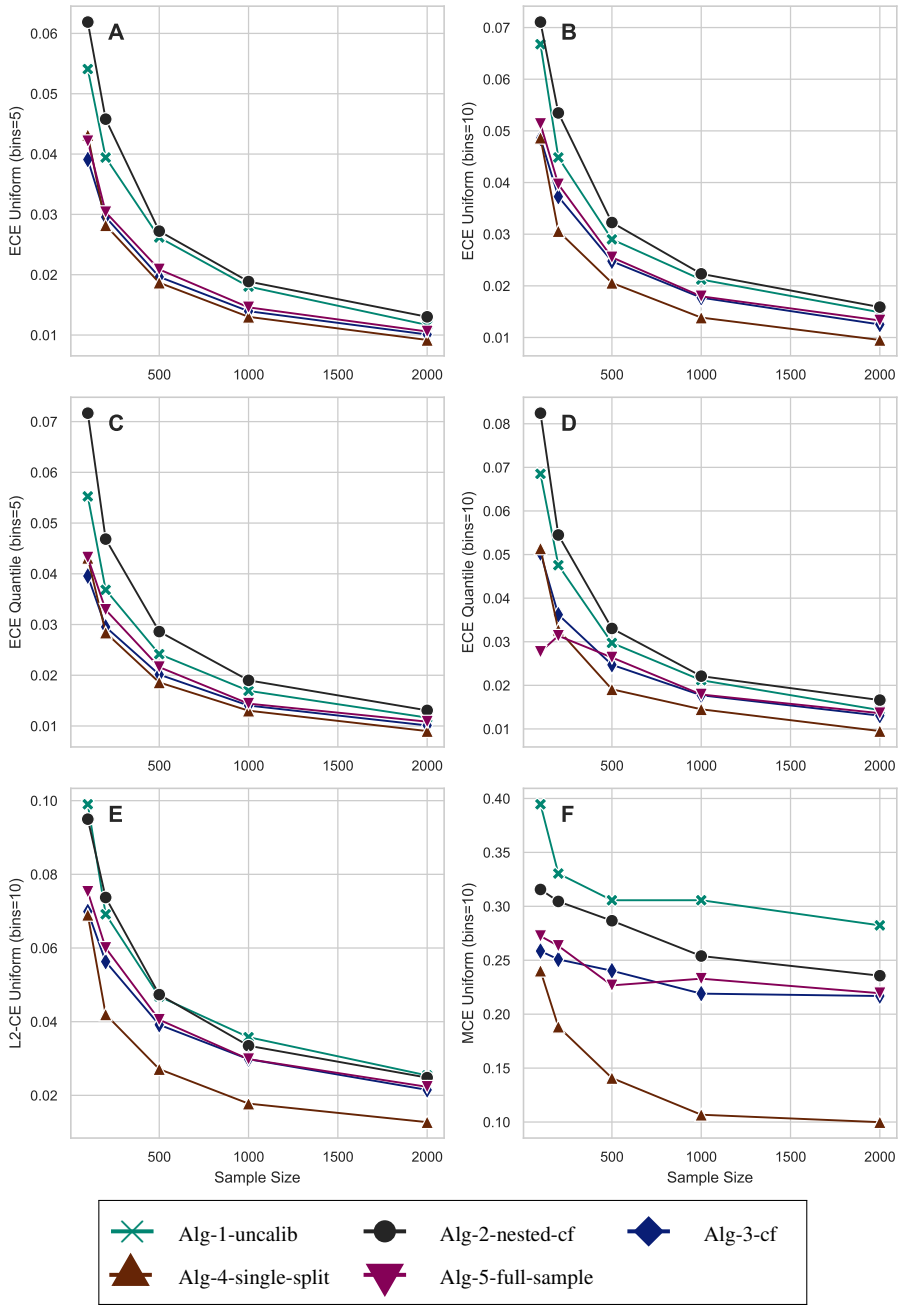

Figure 3.24: DGP 1 IRM, R2D = 0.5, m = LGBM, p = 20, Calibration method for Algorithms 2-5: isotonic regression

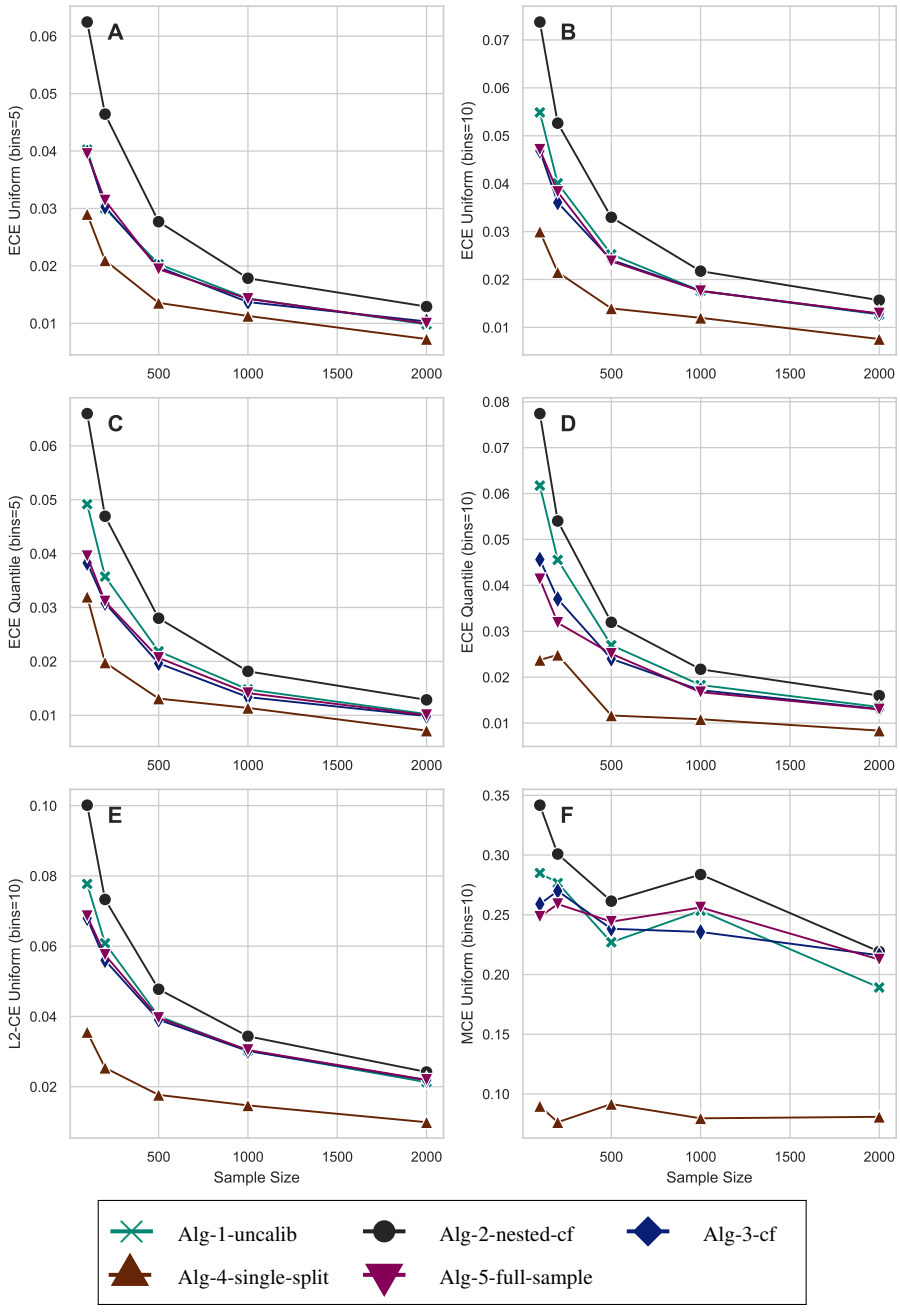

Figure 3.25: DGP 1 IRM, R2D = 0.5, m = RF, p = 20, Calibration method for Algorithms 2-5: isotonic regression

DGP 2 Drug

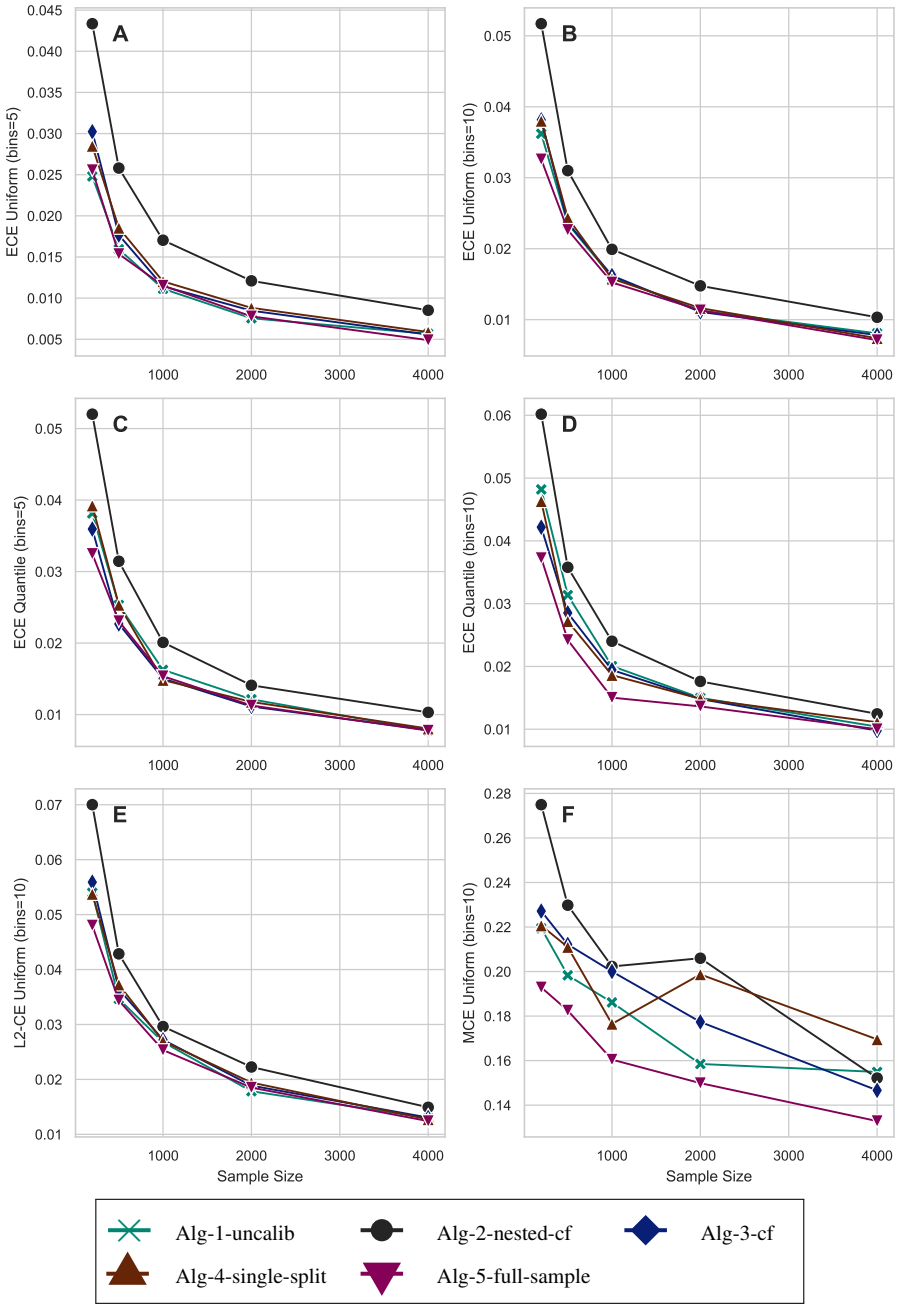

Figure 3.26: DGP 2 Drug, Overlap = 0.5, m = Logit, n = 2000, p = 3, Calibration method for Algorithms 2-4: isotonic regression

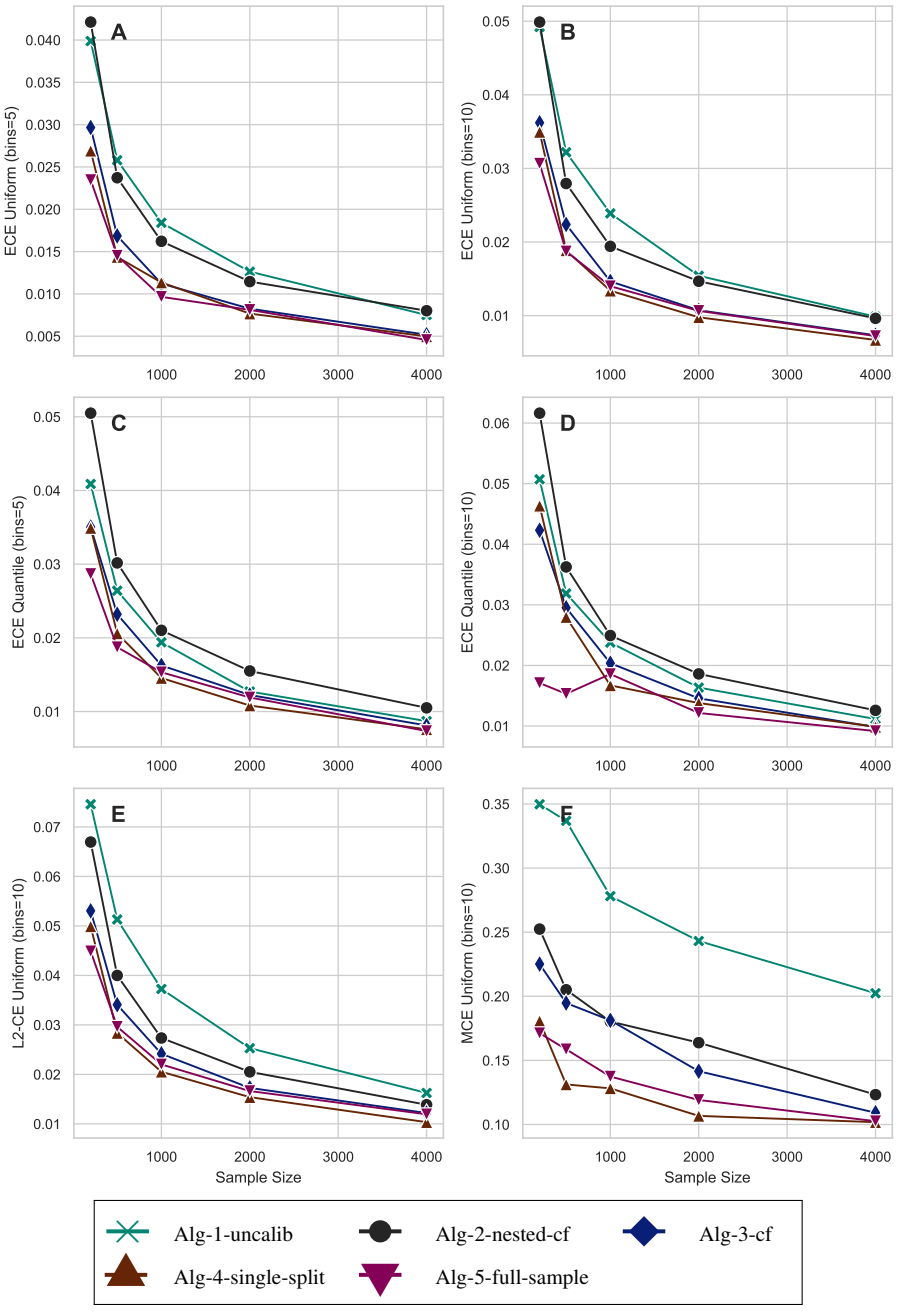

Figure 3.27: DGP 2 Drug, Overlap = 0.5, m = LGBM, p = 3, Calibration method for Algorithms 2-5: isotonic regression

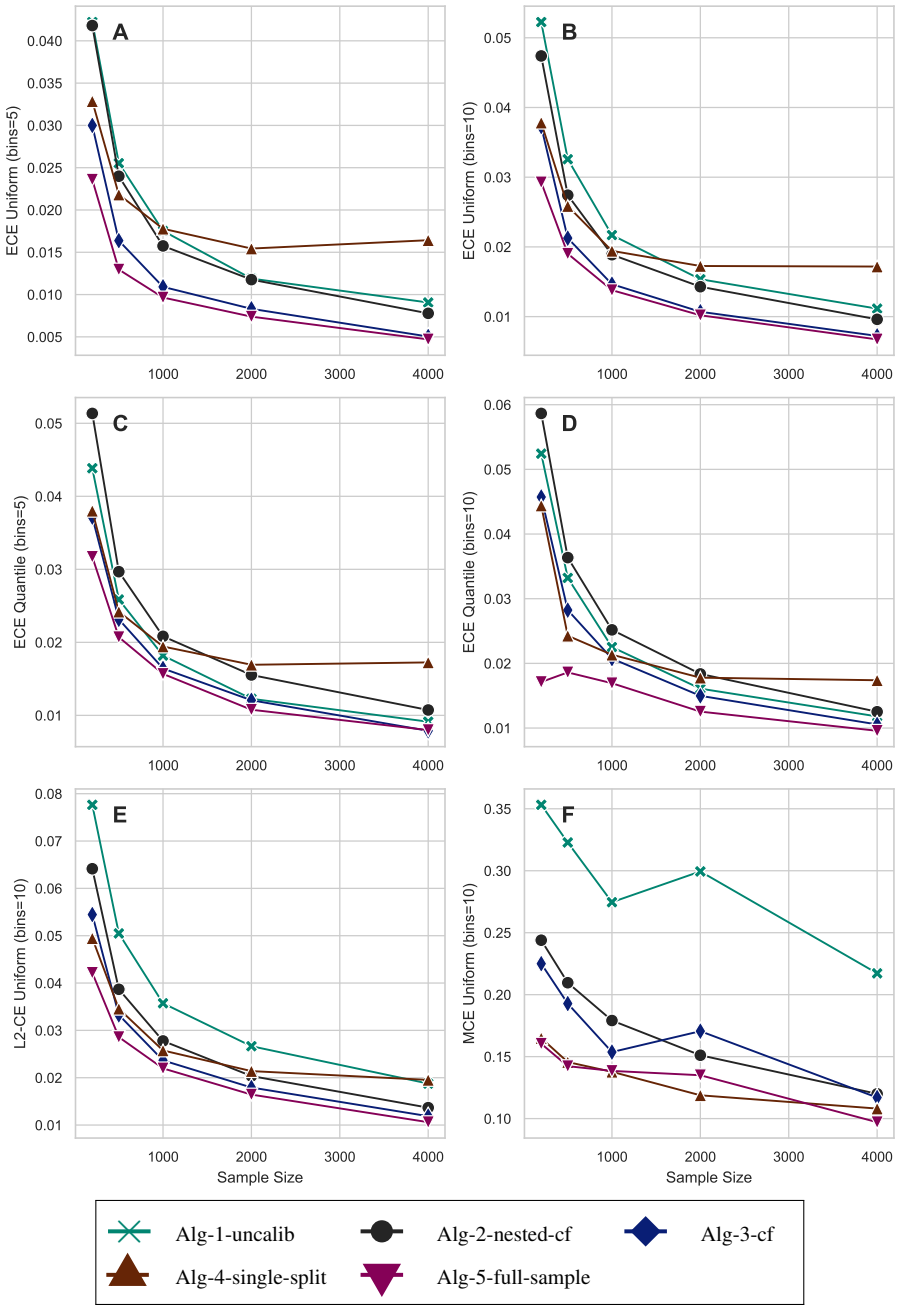

Figure 3.28: DGP 2 Drug, Overlap = 0.5, m = RF, p = 3, Calibration method for Algorithms 2-5: isotonic regression

Nonlinear

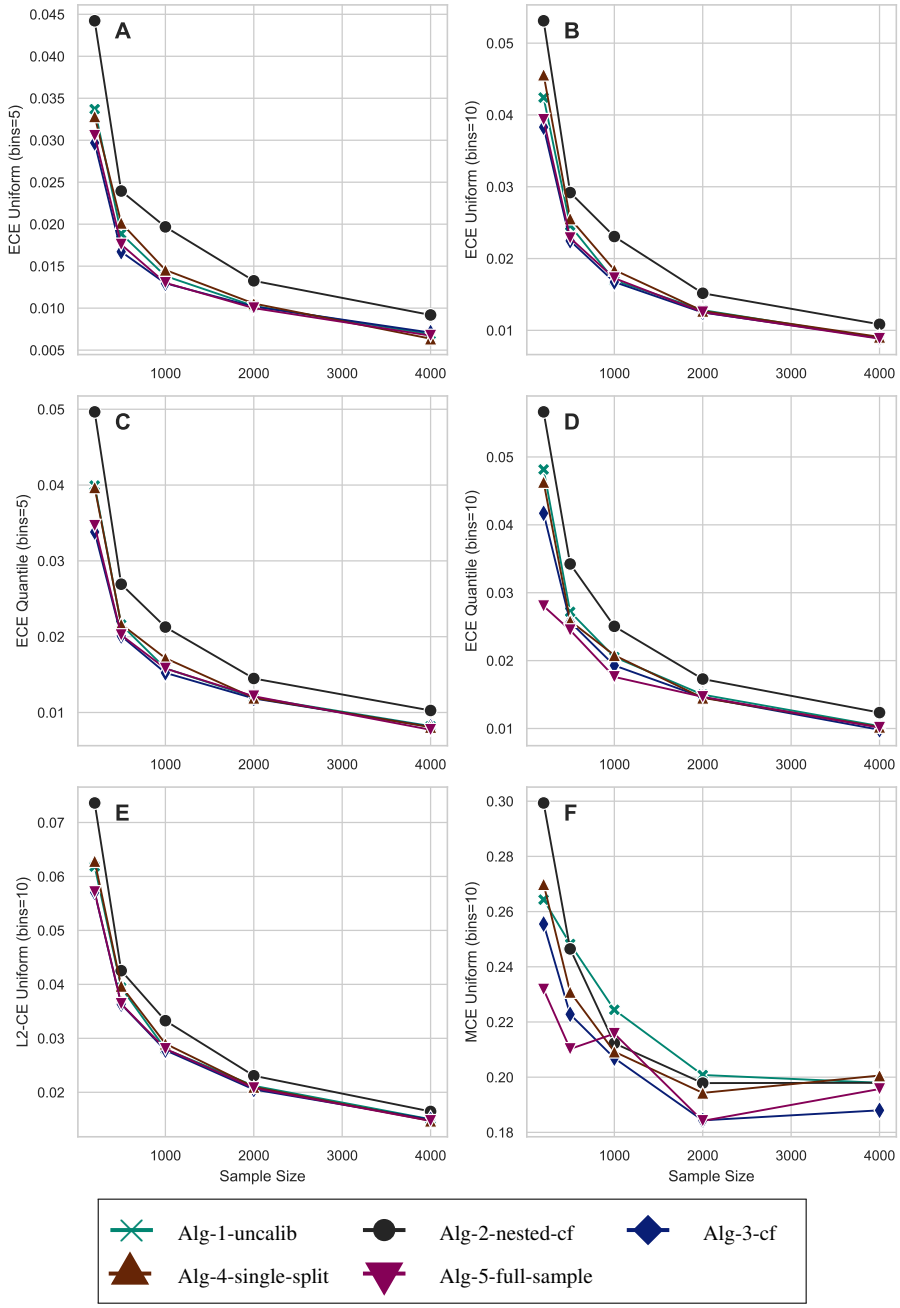

Figure 3.29: DGP 3 Nonlinear, m = Logit, p = 4, Calibration method for Algorithms 2-4: isotonic regression

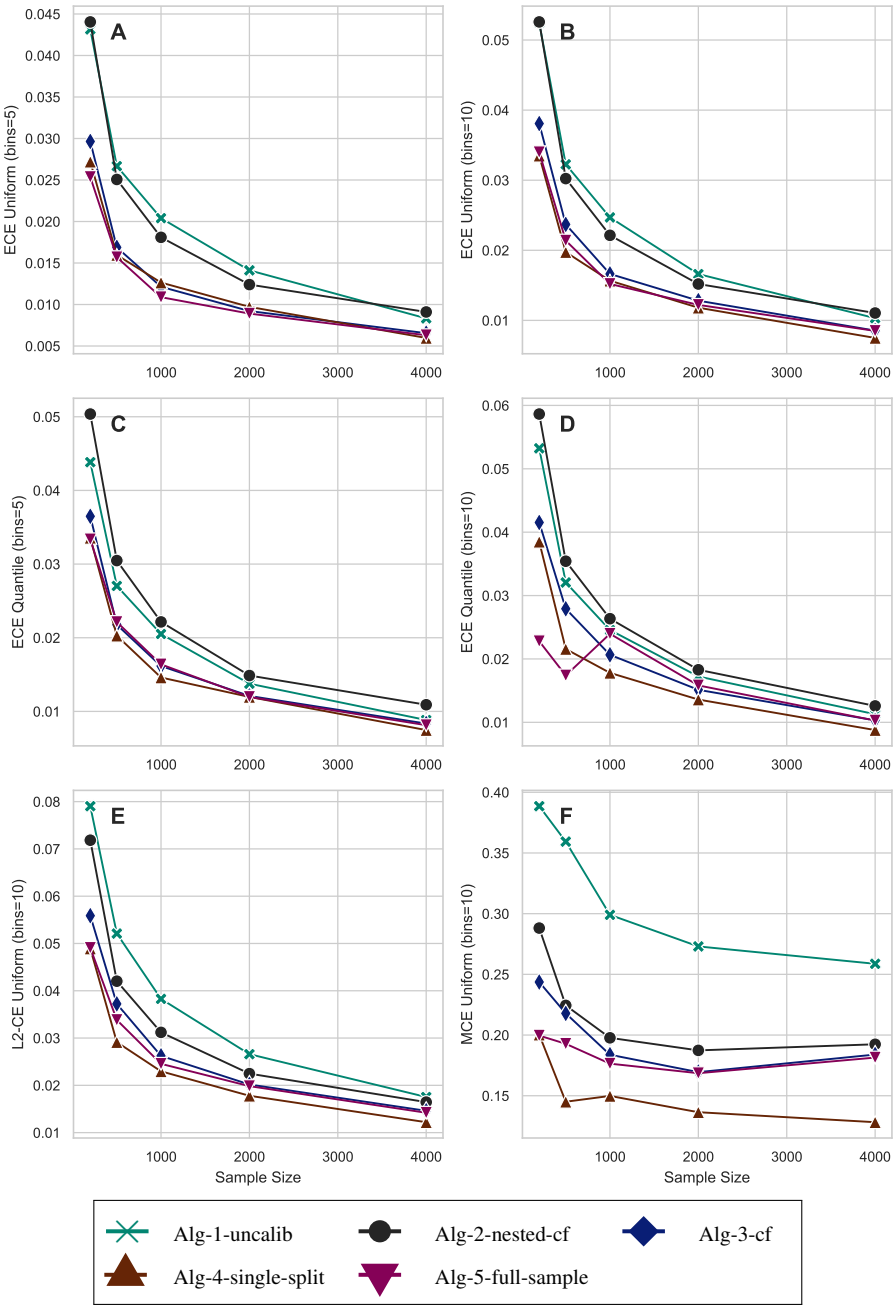

Figure 3.30: DGP 3 Nonlinear, m = LGBM, p = 4, Calibration method for Algorithms 2-5: isotonic regression

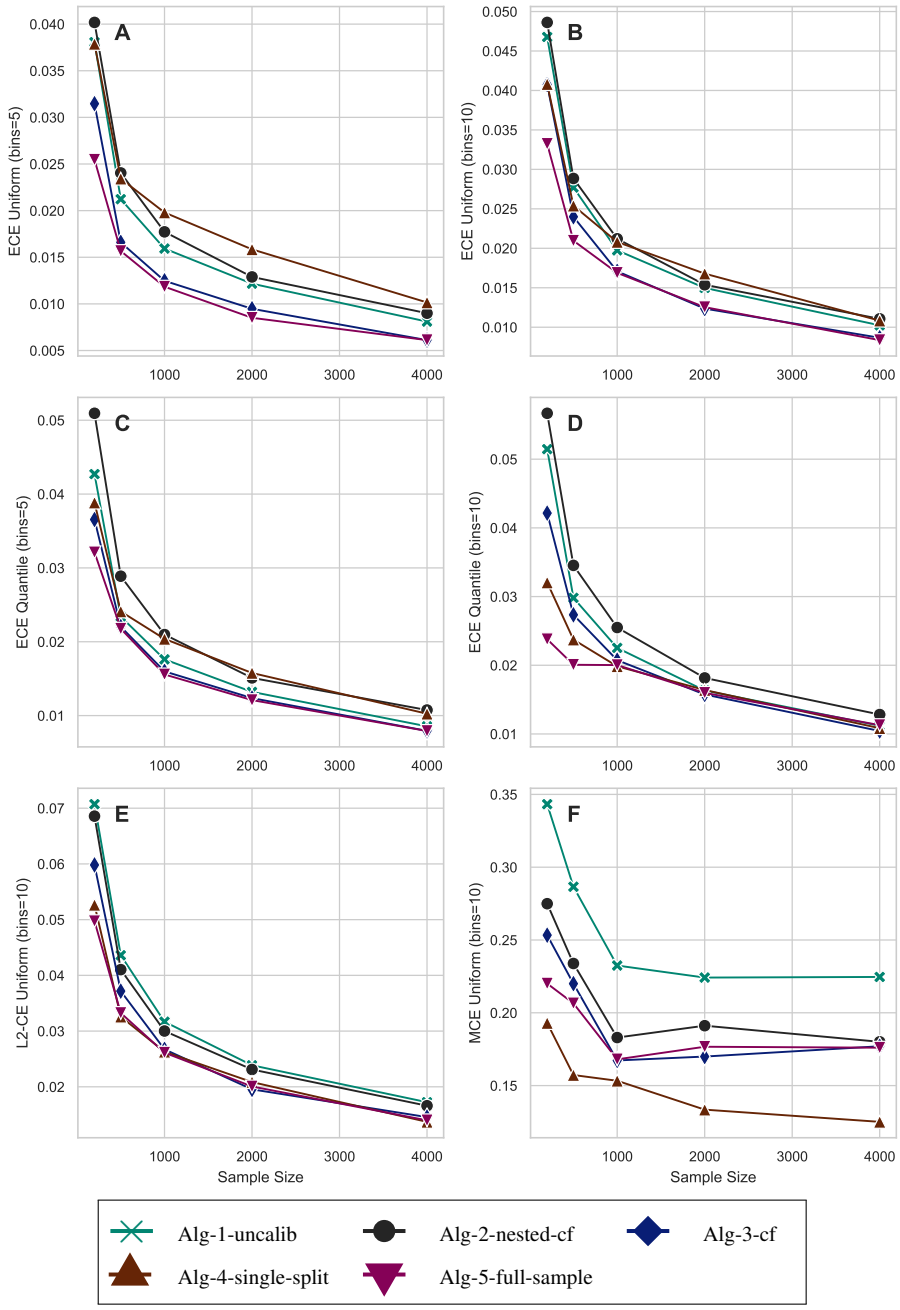

Figure 3.31: DGP 3 Nonlinear, m = RF, p = 4, Calibration method for Algorithms 2-5: isotonic regression

DGP 4 Unbalanced

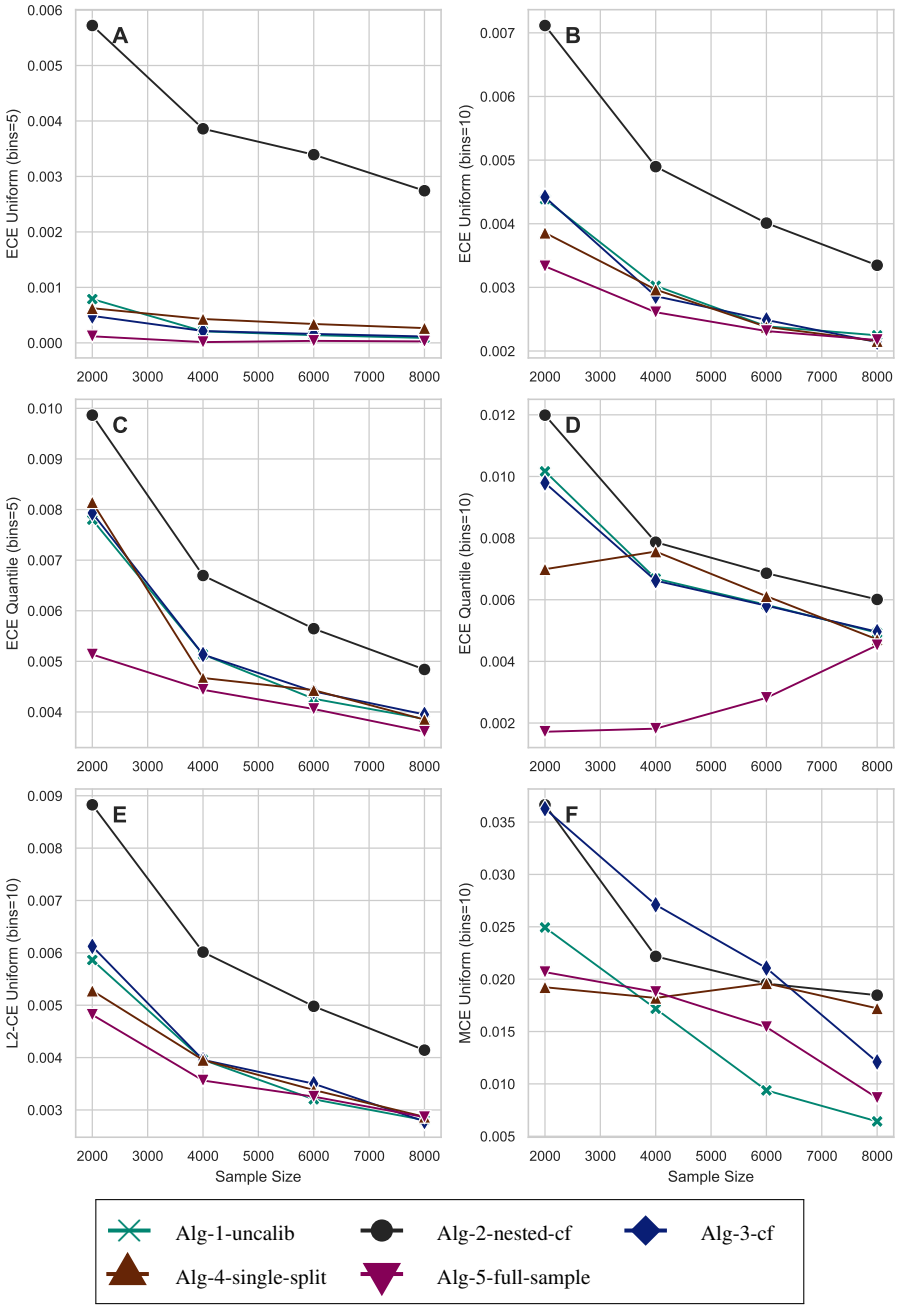

Figure 3.32: DGP 4 Unbalanced, Share_treated = 0.1, m = Logit, p = 20, Calibration method for Algorithms 2-4: isotonic regression

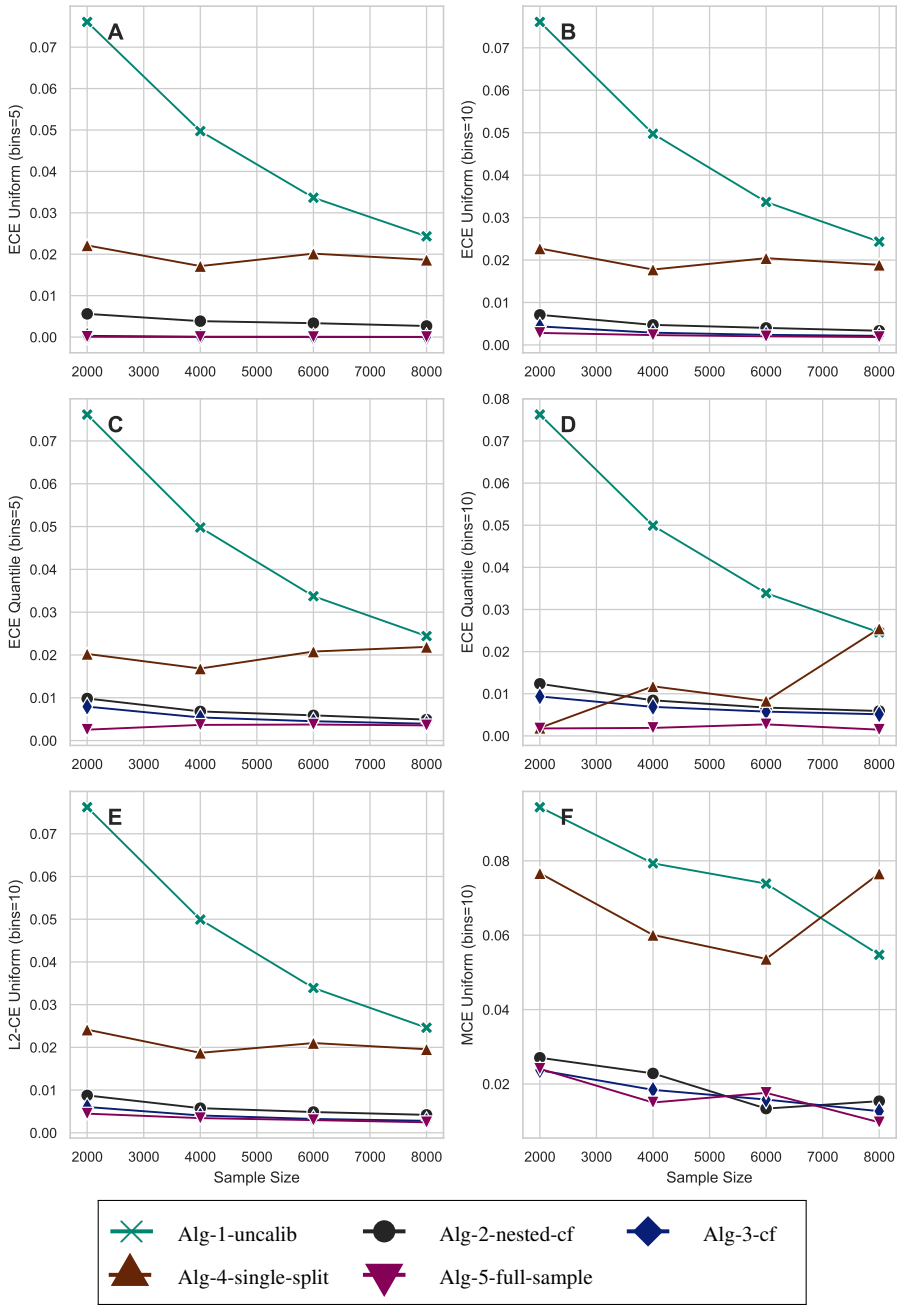

Figure 3.33: DGP 4 Unbalanced, Share_treated = 0.1, m = LGBM, p = 20, Calibration method for Algorithms 2-5: isotonic regression

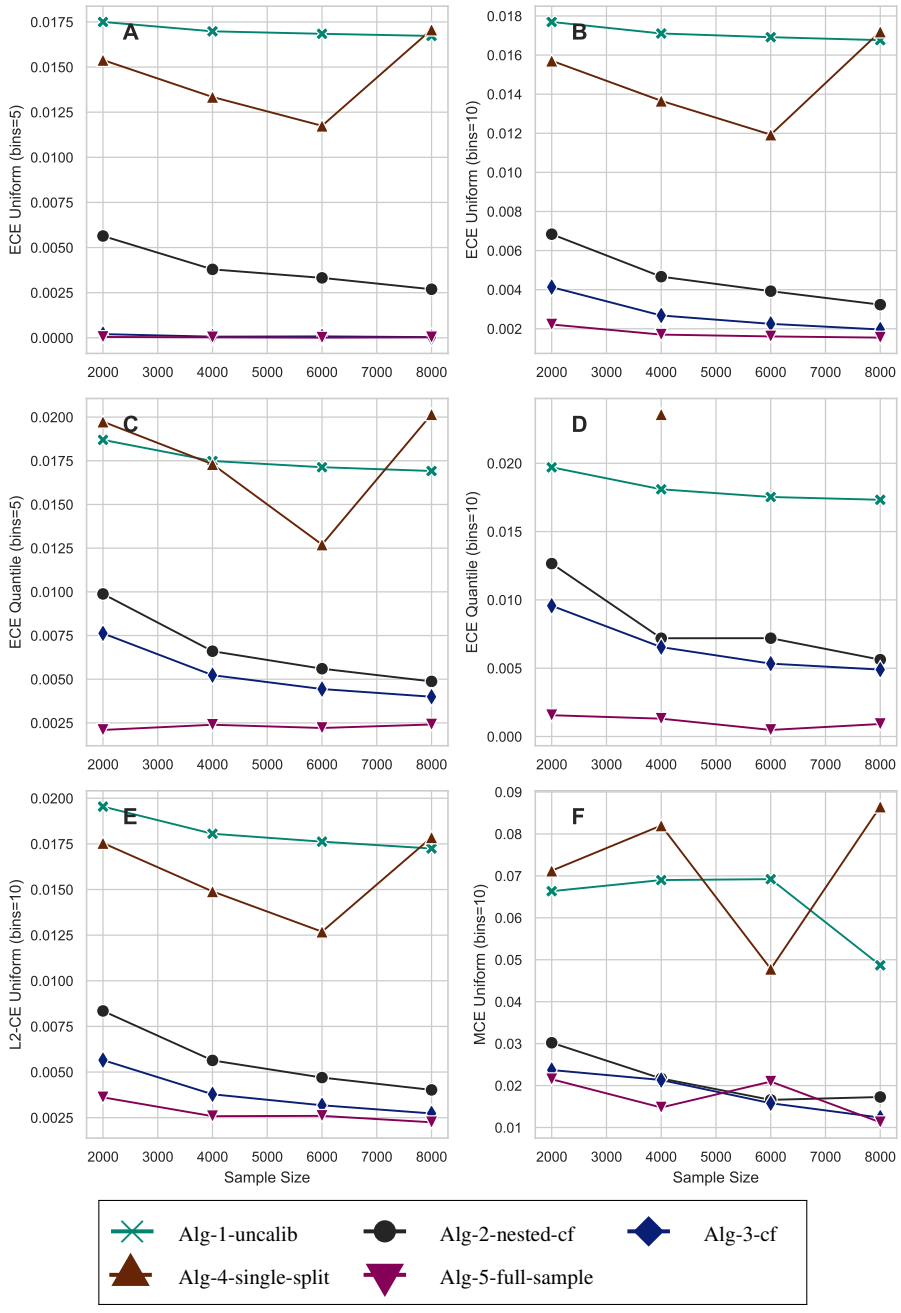

Figure 3.34: DGP 4 Unbalanced, Share_treated = 0.1, m = RF, p = 20, Calibration method for Algorithms 2-5: isotonic regression

ATE Errors on Sample Size

DGP 1 IRM

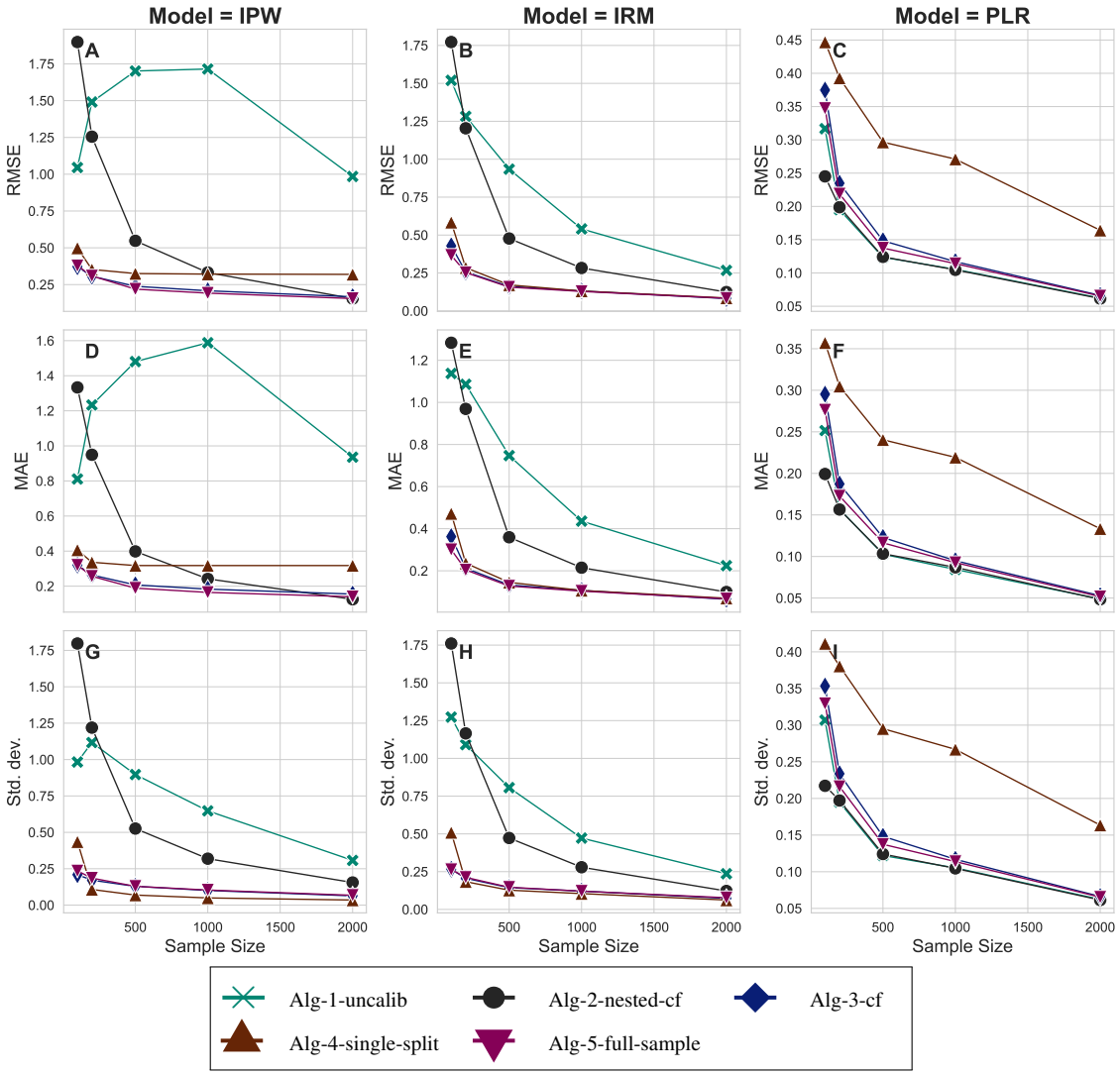

Figure 3.35: DGP 1 IRM, R2D = 0.5, m = LGBM, g = LGBM, p = 20, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

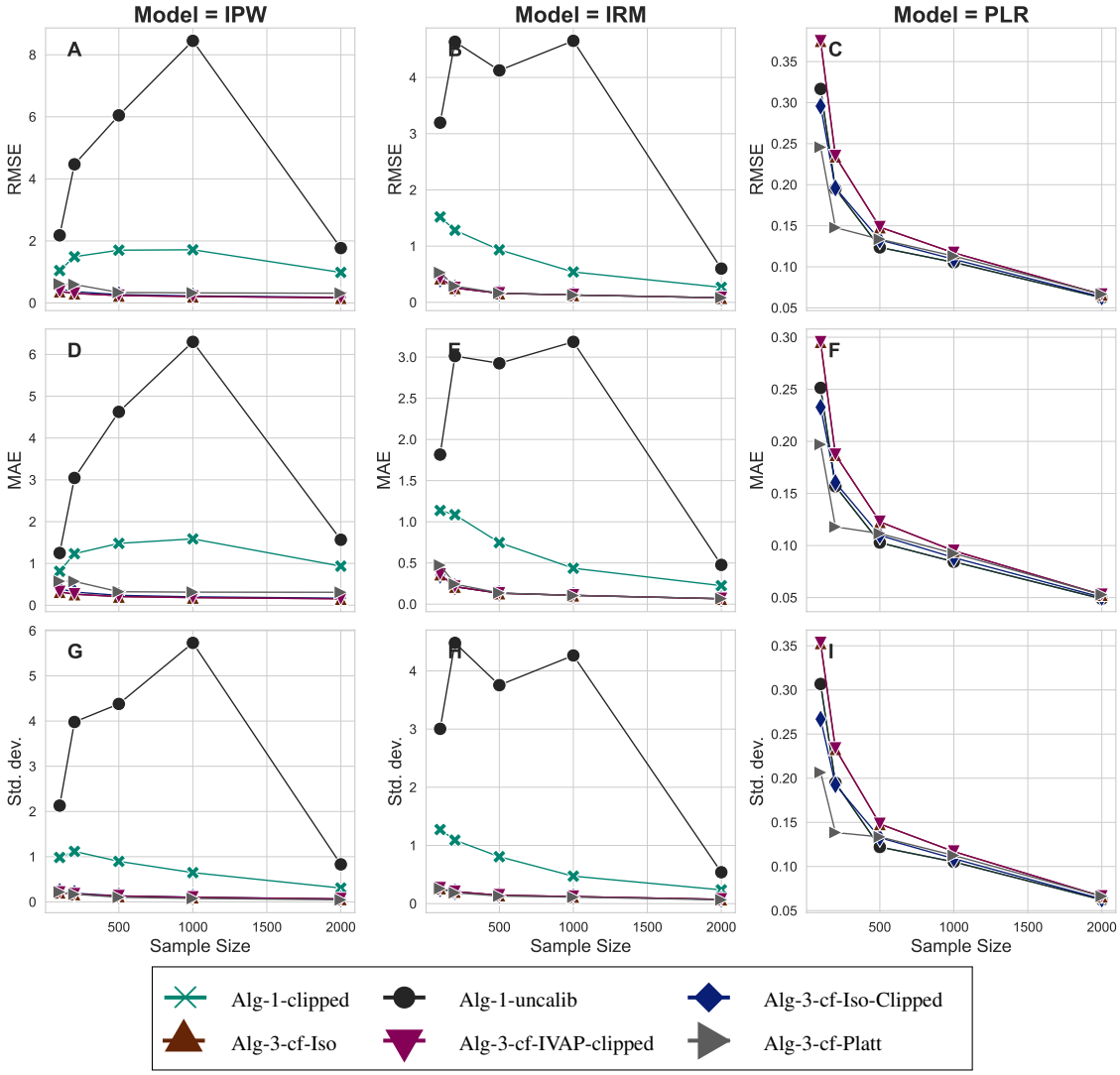

Figure 3.36: DGP 1 IRM, different calibration methods for Algorithm 3, R2D = 0.5, m = LGBM, g = LGBM, p = 20, , Clip = 0.01

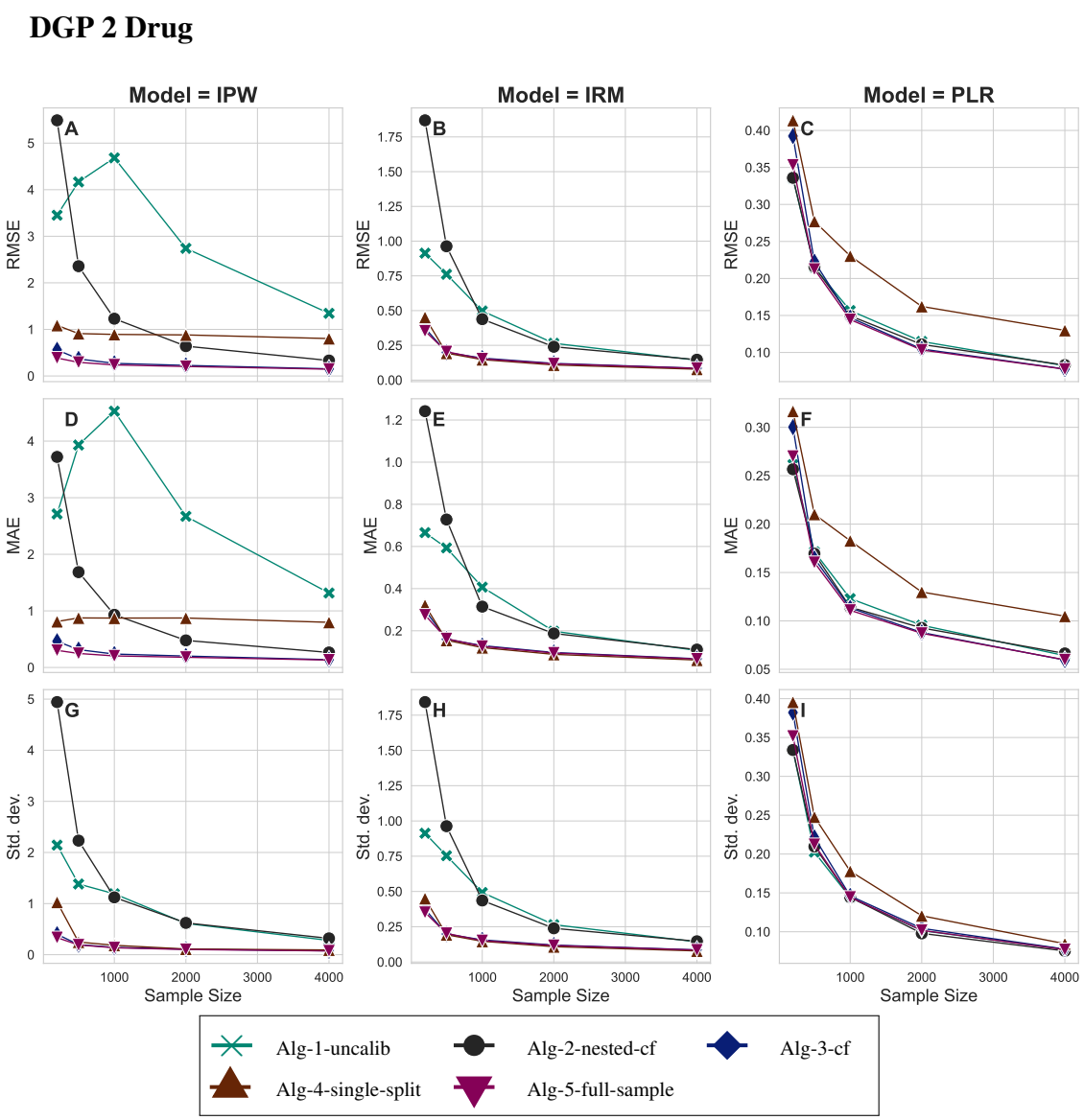

Figure 3.37: DGP 2 Drug, Overlap = 0.5, m = LGBM, g = LGBM, p = 3, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

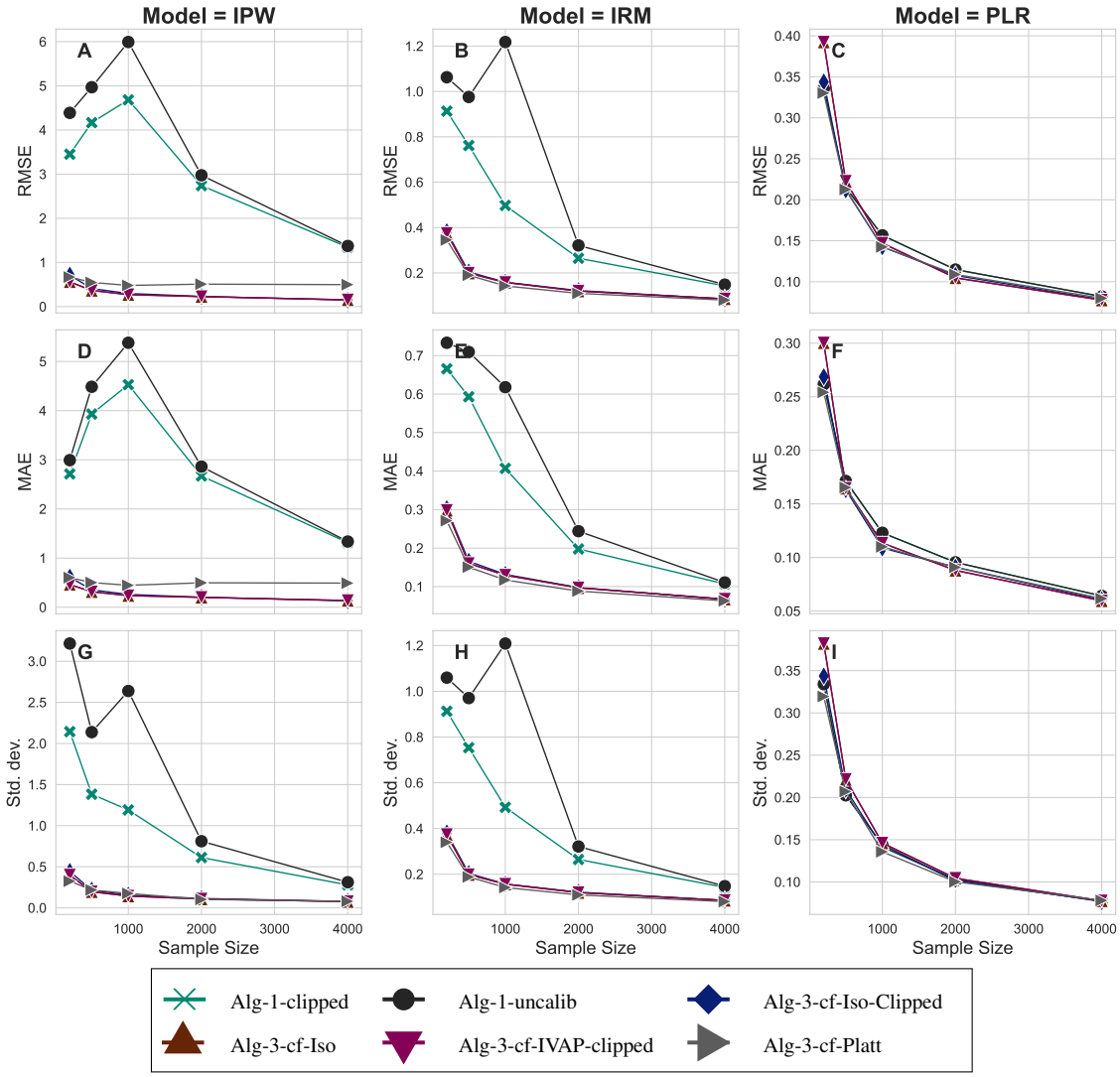

Figure 3.38: DGP 2 Drug, different calibration methods for Algorithm 3, Overlap = 0.5, m = LGBM, g = LGBM, p = 3, Clip = 0.01

DGP 3 Nonlinear

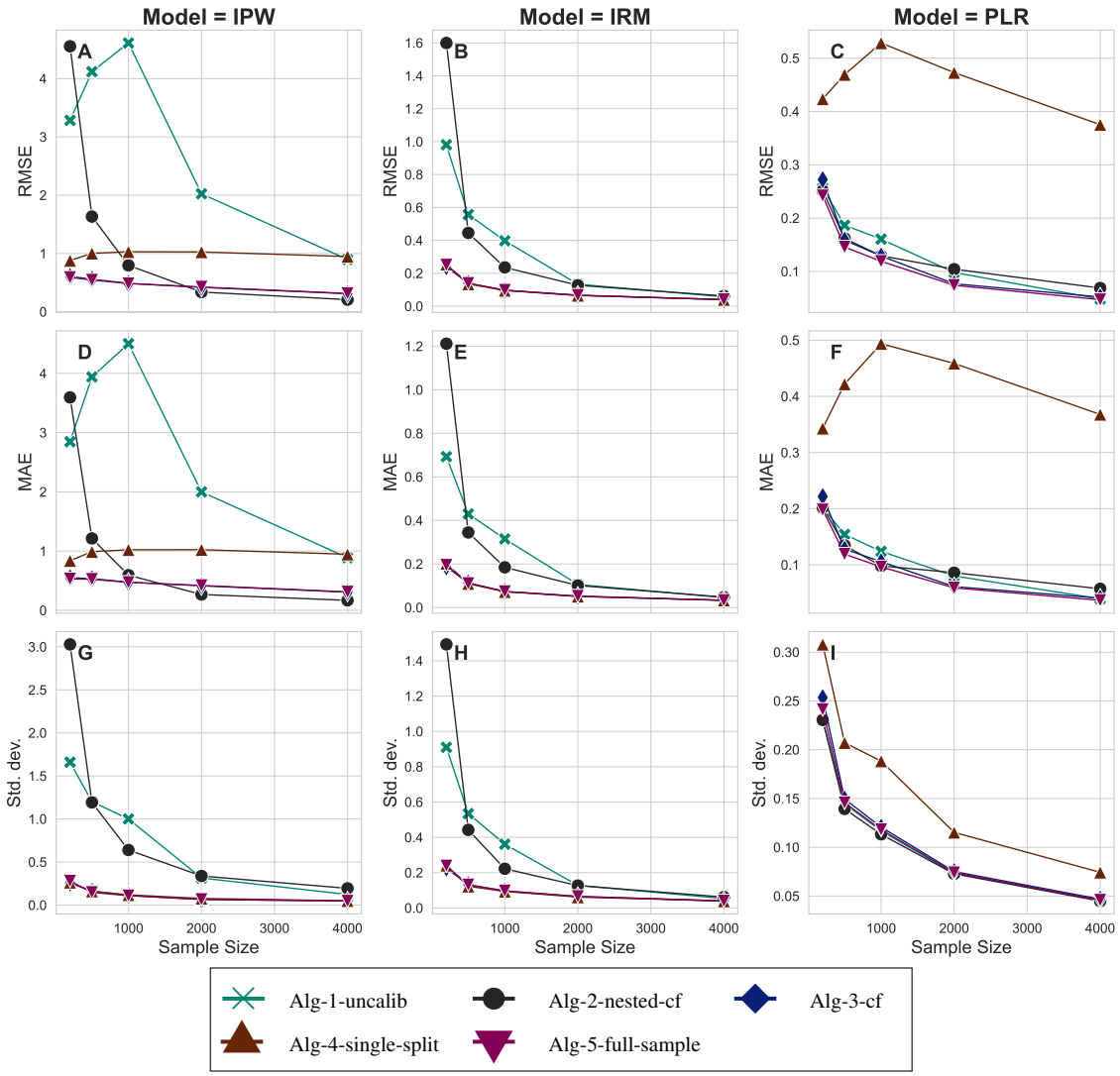

Figure 3.39: DGP 3 Nonlinear, m = LGBM, g = LGBM, p = 4, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

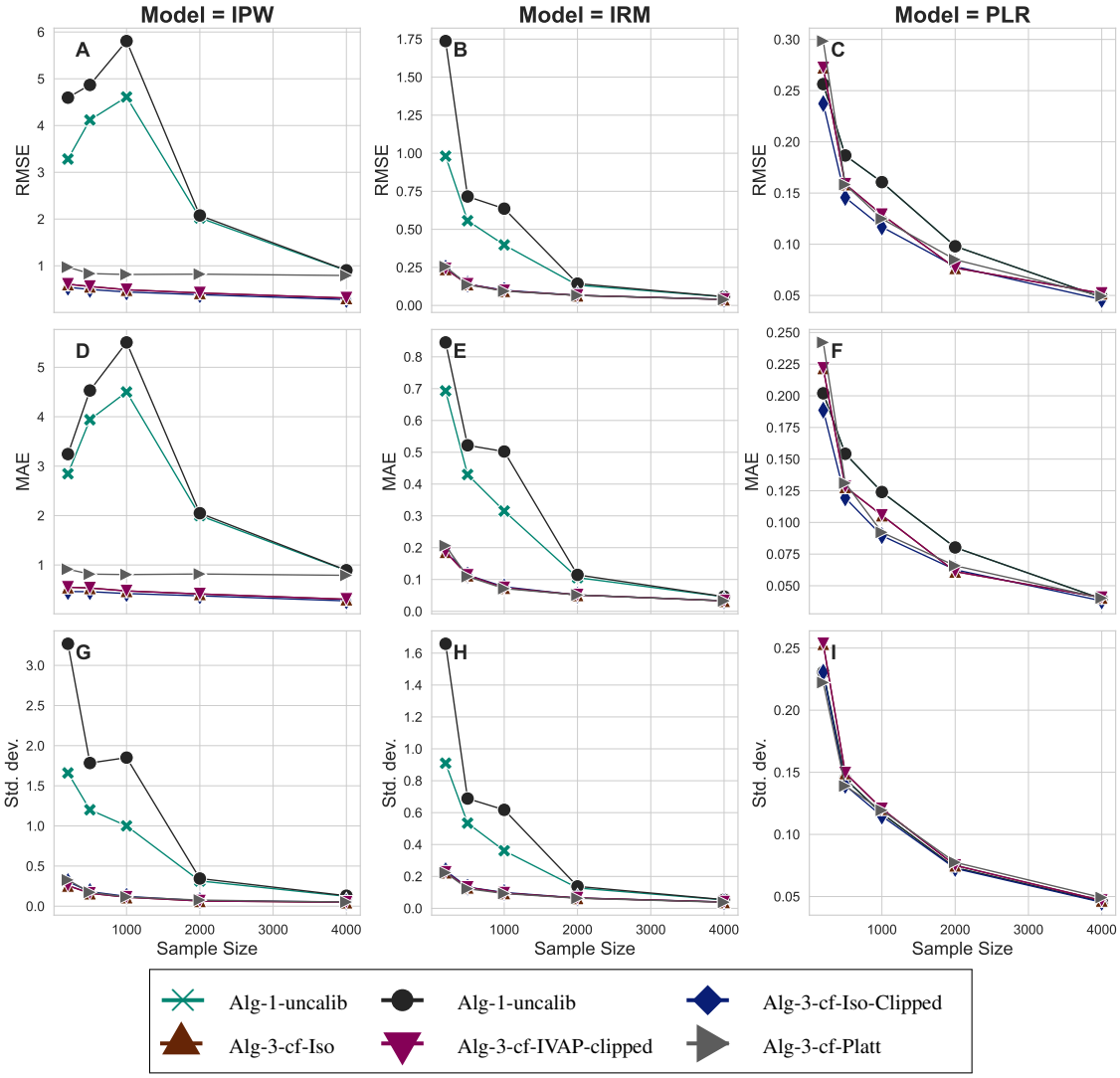

Figure 3.40: DGP 3 Nonlinear, different calibration methods for Algorithm 3, m = LGBM, g = LGBM, p = 4, Clip = 0.01

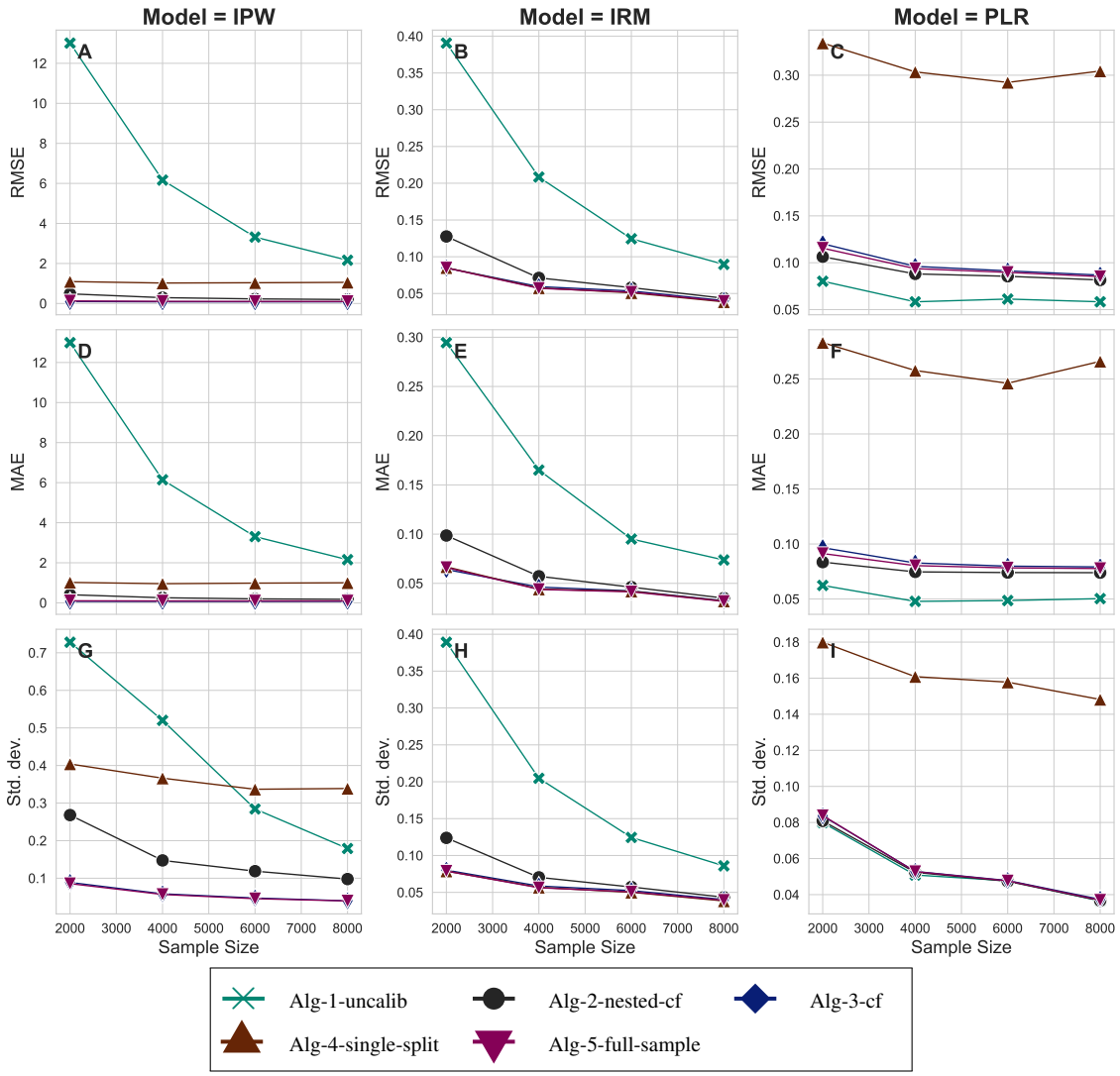

Figure 3.41: DGP 4 Unbalanced, Share_treated = 0.1, m = LGBM, g = LGBM, p = 20, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

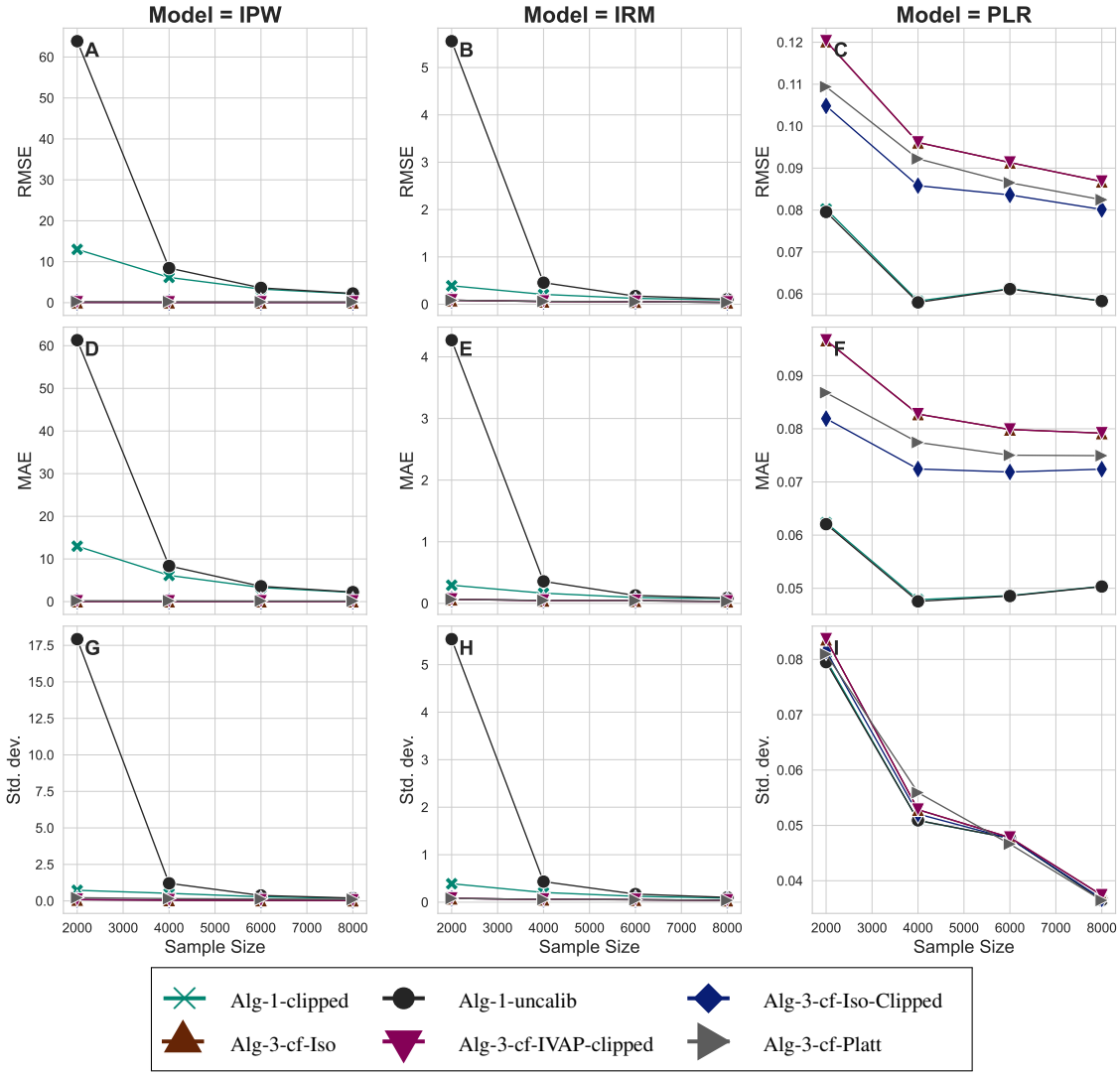

Figure 3.42: DGP 4 Unbalanced, different calibration methods for Algorithm 3, Share_treated = 0.1, m = LGBM, g = LGBM, p = 20, Clip = 0.01

Propensity Learner on ATE Distribution

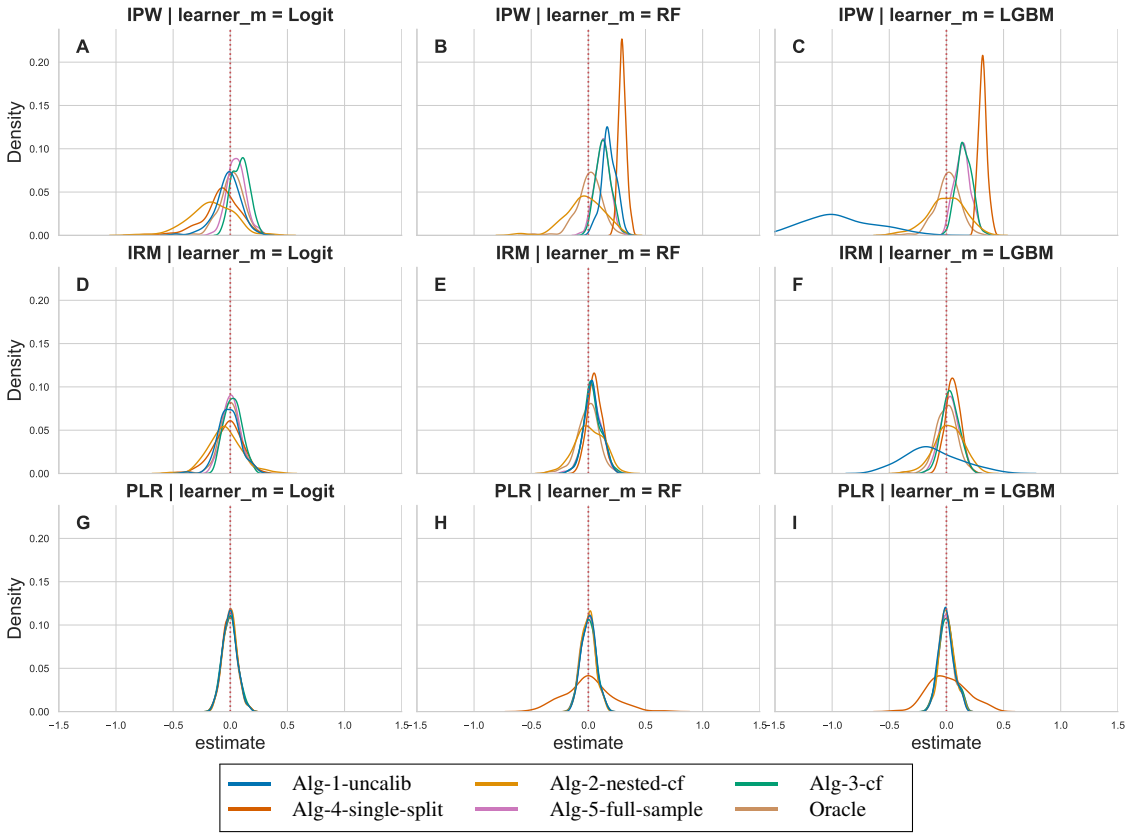

Figure 3.43: DGP 1 IRM, R2D = 0.5, m = LGBM, g = LGBM, n = 2000, p = 20, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

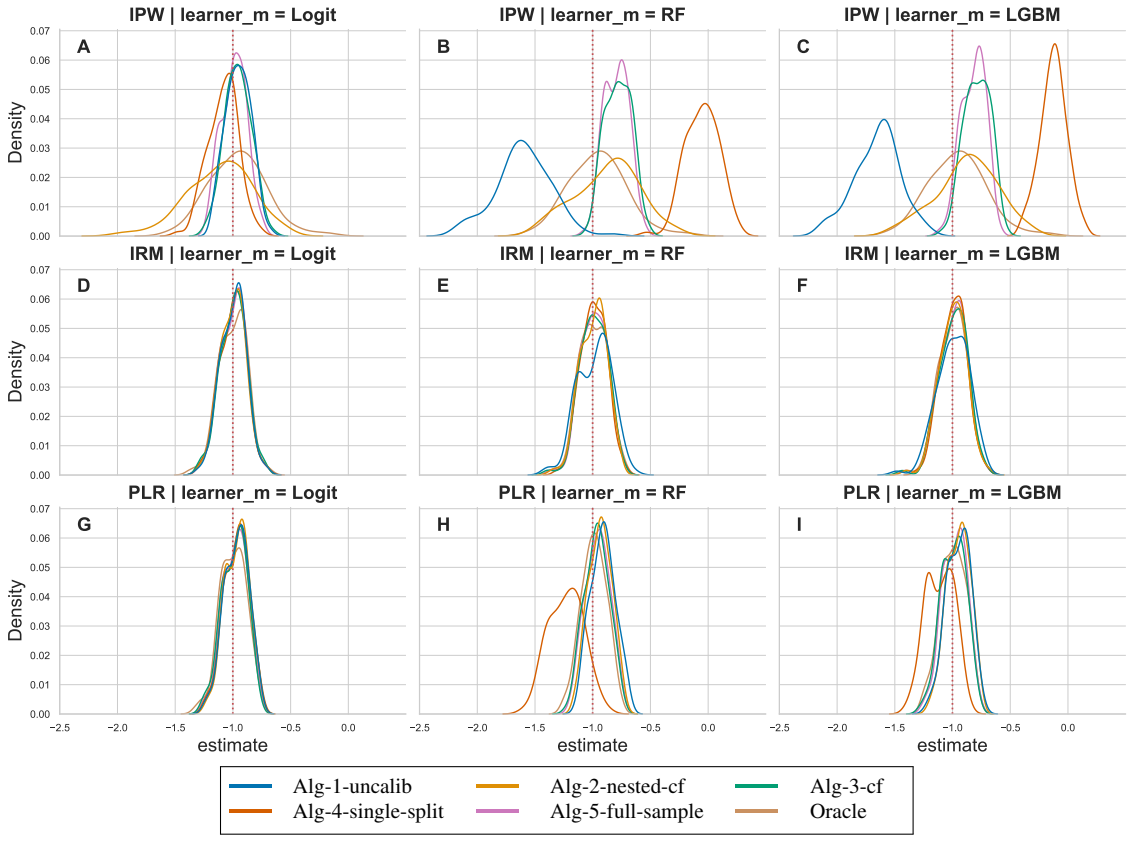

Figure 3.44: DGP 2 Drug, Overlap = 0.5, m = LGBM, g = LGBM, n = 2000, p = 3, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

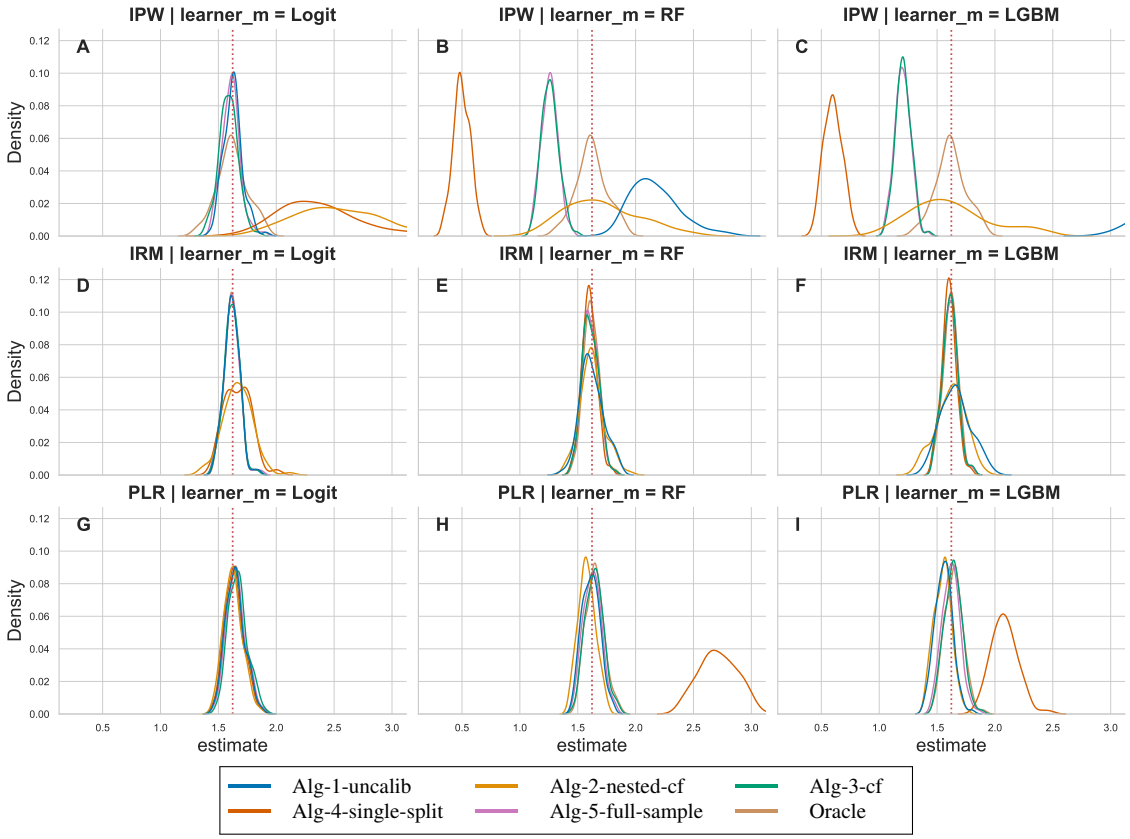

Figure 3.45: DGP 3 Nonlinear, m = LGBM, g = LGBM, n = 2000, p = 4, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

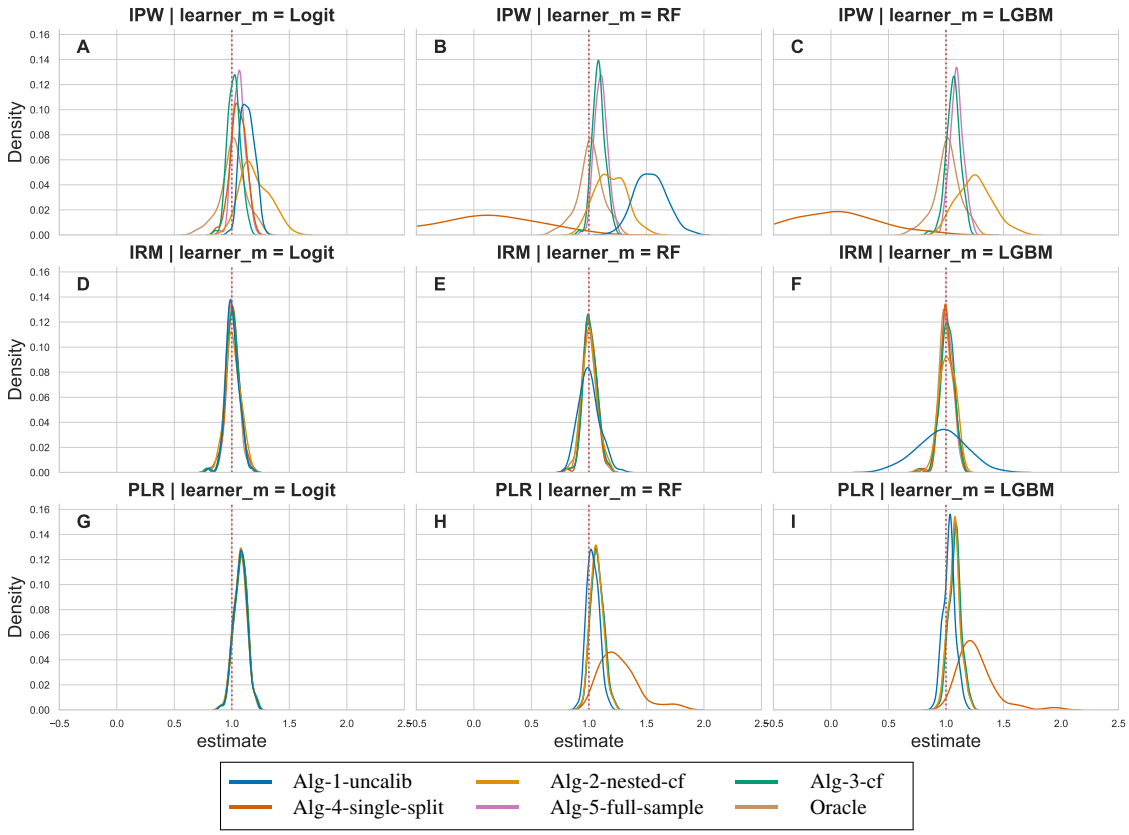

Figure 3.46: DGP 4 Unbalanced, Share_treated = 0.1, m = LGBM, g = LGBM, n = 4000, p = 20, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

Outcome Learner on ATE Distribution

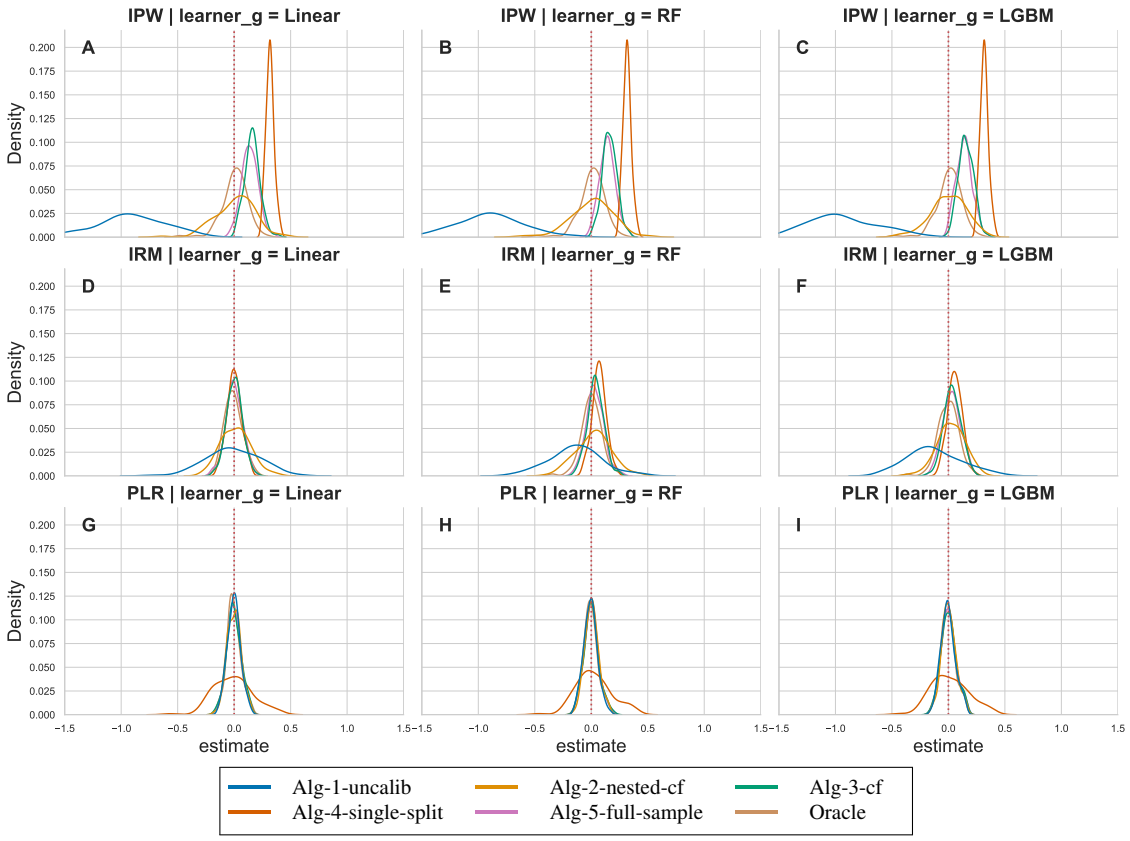

Figure 3.47: DGP 1 IRM, R2D = 0.5, m = LGBM, g = LGBM, n = 2000, p = 20, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

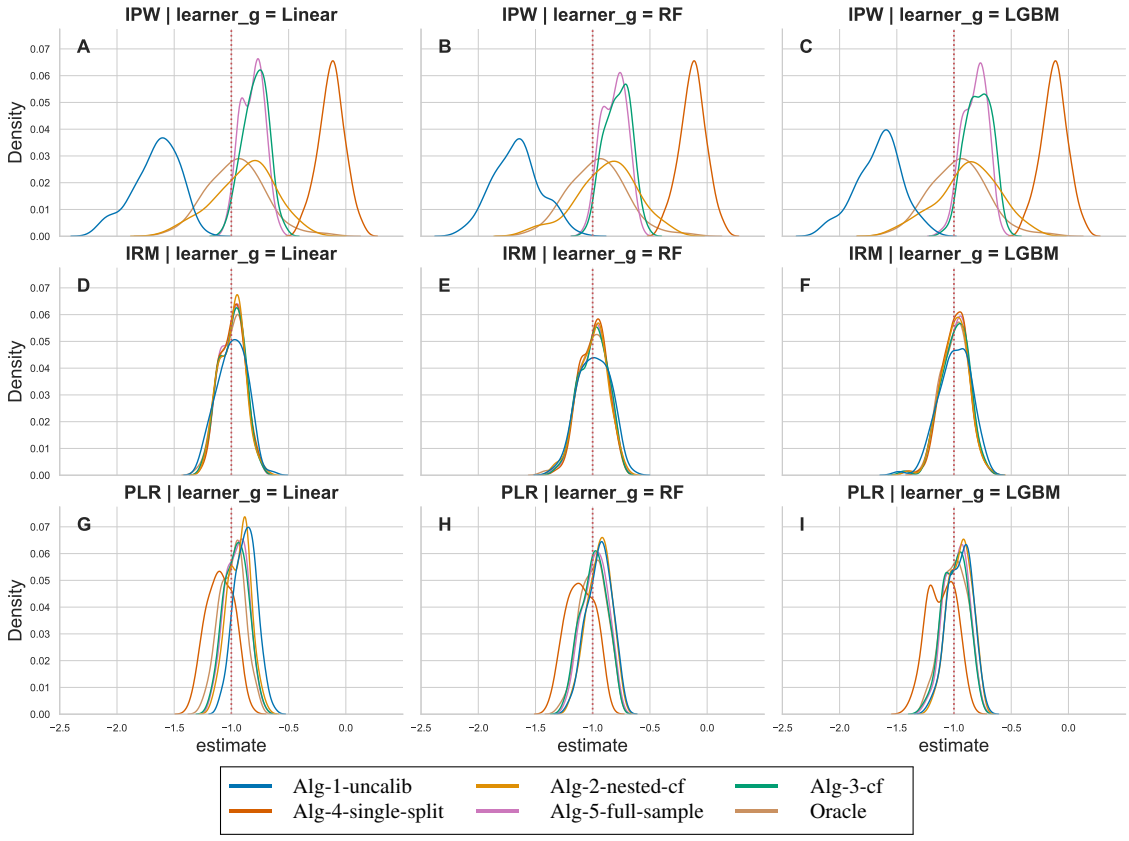

Figure 3.48: DGP 2 Drug, Overlap = 0.5, m = LGBM, g = LGBM, n = 2000, p = 3, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

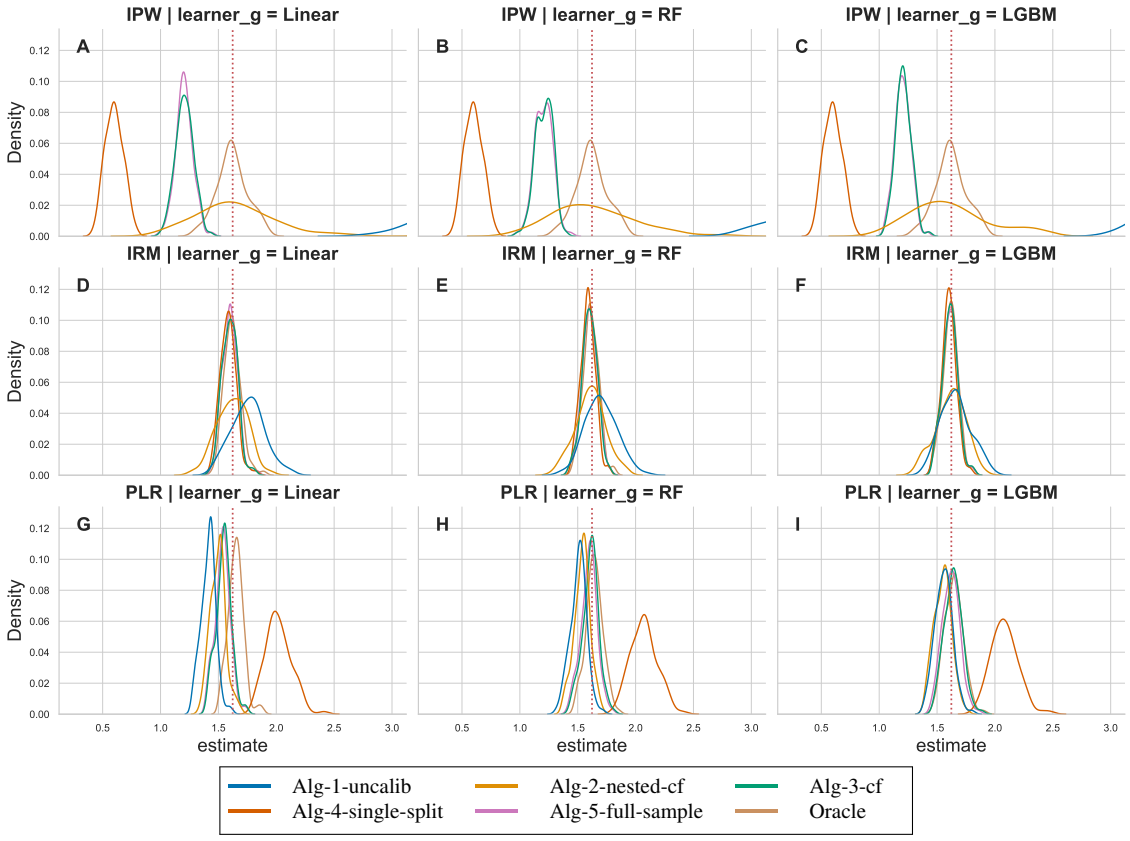

Figure 3.49: DGP 3 Nonlinear, m = LGBM, g = LGBM, n = 2000, p = 4, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

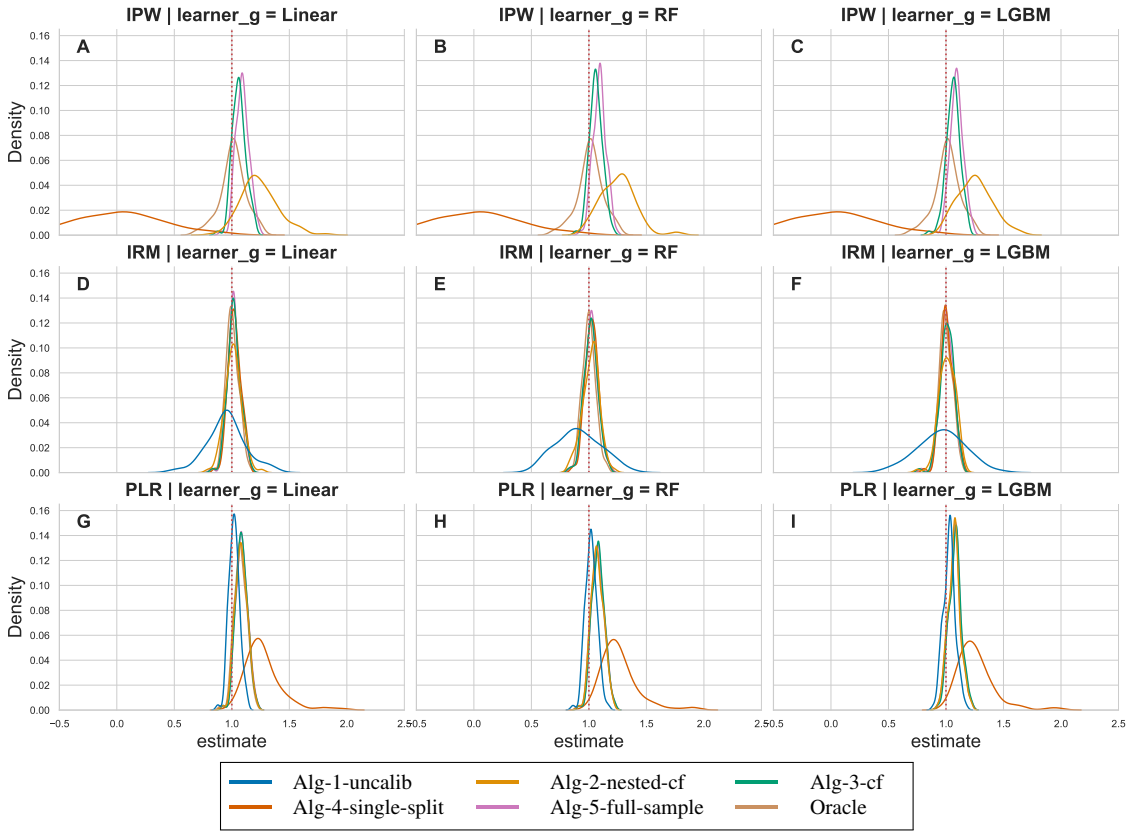

Figure 3.50: DGP 4 Unbalanced, Share_treated = 0.1, m = LGBM, g = LGBM, n = 4000, p = 20, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

Number of Covariates on ATE Distribution

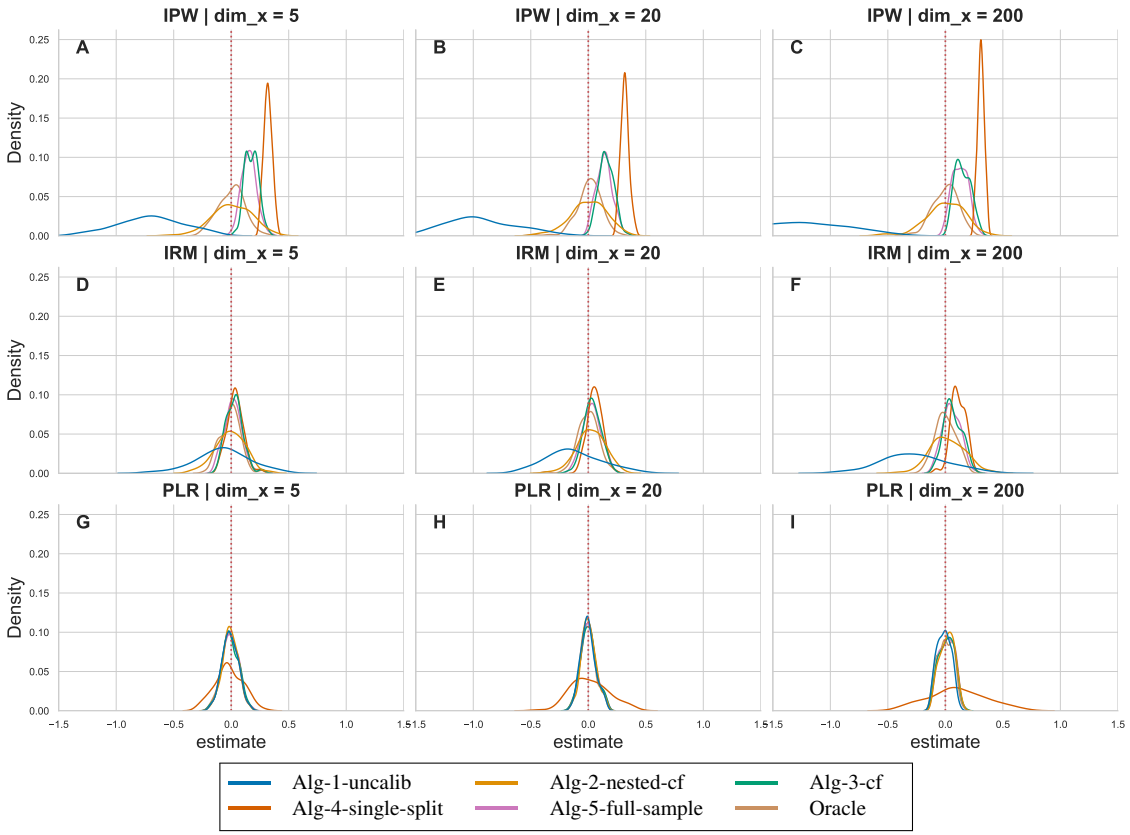

Figure 3.51: DGP 1 IRM, R2D = 0.5, m = LGBM, g = LGBM, n = 2000, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

Clipping Threshold on ATE Distribution

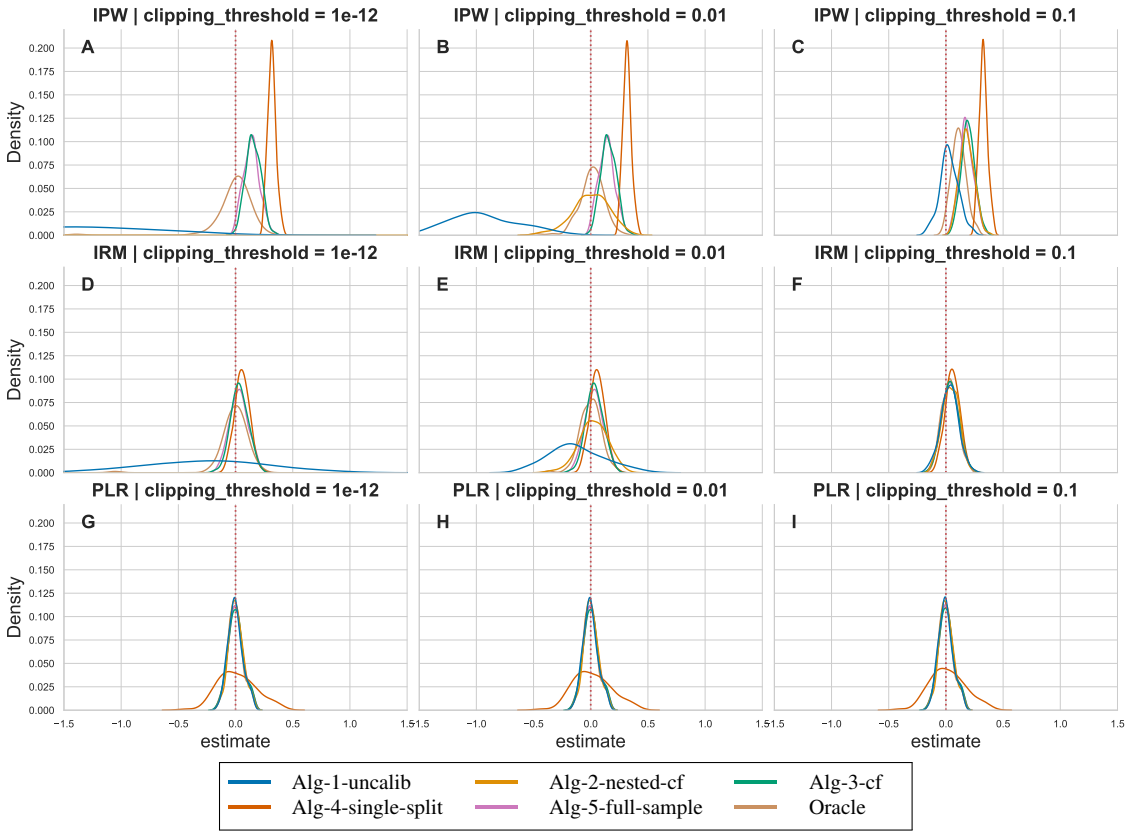

Figure 3.52: DGP 1 IRM, R2D = 0.5, m = LGBM, g = LGBM, n = 2000, p = 20, Calibration method for Algorithms 2-5: isotonic regression

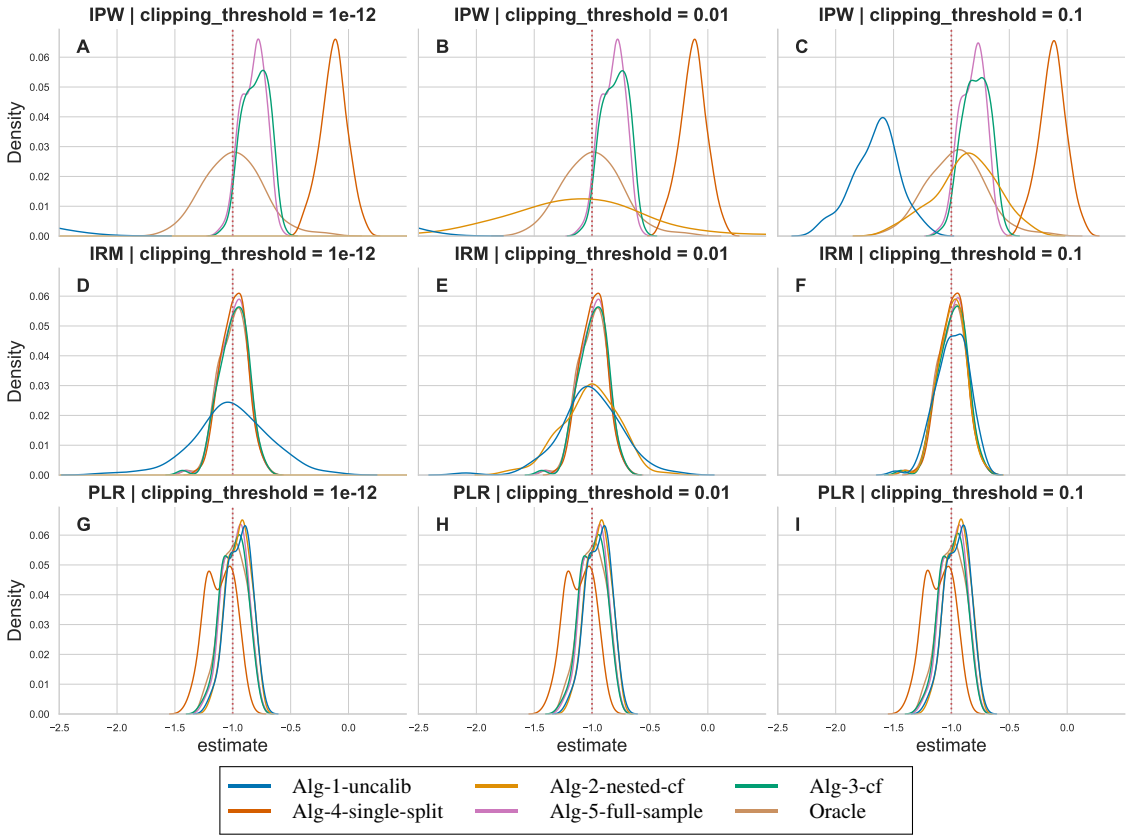

Figure 3.53: DGP 2 Drug, Overlap = 0.5, m = LGBM, g = LGBM, n = 2000, p = 3, Calibration method for Algorithms 2-5: isotonic regression

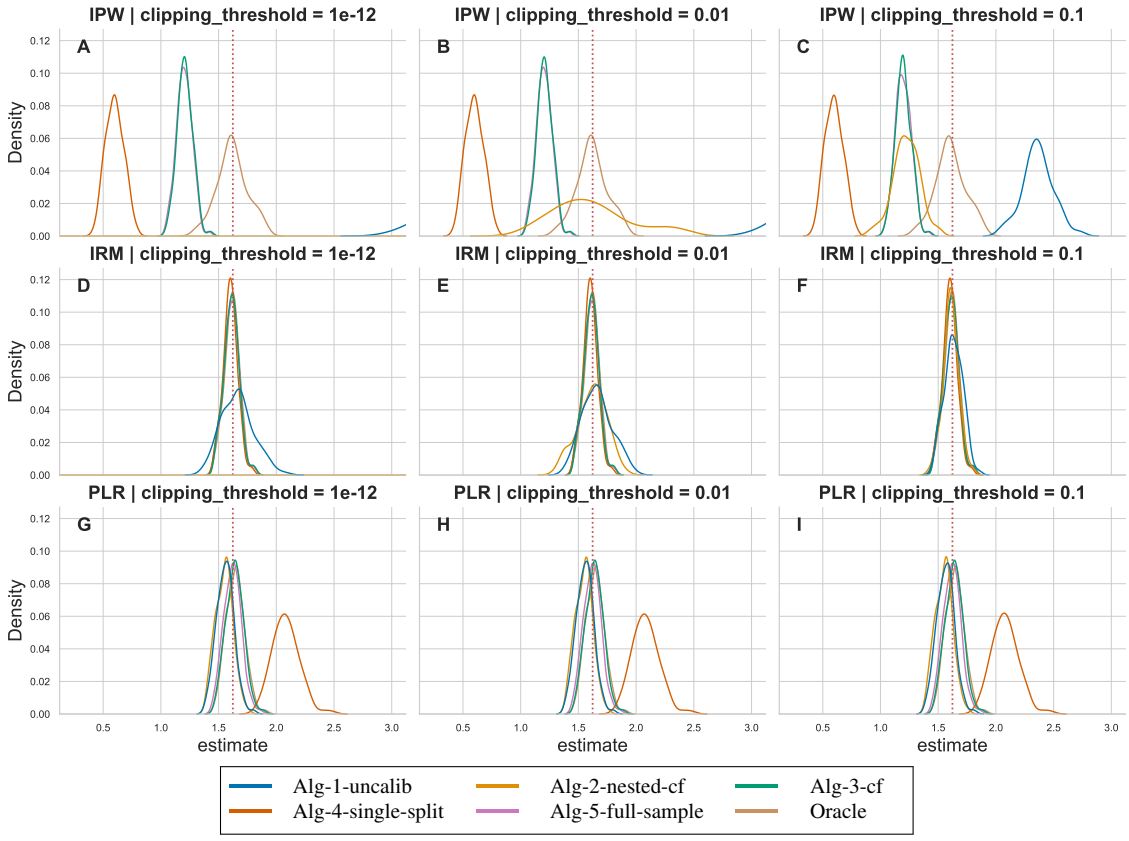

Figure 3.54: DGP 3 Nonlinear, m = LGBM, g = LGBM, n = 2000, p = 4, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

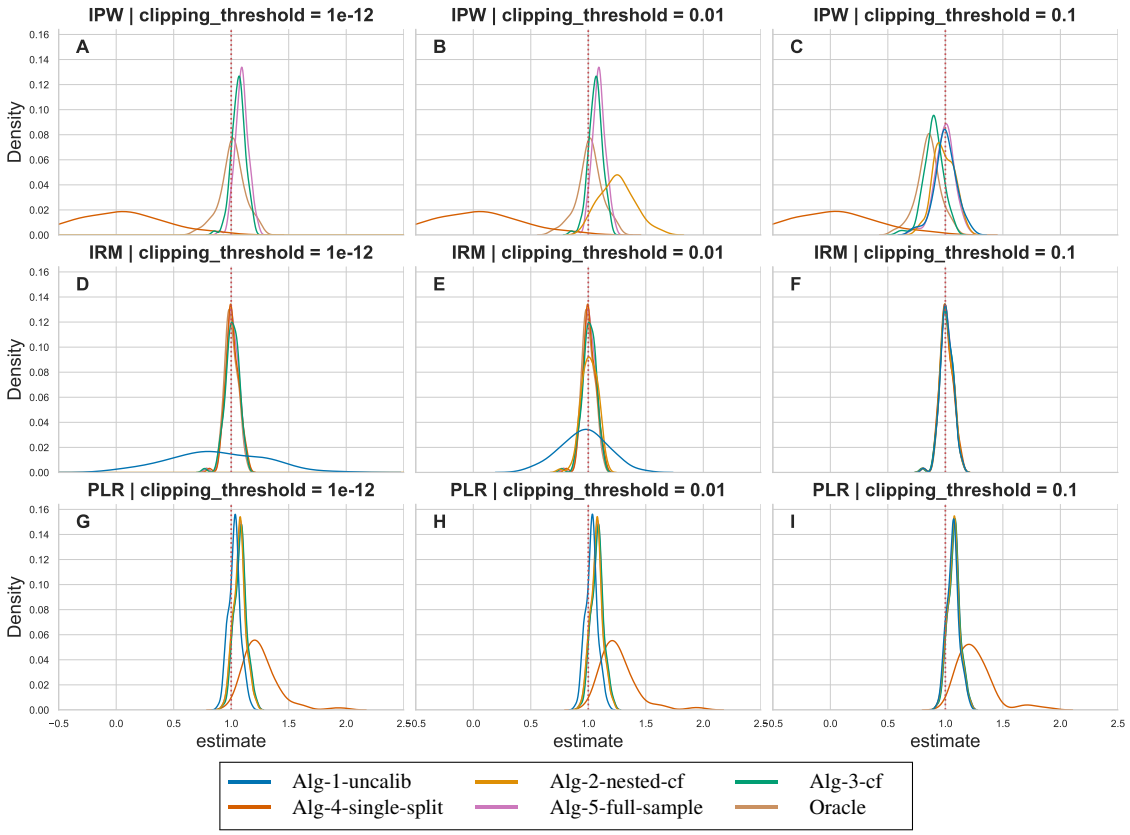

Figure 3.55: DGP 4 Unbalanced, Share_treated = 0.1, m = LGBM, g = LGBM, n = 4000, p = 20, Calibration method for Algorithms 2-5: isotonic regression

DGP Setting on ATE Distribution

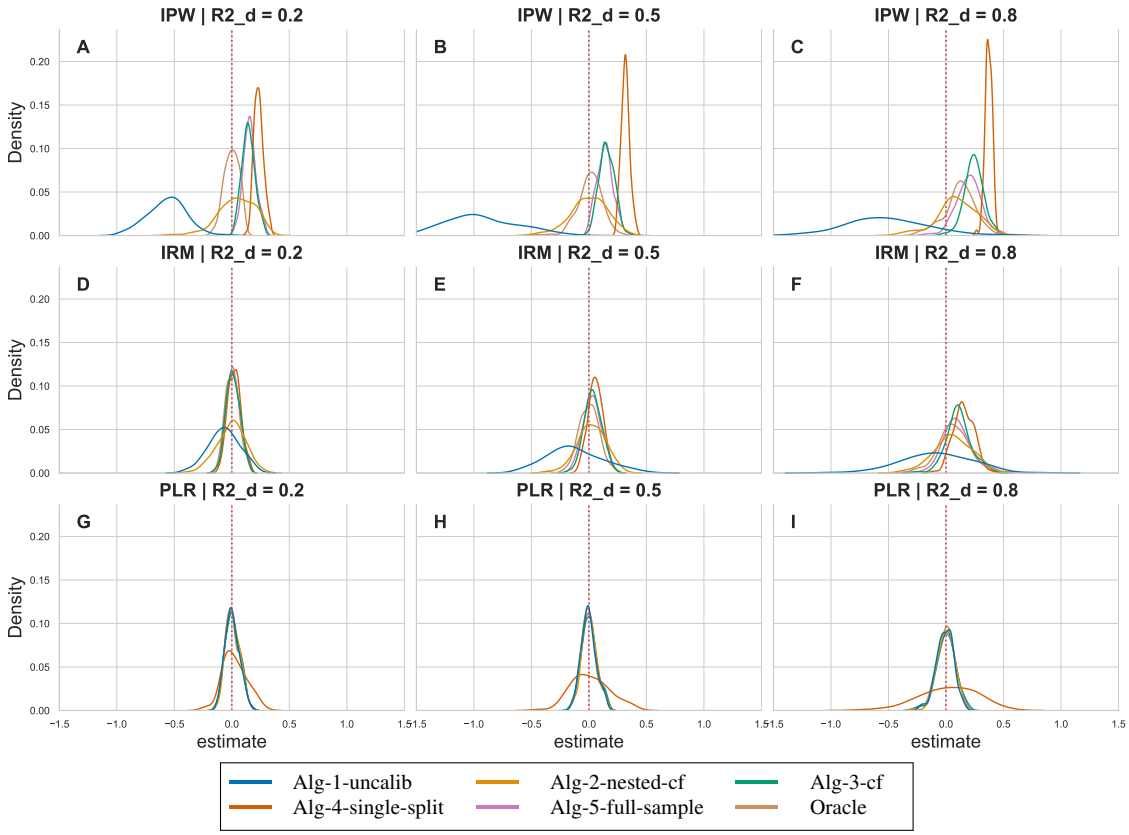

Figure 3.56: DGP 1 IRM, R2D on ATE, m = LGBM, g = LGBM, n = 2000, p = 20, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

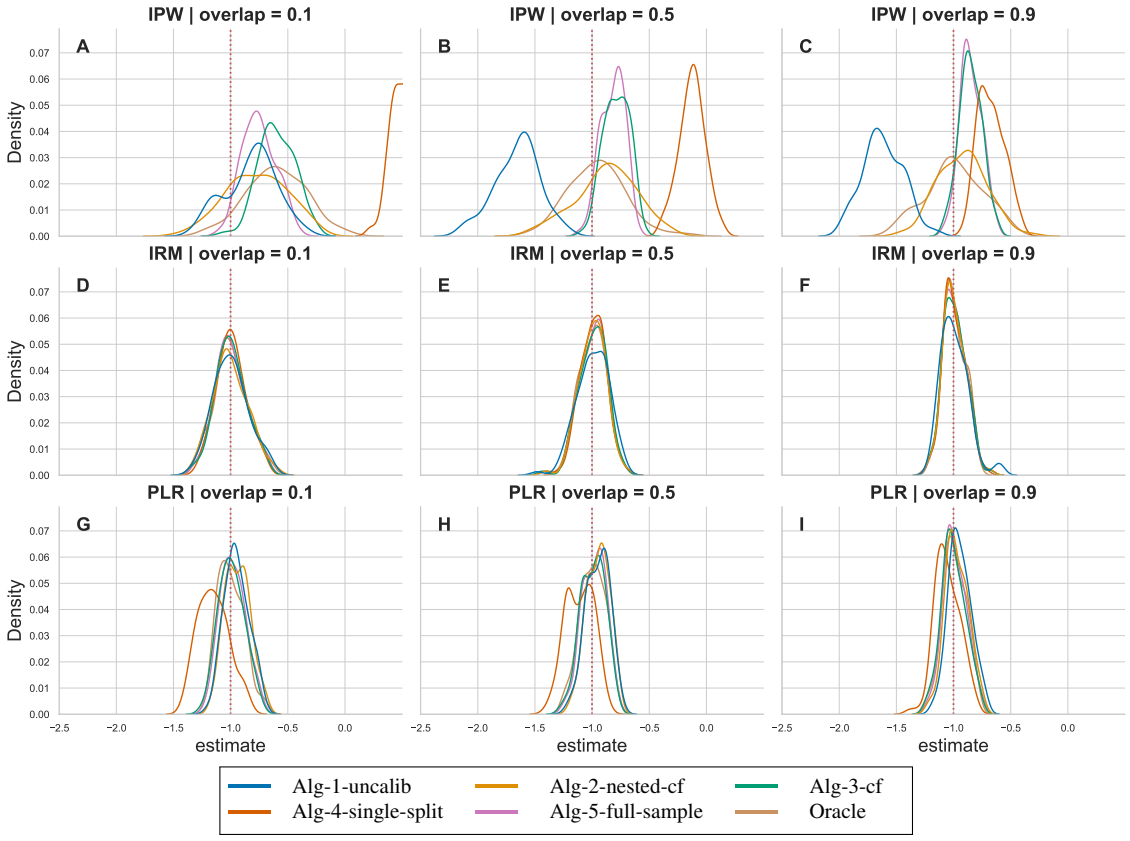

Figure 3.57: DGP 2 Drug, Overlap on ATE, m = LGBM, g = LGBM, n = 2000, p = 3, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

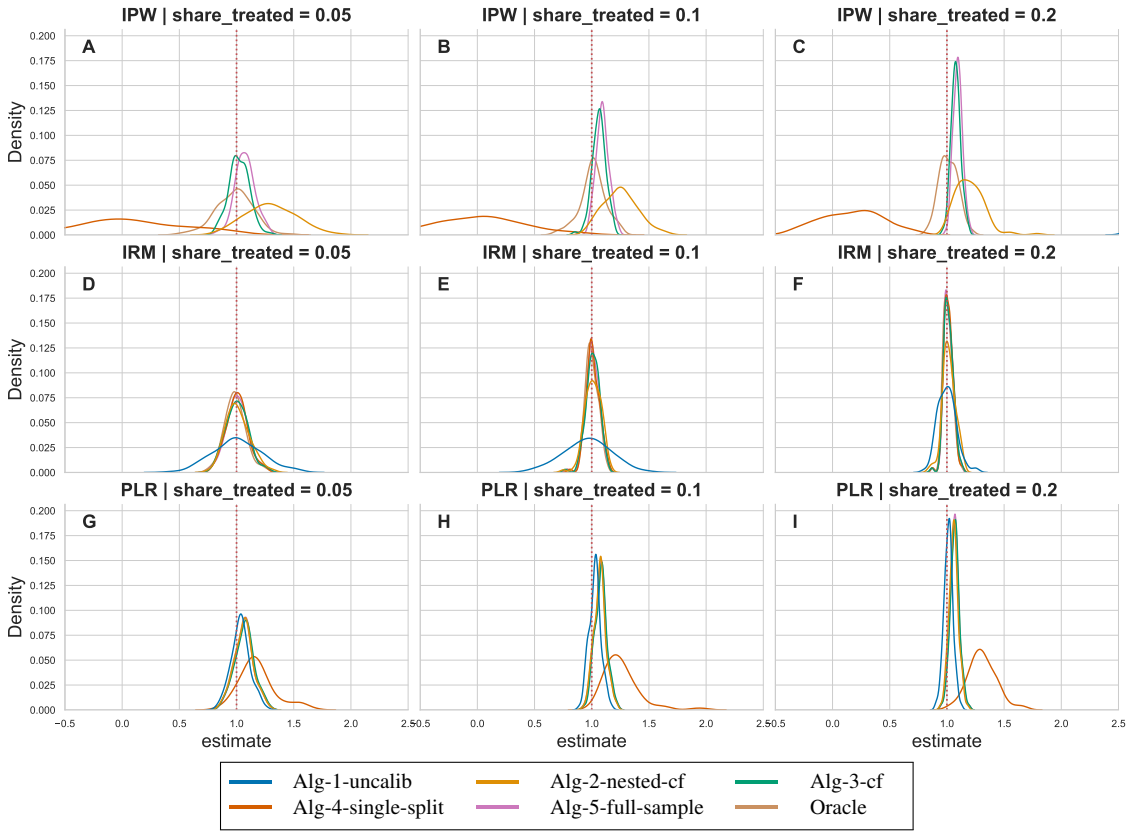

Figure 3.58: DGP 4 Unbalanced, Share_treated on ATE, m = LGBM, g = LGBM, n = 4000, p = 20, Calibration method for Algorithms 2-5: isotonic regression, Clip = 0.01

Chapter 4

Uncertainty Estimation in Insurance Claims Modeling: A Conformal Prediction Approach

Joint work with Michael Merz (University of Hamburg)

Abstract. Claims frequency prediction is crucial for insurance operations, yet standard models lack reliable uncertainty quantification. We address this by developing a conformal prediction framework for count data, providing valid prediction intervals under only the exchangeability assumption. Our contributions are: (1) a comprehensive evaluation of conformal methods for count data, tested on processes mimicking real insurance portfolios; (2) a novel two-stage framework (DGCP Hybrid) for zero-inflated data that balances coverage and efficiency; and (3) a diagnostic tool to cluster portfolios and assess interval performance across risk segments. Simulations show robust coverage where parametric methods fail. In an application to German motor insurance data, our method corrects the severe coverage imbalances found in standard approaches, providing practical, model-agnostic uncertainty quantification for risk differentiation.
Keywords: Prediction Intervals, Conformal Prediction, Count Data, Claims frequency modeling, Car insurance pricing, K-means algorithm

4.1 Introduction

Uncertainty quantification remains a critically understudied area in actuarial science, despite its fundamental importance for high-stakes applications such as insurance pricing, reserving, and risk selection. While predictive models for claims frequency have advanced considerably, from classical Generalized Linear Models (Nelder and Wedderburn 1972, McCullagh and Nelder 1989) to sophisticated machine learning approaches including gradient boosting (Denuit et al. 2020, Ferrario and Hämmerli 2019) and neural networks (Wüthrich and Merz 2019, Richman and Wüthrich 2023b), the accompanying uncertainty quantification has not kept pace. Standard practice produces point estimates without principled measures of prediction reliability, leaving actuaries unable to distinguish between confident predictions and those subject to substantial uncertainty.

Current approaches to uncertainty quantification in actuarial modeling fall into three broad categories. *Bootstrap methods* (Efron 1979, Davison and Hinkley 1997) provide a distribution-free approach by resampling residuals or observations to assess estimation variability. As demonstrated by Wüthrich and Merz (2023), bootstrap approaches can quantify model uncertainty in neural network predictions, revealing coefficients of variation of 15–25% on individual policy predictions even for large portfolios. However, bootstrap methods primarily address parameter

estimation uncertainty rather than providing formal coverage guarantees for prediction intervals. *Quantile regression* (Koenker and Bassett Jr 1978, Takeuchi et al. 2006, Meinshausen 2006) offers a distribution-free alternative that directly estimates conditional quantiles, bypassing distributional assumptions entirely. While theoretically attractive and increasingly adopted in machine learning contexts (Richman and Wüthrich 2021), quantile regression requires careful architectural constraints to ensure monotonicity across quantile levels and does not naturally provide the finite-sample coverage guarantees desirable for regulatory applications. *Bayesian approaches* (Gelman et al. 2013), including variational inference methods (Kingma and Welling 2014, Blundell et al. 2015), provide a principled framework for posterior uncertainty but require specification of prior distributions and can be computationally prohibitive for large-scale insurance portfolios; moreover, the resulting credible intervals lack frequentist coverage guarantees.

Conformal prediction (Vovk et al. 2022, Shafer and Vovk 2008) offers a fundamentally different approach: distribution-free prediction sets with finite-sample validity guarantees requiring only the mild assumption of exchangeability. Unlike parametric methods that may undercover when distributional assumptions fail, or bootstrap approaches that provide asymptotic rather than finite-sample guarantees, conformal prediction delivers mathematically rigorous coverage properties regardless of the underlying data distribution or model complexity. This model-agnostic framework can wrap any predictive algorithm – whether GLM, gradient boosting, or neural network – to provide uncertainty quantification without modifying the underlying model structure.

Despite these attractive properties, conformal prediction for count data remains largely unexplored. The severe class imbalance inherent in insurance claims data, where the majority of policyholders file no claims, creates a fundamental challenge: standard conformal methods achieve marginal coverage by over-covering the majority class while systematically under-covering claimants (Tsoumas and Papadopoulos 2024). This coverage imbalance renders uncertainty estimates unreliable precisely for the cases of greatest actuarial interest.

This work makes three principal contributions to address these gaps. First, we provide a comprehensive study of conformal prediction methods for count data, including extensive simulation studies across four data-generating processes(Poisson, zero-inflated Poisson, negative binomial, and hurdle models) and an empirical application to German motor insurance claims frequency modeling. Second, we develop a novel two-stage hybrid framework specifically designed for zero-inflated count data. Building on the Dynamically Grouped Conformal Prediction (DGCP) approach of Papaioannou et al. (2026) and the binary label-conditional framework of Tsoumas and Papadopoulos (2024), our DGCP Hybrid method combines Mondrian conformal prediction for the zero/non-zero decision with outcome-conditional calibration for positive counts, maintaining valid coverage across targeted claim categories while producing substantially narrower intervals than alternatives. Third, we introduce a covariate-space clustering diagnostic approach that identifies portfolio segments where prediction intervals exhibit systematic under-coverage, providing actionable guidance for model refinement and targeted risk management.

The remainder of this paper is organized as follows. Section 4.2 establishes notation and the conformal prediction framework, including standard inductive conformal prediction and the challenge of imbalanced outcomes for count data. Section 4.4 develops our DGCP Hybrid methodology for zero-inflated counts, including theoretical validity guarantees and connections to related approaches. Section 4.5 presents a comprehensive simulation study across four data-generating processes, demonstrating robust coverage under model misspecification. Section 4.6 applies the framework to German motor insurance data, introducing a cluster-based diagnostic for identifying portfolio segments with elevated prediction uncertainty. Section 4.7 concludes.

4.2 Notation

We establish the notation used throughout this work. Let $(\mathcal{X}, \mathcal{Y})$ denote the feature-outcome space, where $\mathcal{X} \subseteq \mathbb{R}^d$ represents the d-dimensional covariate space and $\mathcal{Y} = \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ denotes the space of non-negative integer-valued claim counts. We observe a dataset $\mathcal{D} = (X_i, Y_i)_{i=1}^N$ of N independent and identically distributed (i.i.d.) policyholder records, where $X_i \in \mathcal{X}$ denotes the feature vector (e.g., age, vehicle type, driving history) and $Y_i \in \mathcal{Y}$ denotes the observed claim count for policyholder i. For inductive conformal prediction, we partition the data into three disjoint sets $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{calib}} \cup \mathcal{D}_{\text{test}}$, with sizes $|\mathcal{D}_{\text{train}}| = n_{\text{train}}$, $|\mathcal{D}_{\text{calib}}| = n$, and $|\mathcal{D}_{\text{test}}| = n_{\text{test}}$. Let $\hat{f} : \mathcal{X} \to \mathbb{R}_+$ denote a fitted prediction model trained on $\mathcal{D}_{\text{train}}$, which outputs the predicted mean $\hat{\mu}_i = \hat{f}(X_i)$ for each observation. In the Poisson regression context standard in actuarial practice, we have $\mathbb{E}[Y_i \mid X_i] = \mu_i$ with $Y_i \mid X_i \sim \text{Poisson}(\mu_i)$. A non-conformity score (NCS) function $s : \mathbb{R}_+ \times \mathcal{Y} \to \mathbb{R}$ measures the discrepancy between a prediction and an observed outcome; for calibration point (X_i, Y_i) with prediction $\hat{\mu}_i$, we write $s_i = s(\hat{\mu}_i, Y_i)$. For a test point X_{n+1} with

Table 4.1: Summary of notation.

Symbol	Description
N	Total number of observations
n	Number of calibration observations
n_{test}	Number of test observations
X_i	Feature vector for observation i
Y_i	Observed claim count for observation i
$\hat{f}$	Fitted prediction model
$\hat{\mu}_i$	Predicted mean for observation i, i.e., $\hat{\mu}_i = \hat{f}(X_i)$
$s(\cdot, \cdot)$	Non-conformity score function
s_i	Non-conformity score for calibration point i
α	Significance level (miscoverage rate)
$1 - \alpha$	Confidence level (target coverage)
$\Gamma^\alpha(\cdot)$	Prediction set at level α
$[\ell, u]$	Prediction interval with lower bound ℓ and upper bound u
$Q_\tau(\cdot)$	τ-quantile function
m	Minimum group size for DGCP calibration
$y_{\max}$	Maximum candidate count value considered
$\mathcal{C}_y$	Set of calibration indices with outcome y
G_y	DGCP group containing outcome y

prediction $\hat{\mu}_{n+1}$, a prediction set at significance level $\alpha \in (0, 1)$ is denoted $\Gamma^\alpha(X_{n+1}) \subseteq \mathcal{Y}$. For count data, we report prediction intervals $[\ell, u]$ where $\ell = \min(\Gamma^\alpha)$ and $u = \max(\Gamma^\alpha)$. The marginal coverage of a prediction set procedure is defined as $\text{Coverage} = \mathbb{P}(Y_{n+1} \in \Gamma^\alpha(X_{n+1}))$, and a procedure achieves validity at level $1 - \alpha$ if $\text{Coverage} \geq 1 - \alpha$. For a grouping function $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{G}$, conditional coverage requires $\mathbb{P}(Y_{n+1} \in \Gamma^\alpha(X_{n+1}) \mid g(X_{n+1}, Y_{n+1}) = g) \geq 1 - \alpha$ for all $g \in \mathcal{G}$. Table 4.1 summarizes the principal notation used throughout this work.

4.3 Conformal Prediction Framework

Conformal prediction (CP) is a distribution-free, model-agnostic framework for uncertainty quantification that provides prediction sets with finite-sample validity guarantees (Vovk et al. 2022). Unlike traditional frequentist or Bayesian approaches that rely on parametric assumptions, CP requires only the mild assumption of *exchangeability* – that is, the joint distribution of data points is invariant to permutations.

Given a fitted prediction model $\hat{f}$, a user-specified significance level $\alpha \in (0, 1)$, and a calibration dataset, CP constructs prediction sets $\Gamma^\alpha(X_{n+1})$ for a new test observation X_{n+1} such that:

$$\mathbb{P}\left(Y_{n+1} \in \Gamma^\alpha(X_{n+1})\right) \geq 1 - \alpha, \tag{4.1}$$

where this coverage guarantee holds marginally over all test points under exchangeability (Shafer and Vovk 2008, Fontana et al. 2023).

In the actuarial context, CP is particularly attractive because it can be applied post-hoc to any claims frequency model, whether Generalized Linear Models (GLMs), gradient boosting machines, or neural networks, without requiring modifications to the underlying model structure. This enables actuaries to obtain uncertainty quantification for models that have been calibrated to meet in-sample portfolio constraints, such as matching the overall claims frequency target.

4.3.1 Inductive Conformal Prediction

The non-conformity score (NCS) measures the disagreement between a predicted value and an observed outcome. For a calibration point (X_i, Y_i) with prediction $\hat{\mu}_i = \hat{f}(X_i)$, the NCS quantifies how "unusual" or "non-conforming" the observation is relative to the model's expectation. For Poisson-distributed claims frequency data with conditional mean μ, we consider four non-conformity measures. The **Poisson deviance residual NCS** is based on the likelihood-ratio statistic from Poisson regression:

$$s_i^{\text{dev}} = \text{sign}(Y_i - \hat{\mu}_i) \cdot \sqrt{2\left[Y_i \log\left(\frac{Y_i}{\hat{\mu}_i}\right) - (Y_i - \hat{\mu}_i)\right]}, \tag{4.2}$$

with the convention that $0 \log(0) = 0$. This score has superior tail behavior compared to Pearson residuals and is the default in our implementation. The **Poisson Pearson residual NCS** provides a variance-stabilizing transformation exploiting the mean-variance relationship of the Poisson distribution:

$$s_i^{\text{pear}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + \epsilon}}, \tag{4.3}$$

where $\epsilon > 0$ is a small constant (typically $\epsilon = 0.1$) for numerical stability when $\hat{\mu}_i \approx 0$. The **zero-adjusted relative NCS** is designed specifically for zero-inflated data common in insurance:

$$s_i^{\text{ZA}} = \frac{|Y_i - \hat{\mu}_i|}{1 + \hat{\mu}_i}, \tag{4.4}$$

and the **absolute error NCS** provides the standard choice for regression problems:

$$s_i^{\text{abs}} = |Y_i - \hat{\mu}_i|. \tag{4.5}$$

The signed scores (deviance, Pearson) enable two-sided prediction intervals through separate lower and upper quantile estimation, while unsigned scores (zero-adjusted, absolute) yield intervals constructed from a single threshold.

Algorithm 1 Standard Inductive Conformal Prediction (ICP) for Count Data

Require: Fitted model $\hat{f}$, calibration set $\mathcal{D}_{\text{calib}} = \{(X_i, Y_i)\}_{i=1}^{n}$, test predictions $\{\hat{\mu}_j\}_{j=1}^{n_{\text{test}}}$, significance level α, NCS function $s(\cdot, \cdot)$, maximum count $y_{\max}$
Ensure: Prediction intervals $\{[\ell_j, u_j]\}_{j=1}^{n_{\text{test}}}$

Step 1: Compute Calibration Scores
for $i = 1$ **to** n **do**
    $s_i \leftarrow s(\hat{\mu}_i, Y_i)$ ▷ Non-conformity score
end for
Step 2: Compute Conformal Quantile
$k \leftarrow \lceil (n+1)(1-\alpha) \rceil / n$
$\hat{q} \leftarrow \text{Quantile}(\{s_1, \ldots, s_n\}, k)$
Step 3: Construct Prediction Sets
for $j = 1$ **to** n_{test} **do**
    $\Gamma_j \leftarrow \emptyset$
    for $\tilde{y} = 0$ **to** $y_{\max}$ **do**
        $s_{\text{test}} \leftarrow s(\hat{\mu}_j, \tilde{y})$
        if $s_{\text{test}} \leq \hat{q}$ **then**
            $\Gamma_j \leftarrow \Gamma_j \cup \{\tilde{y}\}$
        end if
    end for
    $\ell_j \leftarrow \min(\Gamma_j), \quad u_j \leftarrow \max(\Gamma_j)$
end for
return $\{[\ell_j, u_j]\}_{j=1}^{n_{\text{test}}}$

Let $\mathcal{D}_{\text{calib}} = \{(X_i, Y_i)\}_{i=1}^{n}$ denote the calibration set with corresponding NCS values $\{s_1, \ldots, s_n\}$. For a test point X_{n+1} with prediction $\hat{\mu}_{n+1}$, standard ICP constructs prediction sets as described in Algorithm 1.

For signed NCS types, we compute two-sided quantiles $\hat{q}_{\text{lower}} = Q_{\alpha/2}(\{s_i\}_{i=1}^{n})$ and $\hat{q}_{\text{upper}} = Q_{1-\alpha/2}(\{s_i\}_{i=1}^{n})$, and include all candidate values $\tilde{y}$ such that $\Gamma^{\alpha}(X_{n+1}) = \{\tilde{y} \in \mathcal{Y} : \hat{q}_{\text{lower}} \leq s(\hat{\mu}_{n+1}, \tilde{y}) \leq \hat{q}_{\text{upper}}\}$. For unsigned NCS types, we compute the adjusted quantile with finite-sample correction:

$$\hat{q} = Q_{\lceil (n+1)(1-\alpha) \rceil / n}(\{s_i\}_{i=1}^{n}), \tag{4.6}$$

and include candidates satisfying $s(\hat{\mu}_{n+1}, \tilde{y}) \leq \hat{q}$. For discrete count data, we iterate over candidate values $\tilde{y} \in \{0, 1, 2, \ldots, y_{\max}\}$ and report the contiguous interval $[\min(\Gamma^{\alpha}), \max(\Gamma^{\alpha})]$.

4.3.2 The Challenge of Imbalanced Outcomes

While standard ICP guarantees marginal coverage; that is, a proportion of at least $(1 - \alpha)$ of all test predictions containing the true value on average, this guarantee does not extend to subpopulations. In our German car insurance claims data, outcomes exhibit severe class imbalance: approximately 90% of policyholders file no claims in a typical observation period. As shown in Figure 4.1, this creates a fundamental problem: marginal CP can achieve its nominal coverage by over-covering the majority class (non-claimants) while systematically under-covering the minority class (claimants).

Label-Conditional Mondrian Conformal Prediction (LCMICP) addresses class imbalance by calibrating separately within each outcome class (Vovk et al. 2022, Papadopoulos

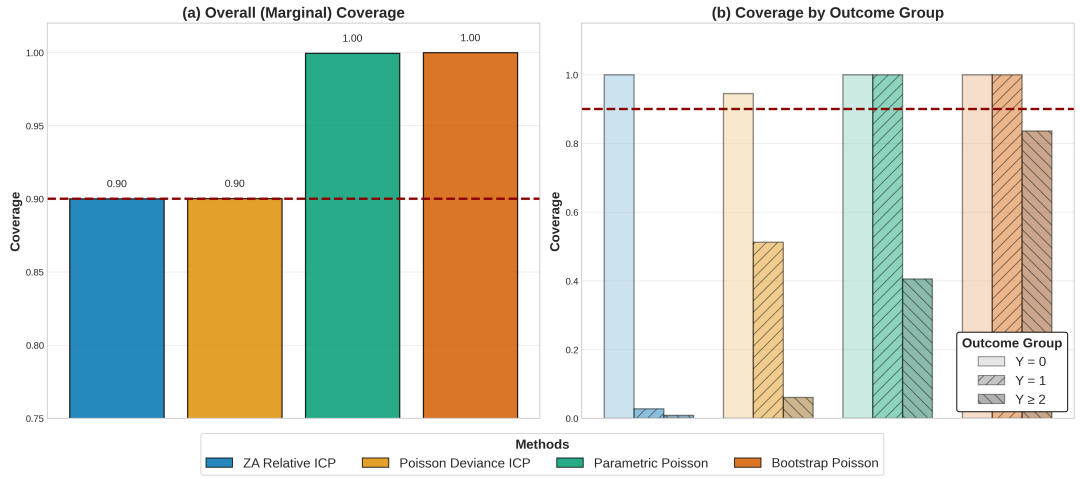

Figure 4.1: Marginal versus outcome-conditional coverage. **(a) Overall coverage** across all test observations. **(b) Coverage by outcome group**: $Y = 0$ (light bars), $Y = 1$ (medium bars, diagonal hatch), $Y \geq 2$ (dark bars, cross hatch). Dashed line indicates 90% target. Standard ICP achieves marginal validity through over-covering non-claimants ($Y = 0$) while under-covering claimants. German motor insurance data; details in Section 4.6.

2008). The calibration set is partitioned by label:

$$\mathcal{C}_y = \{i \in \{1, \ldots, n\} : Y_i = y\}, \quad y \in \mathcal{Y}, \tag{4.7}$$

with $n_y = |\mathcal{C}_y|$ denoting the number of calibration points with outcome y.

For each candidate label $\tilde{y}$, the p-value is computed using *only* calibration points with the same label:

$$p(\tilde{y}) = \frac{|\{i \in \mathcal{C}_{\tilde{y}} : s_i \geq s_{n+1}(\tilde{y})\}| + 1}{n_{\tilde{y}} + 1}. \tag{4.8}$$

The prediction set includes all candidates with p-value exceeding α:

$$\Gamma^{\alpha}(X_{n+1}) = \{\tilde{y} \in \mathcal{Y} : p(\tilde{y}) > \alpha\}. \tag{4.9}$$

This ensures that coverage guaranties hold for each class independently, preventing the majority class from dominating the calibration. Algorithm 2 presents the Mondrian CP procedure adapted for cluster-conditional calibration, where clusters may represent actuarially meaningful risk segments. This is similar to Ding et al. (2023), who use clustered Mondrian CP based on the estimated NCS. In contrast, our cluster construction is independent of the observed outcomes. We argue that, given the low variability in the outcomes and generally small fitted values, we do not observe sparse regions of the NCS that would necessitate clustering. However, we do encounter sparse regions in the covariate space.

The Sparse Calibration Problem. While Mondrian CP provides stronger class-conditional guaranties, it introduces a critical practical limitation: classes with few calibration samples cannot be reliably calibrated. In claims frequency modeling, this manifests as a hierarchy of data availability:

- Abundant data for $Y = 0$ (no claims): typically 90–98% of observations
- Moderate data for $Y = 1$ (single claim): typically 2–8% of observations
- Sparse data for $Y \geq 2$ (multiple claims): often $< 1\%$ of observations

Algorithm 2 Mondrian (Cluster-Conditional) Conformal Prediction

Require: Fitted model $\hat{f}$, calibration set $\mathcal{D}_{\text{calib}}$ with cluster labels $\{c_i\}_{i=1}^{n}$, test predictions $\{\hat{\mu}_j\}$ with clusters $\{c_j^{\text{test}}\}$, significance level α, minimum cluster size $n_{\min}$
Ensure: Prediction intervals $\{[\ell_j, u_j]\}_{j=1}^{n_{\text{test}}}$

Step 1: Partition Calibration by Cluster
for each cluster $c \in \mathcal{C}$ **do**
  $\mathcal{S}_c \leftarrow \{s_i : c_i = c\}$ ▷ Cluster-specific scores
  if $|\mathcal{S}_c| \geq n_{\min}$ **then**
    $\hat{q}_c \leftarrow \text{Quantile}(\mathcal{S}_c, 1 - \alpha)$
  else
    $\hat{q}_c \leftarrow \hat{q}_{\text{global}}$ ▷ Fallback to global quantile
  end if
end for
Step 2: Cluster-Conditional Prediction Sets
for $j = 1$ **to** n_{test} **do**
  $c \leftarrow c_j^{\text{test}}$
  Construct Γ_j using threshold $\hat{q}_c$
  $\ell_j \leftarrow \min(\Gamma_j), \quad u_j \leftarrow \max(\Gamma_j)$
end for
return $\{[\ell_j, u_j]\}_{j=1}^{n_{\text{test}}}$

As Papaioannou et al. (2026) demonstrate in the medical diagnosis context, class-conditional CP fails when calibration samples are insufficient, yielding undefined or unreliable quantile estimates for rare outcomes. The same phenomenon occurs in insurance: attempting to compute the 90th percentile of non-conformity scores for $y = 3$ when only 5 such observations exist in the calibration set produces unstable and invalid prediction intervals.

4.4 Adaptive Conformal Prediction

4.4.1 Dynamically Grouped Conformal Prediction

Dynamically Grouped Conformal Prediction (DGCP), introduced by Papaioannou et al. (2026) for medical diagnosis prediction, addresses the sparse calibration problem through a principled grouping strategy. The key insight is to maintain class-conditional calibration when sufficient data exists, while falling back to group-conditional calibration for underrepresented classes.

Let m denote the minimum number of calibration samples required for reliable class-level calibration. For each outcome class $k \in \mathcal{Y}$:

- If $|\mathcal{C}_k| \geq m$: Apply class-conditional calibration using only samples with $y = k$.
- If $|\mathcal{C}_k| < m$: Group with semantically related classes and apply group-conditional calibration.

For count data without hierarchical structure (unlike medical diagnosis codes which have ICD taxonomies), "semantically related" is operationalized as neighboring count values. Classes with insufficient data expand symmetrically to include $y - 1, y + 1, y - 2, y + 2, \ldots$ until the group contains at least m calibration samples.

A critical adaptation is required when applying DGCP to count data with residual-based non-conformity scores. In the original classification setting of Papaioannou et al. (2026), the NCS is $s_i = 1 - \hat{f}(X_i)_{Y_i}$, which measures disagreement between prediction and true label. When testing a candidate $\tilde{y}$, both the test score and calibration scores use the same functional form.

For count data with Poisson-based NCS (e.g., deviance residuals), a naive implementation would compute calibration scores as $s(\hat{\mu}_i, Y_i)$ using actual outcomes. However, when groups contain multiple count values, this creates heterogeneous score distributions:

$$\mathcal{S}_{G_y}^{\text{naive}} = \{s(\hat{\mu}_i, Y_i) : Y_i \in G_y\} = \{s(\hat{\mu}_i, 4), s(\hat{\mu}_j, 5), s(\hat{\mu}_k, 6), \ldots\} \tag{4.10}$$

When testing candidate $\tilde{y} = 5$, comparing $s(\hat{\mu}_{\text{test}}, 5)$ against this heterogeneous set violates the exchangeability principle underlying conformal prediction.

Solution: Hypothetical Scores. We compute calibration scores *as if* all observations in the group had the candidate outcome value:

$$\mathcal{S}_{G_y}^{\text{hyp}}(\tilde{y}) = \{s(\hat{\mu}_i, \tilde{y}) : Y_i \in G_y\} \tag{4.11}$$

This ensures that when testing $\tilde{y} = 5$:

$$\mathcal{S}_{G_5}^{\text{hyp}}(5) = \{s(\hat{\mu}_i, 5), s(\hat{\mu}_j, 5), s(\hat{\mu}_k, 5), \ldots\} \tag{4.12}$$

The test score $s(\hat{\mu}_{\text{test}}, 5)$ is now compared against scores of the same functional form, restoring the proper conformal comparison. This adaptation follows the conformal prediction principle of testing the null hypothesis "$\tilde{y}$ is the true outcome" by asking: "how would scores behave if $\tilde{y}$ were indeed true?"

DGCP Variants for Insurance Claims. We adapt the DGCP framework specifically for insurance claims frequency, where outcomes are non-negative integers with a zero-inflated distribution. Our implementation provides two variants. Algorithms 3 and 4 present the procedures for DGCP Full and DGCP Hybrid, respectively. DGCP Full applies hypothetical-score calibration uniformly for all counts $y \in \{0, 1, \ldots, y_{\max}\}$. DGCP Hybrid invokes DGCP Full as a subroutine for positive counts while using a separate binary calibration for $Y = 0$.

DGCP Full. Applies the grouping principle with hypothetical scores uniformly across all count values $y \in \{0, 1, 2, \ldots\}$. For each candidate $\tilde{y}$, the conformal p-value is:

$$p(\tilde{y}) = \frac{|\{i : s(\hat{\mu}_i^{\text{cal}}, \tilde{y}) \geq s(\hat{\mu}_{\text{test}}, \tilde{y}), Y_i^{\text{cal}} \in G_{\tilde{y}}\}| + 1}{|G_{\tilde{y}}| + 1} \tag{4.13}$$

where $G_{\tilde{y}}$ is the dynamically constructed group for count $\tilde{y}$. This variant does not distinguish between the claim/no-claim boundary.

DGCP Hybrid. A two-stage approach optimized for zero-inflated count data, combining binary classification CP for the zero/non-zero decision with hypothetical-score DGCP for positive counts. Stage 1 applies Mondrian CP for $Y = 0$. We use a monotone transformation of $\hat{\mu}$ as a score for the zero class; CP validity does not rely on $exp(-\hat{\mu})$ being a calibrated probability, only on exchangeability. While Stage 2 applies DGCP with hypothetical scores for $Y \geq 1$, with neighbor expansion applied only among positive counts.

Algorithm 3 DGCP Full: Outcome-Conditional CP for Counts

Require: Calibration data $\{(\hat{\mu}_i, Y_i)\}_{i=1}^{n}$, test predictions $\{\hat{\mu}_j\}_{j=1}^{n_{\text{test}}}$, significance α, minimum group size m, NCS function $s(\cdot,\cdot)$, minimum count $y_{\min} \in \{0,1\}$

Ensure: Prediction sets $\{\Gamma_j\}_{j=1}^{n_{\text{test}}}$

$y_{\max} \leftarrow \max_i(Y_i)$
Step 1: Build DGCP groups via symmetric neighbor expansion
for $y = y_{\min}$ **to** $y_{\max}$ **do**
    $n_y \leftarrow |\{i : Y_i = y\}|$
    if $n_y \geq m$ **then**
        $G_y \leftarrow \{y\}$
    else
        $G_y \leftarrow \{y\}, r \leftarrow 1$
        while $\sum_{k \in G_y} n_k < m$ **do**
            $G_y \leftarrow G_y \cup \{y - r, y + r\} \cap [y_{\min}, y_{\max}], \quad r \leftarrow r + 1$
        end while
    end if
end for
Step 2: Pre-compute hypothetical scores for each candidate
for $y = y_{\min}$ **to** $y_{\max}$ **do**
    $\mathcal{S}^{\text{hyp}}_{G_y}(y) \leftarrow \{s(\hat{\mu}_i, y) : Y_i \in G_y\}$
end for
// Step 3: Construct prediction sets
for $j = 1$ **to** n_{test} **do**
    $\Gamma_j \leftarrow \emptyset$
    for $\tilde{y} = y_{\min}$ **to** $y_{\max}$ **do**
        $p_{\tilde{y}} \leftarrow \frac{|\{s \in \mathcal{S}^{\text{hyp}}_{G_{\tilde{y}}}(\tilde{y}) : s \geq s(\hat{\mu}_j, \tilde{y})\}| + 1}{|\mathcal{S}^{\text{hyp}}_{G_{\tilde{y}}}(\tilde{y})| + 1}$
        if $p_{\tilde{y}} > \alpha$ **then**
            $\Gamma_j \leftarrow \Gamma_j \cup \{\tilde{y}\}$
        end if
    end for
end for
return $\{\Gamma_j\}_{j=1}^{n_{\text{test}}}$

Theoretical Validity of DGCP Hybrid. The DGCP Hybrid approach requires careful theoretical justification, as it combines two different non-conformity measures within a single prediction procedure. We establish its validity through the following reasoning.

Stage 1: Binary Mondrian CP. For the zero/non-zero decision, we employ the Label-Conditional Mondrian Inductive Conformal Prediction (LCMICP) framework of Tsoumas and Papadopoulos (2024). The binary NCS is defined as:

$$s_{\text{binary}}(\hat{\mu}, y) = 1 - P(Y = y \mid \hat{\mu}) = \begin{cases} 1 - e^{-\hat{\mu}} & \text{if } y = 0 \\ e^{-\hat{\mu}} & \text{if } y > 0 \end{cases} \tag{4.14}$$

Under the Poisson model assumption, this score measures the "surprise" of observing the true class given the predicted mean. Calibrating separately on $\{i : Y_i = 0\}$ and $\{i : Y_i > 0\}$ provides

Algorithm 4 DGCP Hybrid: Two-Stage CP for Zero-Inflated Counts

Require: Calibration data $\{(\hat{\mu}_i, Y_i)\}_{i=1}^{n}$, test predictions $\{\hat{\mu}_j\}_{j=1}^{n_{\text{test}}}$, significance α, minimum group size m, NCS function $s(\cdot, \cdot)$
Ensure: Prediction intervals $\{[\ell_j, u_j]\}_{j=1}^{n_{\text{test}}}$

Partition calibration data
$\mathcal{C}_0 \leftarrow \{i : Y_i = 0\}, \quad \mathcal{C}_+ \leftarrow \{i : Y_i > 0\}$
// Stage 1: Binary Mondrian calibration for $Y = 0$
$\mathcal{S}_0 \leftarrow \{1 - \exp(-\hat{\mu}_i) : i \in \mathcal{C}_0\}$
Stage 2: DGCP for positive counts
$\{\Gamma_j^{(+)}\}_{j=1}^{n_{\text{test}}} \leftarrow \text{DGCP-FULL}(\{(\hat{\mu}_i, Y_i) : i \in \mathcal{C}_+\}, \{\hat{\mu}_j\}, \alpha, m, s, y_{\min} = 1)$
Construct hybrid prediction sets
for $j = 1$ **to** n_{test} **do**
    $\Gamma_j \leftarrow \Gamma_j^{(+)}$
    $s_0 \leftarrow 1 - \exp(-\hat{\mu}_j)$
    $p_0 \leftarrow \frac{|\{s \in \mathcal{S}_0 : s \geq s_0\}| + 1}{|\mathcal{S}_0| + 1}$
    if $p_0 > \alpha$ **then**
        $\Gamma_j \leftarrow \Gamma_j \cup \{0\}$
    end if
    $[\ell_j, u_j] \leftarrow [\min(\Gamma_j), \max(\Gamma_j)]$
end for
return $\{[\ell_j, u_j]\}_{j=1}^{n_{\text{test}}}$

the Mondrian guarantee:

$$\mathbb{P}(Y \in C(X) \mid Y = 0) \geq 1 - \alpha \quad \text{and} \quad \mathbb{P}(Y \in C(X) \mid Y > 0) \geq 1 - \alpha. \tag{4.15}$$

Stage 2: Conditional Count CP. For positive counts, we apply DGCP with hypothetical Poisson NCS, calibrated exclusively on observations with $Y > 0$. Within this stratum, the method functions as a Mondrian conformal predictor where the dynamically formed groups G_y act as the conditioning categories. Under Assumption 17 (which holds exactly for singleton groups; see the proof in Appendix 4.8.1), for each candidate count $\tilde{y} \geq 1$ the conformal p-value $p(\tilde{y})$ satisfies:

$$\mathbb{P}\big(p(\tilde{y}) \leq \alpha \mid Y_{\text{test}} = \tilde{y},\, Y_{\text{test}} > 0\big) \leq \alpha. \tag{4.16}$$

When sufficient data exists such that $G_y = \{y\}$ (i.e., class-specific calibration), this yields an exact per-class guarantee without requiring Assumption 17.

Combined Validity. The hybrid prediction set is constructed as the union of two disjoint sets:

$$C_{\text{hybrid}}(X) = C_{\text{binary}}^{(0)}(X) \cup C_{\text{count}}^{(+)}(X) \tag{4.17}$$

where $C_{\text{binary}}^{(0)}(X) \in \{\emptyset, \{0\}\}$ and $C_{\text{count}}^{(+)}(X) \subseteq \{1, 2, 3, \ldots\}$.

Assumption 17 (Conditional Exchangeability of Scores)**.** *Conditional on the outcome vector* $(Y_1, \ldots, Y_n, Y_{test})$ *and on* $\mathcal{I}_{y^*}$*, the random variables* $\{s(\hat{f}(X_i), y^*)\}_{i \in \mathcal{I}_{y^*}} \cup \{s(\hat{f}(X_{test}), y^*)\}$ *are exchangeable.*

This assumption holds exactly when $G_{y^*} = \{y^*\}$: in that case, all indices in $\mathcal{I}_{y^*} \cup \{\text{test}\}$ share the same outcome value y^*, so exchangeability of the underlying (X, Y) sequence implies

exchangeability of the X-values among these indices (see Aldous 1985, Section 1), and hence of the scores $s(\hat{f}(\cdot), y^*)$ applied to those X-values. More generally, the assumption is satisfied whenever the conditional distribution of $\hat{f}(X)$ given $Y = y$ is identical for all $y \in G_{y^*}$. The DGCP grouping mechanism supports this condition by construction: groups are formed from neighboring count values, which in regression settings arise from similar regions of the covariate space and therefore tend to have similar distributions of predicted values $\hat{f}(X)$.

Theorem 6 (Marginal Validity of DGCP Hybrid)**.** *Under the exchangeability of the pooled calibration and test data and Assumption 17, the DGCP Hybrid method (Algorithm 4) at significance level $\alpha \in (0, 1)$ guarantees marginal coverage:*

$$\mathbb{P}\left(Y_{test} \in C_{hybrid}(X_{test})\right) \geq 1 - \alpha.$$

Remark 7 (Per-Candidate vs. Class-Conditional Guarantees)**.** *The DGCP framework provides coverage guarantees* per candidate count value *$\tilde{y}$: for each $\tilde{y}$, the conformal p-value $p(\tilde{y})$ is computed using the calibration neighborhood $G_{\tilde{y}}$, and is valid under exchangeability within that neighborhood. For counts y with abundant calibration data ($|\mathcal{C}_y| \geq m$), the neighborhood reduces to a singleton $G_y = \{y\}$, yielding a per-class guarantee. For rare counts, neighboring values are grouped together; the resulting neighborhoods may overlap across different candidate values, but the per-candidate validity is unaffected. This distinction is practically important: in typical insurance data, $Y = 0$ and $Y = 1$ often have sufficient calibration samples for class-specific guarantees, while $Y \geq 2$ are typically grouped together. The marginal validity guarantee (Theorem 6) holds regardless of the group structure, but practitioners should interpret conditional coverage results at the appropriate granularity; per candidate value rather than over a partition of the outcome space.*

Remark 8 (On Assumption 17 in Practice)**.** *In practice, when G_{y^*} contains multiple count values but the conditional distributions of $\hat{f}(X)$ given $Y = y$ are similar across these values, Assumption 17 holds approximately. This is the typical situation in insurance claims frequency modeling: policyholders filing 2 versus 3 claims tend to occupy similar regions of the covariate space, yielding similar distributions of predicted means. The simulation results in Section 4.5 confirm that DGCP Hybrid maintains nominal coverage across all tested configurations, including settings where non-singleton groups are formed, providing empirical support for the assumption in the count data context. Moreover, when the minimum group size m is large enough that all positive counts collapse into a single group (as with $m = 200$ in our application), the method reduces to binary label-conditional CP (Tsoumas and Papadopoulos 2024), for which validity follows from the standard Mondrian guarantee without requiring Assumption 17.*

Scope of Coverage Guarantees. We emphasize the precise scope of the guarantees provided by DGCP Hybrid.

- **Guaranteed (Theorem 6):** Marginal validity of the final hybrid prediction set, i.e., $\mathbb{P}(Y_{\text{test}} \in C_{\text{hybrid}}(X_{\text{test}})) \geq 1 - \alpha$, under exchangeability of the pooled calibration and test data.
- **Guaranteed (per-stage):** Within Stage 1, label-conditional validity for the binary partition $\{Y = 0\}$ vs. $\{Y > 0\}$, by the Mondrian CP guarantee. Within Stage 2, for each candidate count $\tilde{y}$, the conformal p-value $p(\tilde{y})$ is valid with respect to the calibration neighborhood $G_{\tilde{y}}$ under exchangeability and Assumption 17 (the latter holding exactly for singleton groups).
- **Not guaranteed:** Full covariate-conditional coverage $\mathbb{P}(Y \in C(X) \mid X = x) \geq 1 - \alpha$ for all $x \in \mathcal{X}$. The cluster-based diagnostics in Section 4.6.3 provide empirical evidence of

approximate conditional coverage across portfolio segments, but this remains a diagnostic finding rather than a theoretical guarantee. Achieving distribution-free conditional coverage is known to be impossible without further structural assumptions (Vovk et al. 2022, Foygel Barber et al. 2020).

On the Structure of DGCP Groups. Algorithm 3 constructs, for each candidate count value $\tilde{y}$, a calibration neighborhood $G_{\tilde{y}}$ via symmetric expansion. These neighborhoods are *not* necessarily disjoint: two distinct candidates $\tilde{y} \neq \tilde{y}'$ may share calibration points if their neighborhoods overlap (i.e., $G_{\tilde{y}} \cap G_{\tilde{y}'} \neq \emptyset$). Crucially, this does not invalidate the coverage guarantee. The prediction set is constructed as

$$\Gamma^{\alpha}(X_{n+1}) = \big\{\tilde{y} \in \mathcal{Y} : p(\tilde{y}) > \alpha\big\}, \tag{4.18}$$

where each conformal p-value $p(\tilde{y})$ is computed independently using the hypothetical scores from $G_{\tilde{y}}$. As shown in the proof of Theorem 6, the group-construction rule is a symmetric function of the calibration outcomes. For singleton groups ($G_{y^*} = \{y^*\}$), conditioning on group membership preserves the exchangeability of scores by the exchangeability of the underlying data alone. For non-singleton groups, Assumption 17 ensures exchangeability of the hypothetical scores, which is a mild condition when groups contain neighboring count values (see Remark 8). In both cases, the p-value $p(y^*)$ for the true outcome y^* is super-uniform, and the overlapping structure of neighborhoods does not affect validity.

When the groups happen to form a partition (e.g., when m is large enough that all positive counts collapse into a single group, or when each count has sufficient data for singleton groups), the stronger group-conditional guarantee of Remark 7 applies. In general, however, the per-candidate validity holds without requiring a partition structure.

Why Independent Stages Are Valid. A natural concern is whether combining two separate conformal procedures compromises validity. The independence is justified because:

1. **Disjoint Calibration Sets:** Stage 1 uses only observations with $Y_i = 0$ for calibration. Stage 2 uses only observations with $Y_i > 0$. No calibration point informs both stages, preserving exchangeability within each.

2. **Separate Hypothesis Tests:** Stage 1 tests the single hypothesis $H_0 : Y = 0$. Stage 2 tests a family of hypotheses $H_0 : Y = \tilde{y}$ for $\tilde{y} \geq 1$ but does so within its own stratum ($Y > 0$). There is no multiple testing across strata because the calibration data and null hypotheses are segregated.

3. **Appropriate Score Separation:** The binary NCS $1 - e^{-\hat{\mu}}$ directly measures the model's assessed probability of the event $Y = 0$. The Poisson NCS $s(\hat{\mu}, y)$ measures the deviance for a specific count. Using distinct, purpose-built scores for the zero class and positive counts improves efficiency without invalidating the marginal guarantee.

An important special case arises when the minimum group size m is sufficiently large that all positive counts are grouped together. In this regime, DGCP Hybrid with large m reduces to the binary label-conditional CP described by Tsoumas and Papadopoulos (2024) for the claims/no-claims classification problem. Specifically, with $m = 200$ in our application in Section 4.6, all positive count values ($Y \in \{1, 2, 3\}$) collapse into a single "claims" group due to the scarcity of multiple-claim observations. The hypothetical scores for any candidate $\tilde{y} \geq 1$ are then computed using all positive-count calibration μ values:

$$\mathcal{S}^{\text{hyp}}_{\text{all}+}(\tilde{y}) = \{s(\hat{\mu}_i, \tilde{y}) : y_i > 0\} \tag{4.19}$$

This makes DGCP Hybrid with large m functionally equivalent to:

1. Binary Mondrian CP for the claim/no-claim decision
2. A count interval for claimants calibrated on all $Y > 0$ observations jointly

The Stabilizing Effect of Large m. Our simulation study in Section 4.5 reveals a subtle but important phenomenon: when m is small, the prediction intervals exhibit *contiguity-induced conservatism*. Although DGCP tests each candidate count value independently, the final prediction set is reported as a contiguous interval from $\ell = \min(\Gamma)$ to $u = \max(\Gamma)$. With small m, the limited calibration data for each count value leads to noisy quantile estimates. This occasionally includes isolated count values at the distribution tails (e.g., $y = 3$ and $y = 48$ might be included while intermediate values are excluded). The interval $[3, 48]$ then substantially over-covers, explaining the observed width-coverage trade-off in Figure 4.18. Conversely, large m values (e.g., $m = 200$) group more claims together, providing sufficient calibration data for stable quantile estimation. This eliminates sporadic tail inclusions and produces intervals that closely match the high-density region of the count distribution. For insurance claims frequency with typical calibration sizes ($n \geq 20\,000$), we recommend $m \geq 100$ to avoid contiguity-induced over-coverage. This ensures stable grouping that reflects the inherent structure of claims data – most claimants have 0 or 1 claims, with few having multiple claims. Thus, contrary to the medical diagnosis setting of Papaioannou et al. (2026) where $m = 10 - 20$ suffices, insurance claims frequency benefits from larger m values that leverage the limited count range to achieve both validity and efficiency.

4.4.2 Related Approaches

Two-Stage Conformal Approaches Two-stage conformal prediction has been applied in related settings, though neither addresses the specific challenge of outcome-conditional CP for discrete count data.

Graziadei et al. (2024) developed two-stage split conformal prediction for frequency-severity modeling in insurance, where the first stage predicts claim frequency (count) and the second stage predicts claim severity (continuous). Importantly, their conformal prediction intervals are constructed only for the continuous severity outcome, using locally-weighted conformity scores $R_i = |Y_i - \hat{\psi}(X_i, \hat{\mu}(X_i))|/\hat{\sigma}(X_i, \hat{\mu}(X_i))$. The two-stage structure serves to incorporate predicted frequency as an input feature to the severity model, rather than to provide uncertainty quantification for discrete outcomes. Diaz-Rincon et al. (2024) proposed a two-stage approach for zero-inflated continuous medication dosage prediction in Parkinson's disease. Their first stage uses classification to identify patients requiring medication changes, while the second stage applies conformal regression for the magnitude of change. They address zero-inflation through an adjusted conformal quantile:

$$\gamma = \max\left\{0, \min\left\{1, \frac{1 - \hat{\beta} - r}{1 - r}\right\}\right\} \tag{4.20}$$

where $\hat{\beta}$ estimates the proportion of true zeros among predicted zeros. While conceptually similar to our hybrid approach in addressing zero-inflation, their outcome remains continuous rather than discrete, and they do not provide outcome-conditional guarantees for specific count values.

Our two-stage adaptation differs from both approaches in three key respects: (1) we construct prediction sets for discrete count outcomes, requiring the hypothetical scores adaptation to maintain exchangeability; (2) we provide outcome-conditional coverage guarantees stratified by claim count, not just marginal or binary coverage; and (3) we employ Mondrian calibration within the binary stage while using DGCP grouping with neighbor expansion for the count stage.

Bin-Conditional Conformal Prediction Randahl et al. (2026) propose Bin-Conditional Conformal Prediction (BCCP) as an alternative approach for real-valued outcomes where the label space does not have natural categories. Rather than conditioning on exact outcome values or relying on semantic hierarchies, BCCP partitions the outcome space into user-defined bins and ensures coverage within each bin.

The key insight is that when outcomes are heavily right-skewed (as in conflict fatalities or insurance claims), standard CP systematically over-covers the dense region near zero and under-covers the sparse tail. BCCP addresses this by first defining bins based on ranges of substantive interest to the analyst, then computing separate quantile thresholds within each bin using only calibration points falling in that bin, and finally constructing prediction intervals that achieve coverage guarantees per bin. For count data, natural bins might be $Y = 0$ (no claims), $Y = 1$ (single claim), $Y \in [2, 5]$ (moderate frequency), and $Y \geq 6$ (high frequency).

The fundamental trade-off in BCCP is between interval width and coverage uniformity. Fewer bins lead to narrower intervals but may fail to provide consistent coverage across the outcome space, while more bins improve local calibration but result in wider intervals. As Randahl et al. (2026) demonstrate, the choice of bins should balance accurate local coverage with reasonably tight prediction intervals. BCCP may produce discontiguous prediction sets when intervals from different bins do not overlap, which can be addressed by either reporting the union of disjoint intervals or by contiguizing the final interval (taking the overall minimum and maximum), the latter approach being more conservative. BCCP and DGCP are complementary: BCCP allows arbitrary bin definitions without requiring semantic similarity between outcomes, while DGCP's dynamic grouping based on calibration data availability is more adaptive. We focus on DGCP as it naturally exploits the ordering of count values for neighbor-based grouping.

4.5 Simulation Study for Count Data Prediction Intervals

4.5.1 Introduction and Motivation

Evaluating prediction interval methods for count data presents a fundamental challenge: the true conditional distribution of claims is unknown in observational data, making direct assessment of interval validity impossible. To overcome this limitation, we conduct a comprehensive simulation study using four distinct data-generating processes (DGPs) that capture the key distributional features encountered in insurance claims frequency modeling. Unlike real-world applications where claim counts are typically concentrated at zero with occasional small positive values, the DGPs employed here are designed to generate a broader range of count outcomes. This design choice enables a rigorous comparison of conformal prediction methods against parametric benchmarks across the full spectrum of count distributions, including scenarios with substantial counts that stress-test the methods' tail behavior.

The simulation study serves three primary objectives. First, we assess whether each method achieves the nominal marginal coverage guarantee of $1 - \alpha = 0.90$ across different distributional assumptions. Second, we evaluate outcome-conditional coverage to determine whether methods maintain validity across outcome bins that reflect varying claim frequencies. Third, we examine the efficiency-validity trade-off by comparing interval widths among methods that achieve valid coverage.

Our selection of five benchmarks spans key methodological dimensions: Two methods assume equidispersion (Poisson-based), while three accommodate overdispersion (NB-based), allowing assessment of conformal methods under both correct and misspecified distributional assumptions.

The normal approximation methods ($\tilde{\Gamma}_1$, $\tilde{\Gamma}_3$) provide computationally efficient baselines, while bootstrap methods ($\tilde{\Gamma}_2$, $\tilde{\Gamma}_5$) offer more complete uncertainty quantification at higher computational cost. The Chebyshev method ($\tilde{\Gamma}_4$) represents a conservative, moment-based approach that provides valid coverage under weak assumptions, contrasting with the tighter but potentially undercovering normal approximations. A summary of the literature on prediction-interval approaches for the Poisson and Negative Binomial distributions is provided in Appendix 4.8.2. Table 4.2 summarizes the key characteristics of each approach.

Table 4.2: Summary of prediction interval methods.

Method	**Description**	**Guarantee**
Conformal Methods		
Poisson Deviance ICP	Standard inductive CP with Poisson deviance NCS	Marginal
ZA Relative ICP	Standard ICP with zero-adjusted relative NCS	Marginal
Cluster Conditional CP	Mondrian CP using actuarial risk segments	Cluster-conditional
Binary Marginal CP	Marginal CP for binary claim/no-claim classification	Marginal
DGCP Hybrid m20	Two-stage DGCP with $m = 20$ minimum group size	Group-conditional
DGCP Hybrid m200	Two-stage DGCP with $m = 200$; approximates binary CP	Group-conditional
DGCP Hybrid m200 max1	Two-stage DGCP with $m = 200$, maximum count $y_{\max} = 1$	Binary-conditional
DGCP Full m20	Full DGCP on all counts with $m = 20$	Group-conditional
DGCP Full m200	Full DGCP on all counts with $m = 200$	Group-conditional
Parametric Baseline Methods		
Parametric Poisson	Normal approximation Poisson intervals; lower coverage for small n; assumes equidispersion	Asymptotic†
Bootstrap Poisson	Residual bootstrap from Pearson residuals; computationally intensive	Parametric*
NB Bootstrap	Negative binomial parametric bootstrap; computationally intensive	Asymptotic†
NB Plug-in	Extends normal approximation to overdispersed data; requires α estimation	Parametric**
NB Chebyshev	Conservative one-sided; only requires first two moments; robust to model misspecification	Parametric*

†Requires correct model specification and large sample

*Finite-sample valid under correct model specification

**Conservative finite-sample under weak moment assumptions

4.5.2 Data Generating Processes

We consider four Data Generating Processes (DGPs) representing progressively complex departures from the standard Poisson assumption, each generating claim counts $Y_i \in \mathbb{N}_0$ conditional on a d-dimensional covariate vector X_i and exposure $v_i > 0$. The covariate effects enter through a linear predictor $\eta_i = X_i^\top \beta$ with heterogeneous coefficient magnitudes designed to reflect actuarial practice. The **Poisson DGP** provides the baseline specification with $Y_i \mid X_i \sim \text{Poisson}(\mu_i)$ where $\mu_i = v_i \cdot \exp(X_i^\top \beta)$; while conditionally equidispersed, covariate heterogeneity induces marginal overdispersion. The **Zero-Inflated Poisson (ZIP) DGP** introduces structural zeros through a

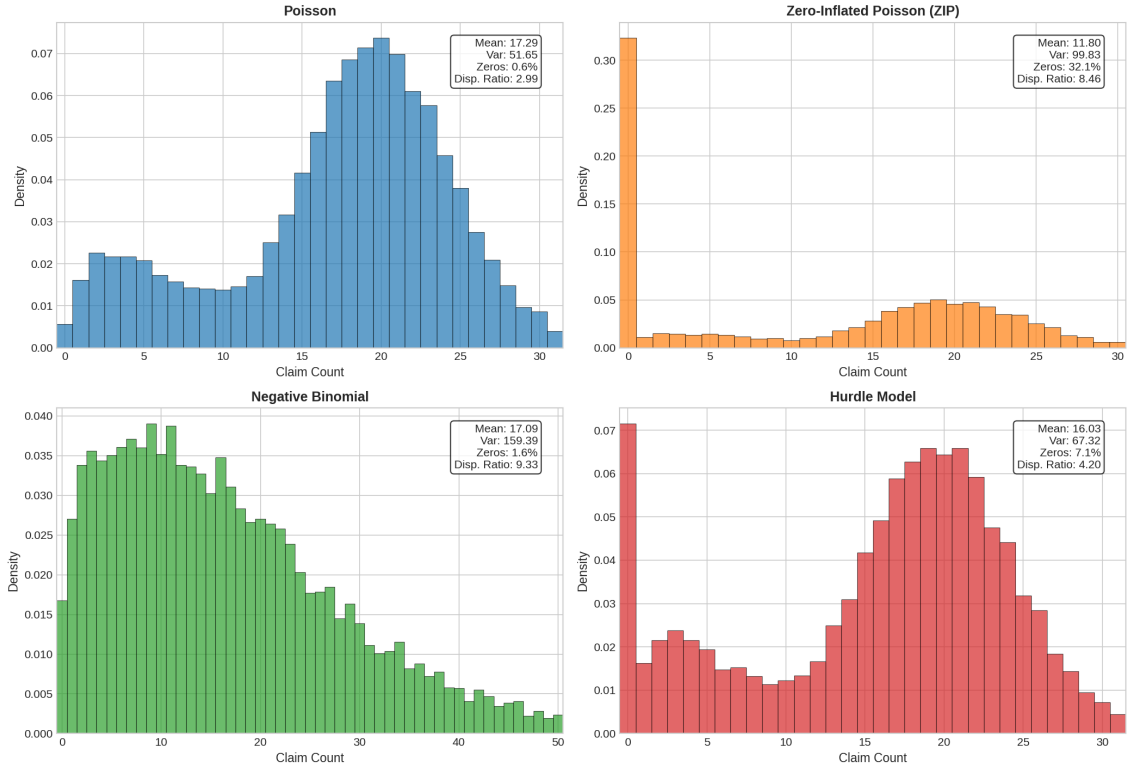

Figure 4.2: Distribution of claim counts under each DGP at baseline configuration, showing empirical density with summary statistics.

two-component mixture with covariate-dependent zero-inflation probability π_i, inducing both excess zeros and overdispersion. The **Negative Binomial DGP** accommodates overdispersion through gamma-mixed Poisson formulation with shape parameter θ, yielding variance $\mu_i + \mu_i^2/\theta$. The **Hurdle DGP** separates the zero-generating process from positive counts via a zero-truncated Poisson distribution, allowing distinct mechanisms for claim occurrence and frequency. Figure 4.2 displays the empirical distributions of claim counts under each DGP at the baseline configuration. The Poisson DGP exhibits a right-skewed distribution centered around a mean of approximately 17 claims, with variance exceeding the mean due to the heterogeneity induced by the covariate structure and interaction effects. The ZIP DGP shows a pronounced spike at zero comprising 32.1% of observations, substantially higher than the Poisson baseline, with the remaining mass distributed across positive counts. The Negative Binomial DGP demonstrates the most pronounced overdispersion, with a variance-to-mean ratio of 9.33 and a heavier right tail. The Hurdle DGP produces an intermediate zero proportion of 7.1% with a distinct separation between the zero mass and the positive count distribution. Figure 4.3 provides a comparative view of the key distributional characteristics across DGPs. The left panel confirms that the ZIP DGP generates the highest proportion of zeros, followed by the Hurdle model, while the Poisson and Negative Binomial DGPs produce zero rates consistent with their respective distributional assumptions. The center panel illustrates the dispersion ratios, with all DGPs exhibiting overdispersion relative to the classical Poisson equidispersion benchmark due to the heterogeneity induced by covariate effects and interactions. The Negative Binomial DGP shows the most extreme departure from equidispersion, followed by the ZIP model. The right panel displays boxplots of the count distributions, highlighting the differences in central tendency, spread, and tail behavior across DGPs.

The covariate structure combines continuous variables (policyholder age, vehicle age, policy duration) with categorical factors (geographic region, coverage type), designed to mimic realistic

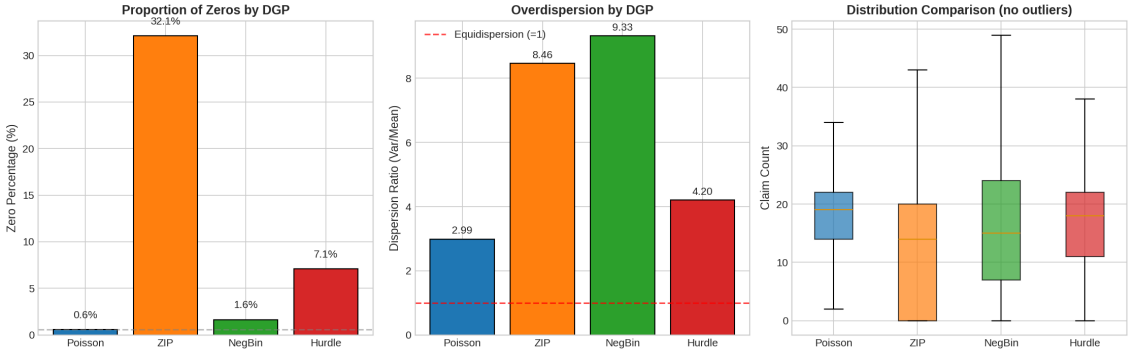

Figure 4.3: Comparison of DGP characteristics: proportion of zeros (left), dispersion ratio (center), and count distributions (right).

insurance rating factors. The baseline configuration specifies $N = 10{,}000$ observations with $d = 10$ covariates. Sensitivity analyses vary sample size ($N \in \{1000, \ldots, 10000\}$), covariate dimension ($d \in \{10, 20, 50\}$), zero-inflation level ($\pi_0 \in \{0.0, 0.1, 0.3, 0.5\}$), overdispersion ($\theta \in \{2.0, 3.0, 4.0, 5.0\}$), and the inclusion of interaction and spatial effects. Full details of the covariate distributions and parameter configurations are given in Appendix 4.8.3. The simulation is run on the HPC Hummel Cluster at the University of Hamburg with 96 repetitions using different random seeds. Additional computational information is reported in Appendix 4.8.3.

4.5.3 Prediction Models

For each simulated dataset, we partition the observations into training (60%), calibration (20%), and test (20%) sets. Three prediction models are fitted on $\mathcal{D}_{\text{train}}$ to obtain the predicted mean $\hat{\mu}_i = \hat{f}(X_i)$ for each observation. For the implementation of the models, we follow Wüthrich et al. (2026). The **Poisson GLM** specifies the canonical log-linear model

$$\log(\mu_i) = X_i^\top \beta + \log(v_i), \tag{4.21}$$

with coefficients β estimated by maximum likelihood (Nelder and Wedderburn 1972, McCullagh and Nelder 1989). This represents the standard actuarial approach and serves as the correctly specified model under the Poisson DGP. LightGBM (Ke et al. 2017) with a Poisson deviance objective provides a gradient-boosting alternative that can capture nonlinear relationships and interactions without explicit specification.[1]. Random Forest (Breiman 2001) with a Poisson splitting criterion offers a bagging-based ensemble method with automatic regularization[2]. All models output predictions $\hat{\mu}_i$ are clipped to $[10^{-6}, 100]$ to ensure numerical stability in downstream non-conformity score calculations.

We note that this is not an extensive modeling approach but rather serves to test conformal prediction methods under different underlying model structures and assumptions. For more advanced approaches to claim frequency modeling, one could consider neural network extensions such as the Combined Actuarial Neural Network (CANN) architecture of Wüthrich and Merz (2019), which embeds classical GLM structures into feedforward networks while preserving interpretability; the LocalGLMnet of Richman and Wüthrich (2023b,a), which learns feature-dependent regression coefficients through attention mechanisms; or the systematic nested approach of Schelldorfer

[1] Hyperparameters selected via 5-fold cross-validation: 200 trees, 31 leaves, learning rate 0.05, and L1/L2 regularization parameters of 2.0.

[2] Hyperparameters selected via 5-fold cross-validation: 100 trees, maximum depth 10, minimum leaf size 20, and square-root feature subsampling.

and Wüthrich (2019), which demonstrates how neural networks can identify missing interactions in traditional actuarial models. These architectures offer potential improvements in predictive performance while maintaining connections to interpretable GLM foundations.

4.5.4 Evaluation Metrics

We assess each method using metrics that capture both validity and efficiency. The empirical **marginal coverage** on the test set is $\widehat{\text{Cov}} = n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\{Y_i \in [\ell_i, u_i]\}$, where $[\ell_i, u_i]$ is the prediction interval for test observation i. The target is $1 - \alpha = 0.90$, and we consider a method valid if $\widehat{\text{Cov}} \geq 0.88$, allowing for sampling variability. For outcome group $\mathcal{G} \subseteq \mathcal{Y}$, the **outcome-conditional coverage** is $\widehat{\text{Cov}}_{\mathcal{G}} = \sum_{i:Y_i \in \mathcal{G}} \mathbf{1}\{Y_i \in [\ell_i, u_i]\} / \sum_{i:Y_i \in \mathcal{G}} 1$. We evaluate coverage for $\mathcal{G} \in \{\{0\}, \{1\}, \{2, 3, \ldots\}\}$ as well as binned groups $\{1, \ldots, 5\}$, $\{6, \ldots, 10\}$, $\{11, \ldots, 20\}$, and $\{21, \ldots, 50\}$. Efficiency is measured by mean **interval width** $\overline{W} = n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} (u_i - \ell_i)$. The **interval score** (Winkler 1972, Gneiting and Raftery 2007) provides a proper scoring rule that jointly penalizes width and miscoverage:

$$\text{IS}_\alpha = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left[(u_i - \ell_i) + \frac{2}{\alpha}(\ell_i - Y_i)\mathbf{1}\{Y_i < \ell_i\} + \frac{2}{\alpha}(Y_i - u_i)\mathbf{1}\{Y_i > u_i\} \right]. \tag{4.22}$$

4.5.5 Simulation Results

Figure 4.4 presents the coverage-width trade-off averaged across all four DGPs in the baseline configuration. The results reveal a clear stratification of methods into three behavioral clusters that align with their underlying design principles.

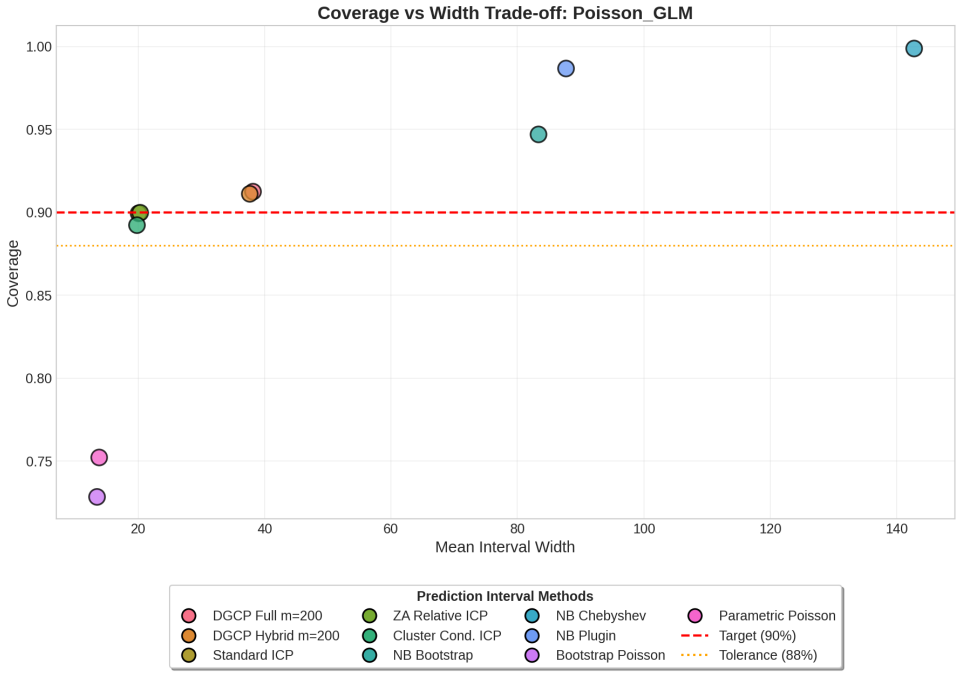

Figure 4.4: Coverage vs. mean interval width trade-off averaged across all DGPs in the baseline simulation setting. Methods cluster into three groups: efficient ICP methods achieving target coverage with narrow intervals, DGCP methods producing wider intervals with mild over-coverage, and parametric benchmarks exhibiting either under-coverage (Poisson) or extreme width (Negative Binomial).

The inductive conformal prediction methods, with Poisson deviance nonconformity scores (Standard ICP) and ZA Relative ICP, achieve the lowest mean interval widths while maintaining the target 90% marginal coverage. These methods exploit the concentration of the count distribution around small values to produce tight intervals that satisfy the coverage guarantee through their calibration on the empirical residual distribution. Cluster Conditional ICP produces intervals of similar width but exhibits mild under-coverage of approximately 2 percentage points, a consequence of the finite-sample variability introduced by stratified calibration across clusters with potentially imbalanced sizes.

The DGCP methods occupy an intermediate position in the trade-off space, producing wider intervals than standard ICP while achieving slight over-coverage. This behavior reflects the intentional design choice underlying DGCP: by calibrating separately within outcome-based groups, these methods sacrifice some marginal efficiency to ensure validity across the outcome distribution. The over-coverage arises because the proposed methods target the most conservative threshold across outcome groups, and the additional width represents the price paid for outcome-conditional guarantees[3]. DGCP Hybrid produces narrower intervals than DGCP Full, reflecting its more targeted treatment of zero outcomes versus positive counts.

The parametric benchmark methods reveal the consequences of distributional misspecification in opposite directions. The Poisson-based methods (Parametric Poisson, Bootstrap Poisson) assume equidispersion and produce systematic under-coverage when applied to the overdispersed and zero-inflated DGPs that characterize the simulation design. The intervals they construct are too narrow because they underestimate the true conditional variance, particularly for observations with high predicted means where the quadratic variance component of overdispersion becomes dominant. The Negative Binomial benchmarks err in the opposite direction, achieving or exceeding target coverage but at the cost of extremely wide intervals that provide little practical utility. These methods overestimate uncertainty by assuming a level of overdispersion calibrated to accommodate worst-case scenarios, resulting in intervals that are conservative to the point of being uninformative.

Stability Across Base Learners A key practical consideration for conformal prediction methods is whether their coverage properties depend on the choice of underlying predictive model. Figure 4.5 addresses this question by displaying coverage and width metrics separately for Poisson GLM, LightGBM, and Random Forest base learners across the conformal prediction methods.

The results demonstrate stability across base learners. Standard ICP and ZA Relative ICP achieve approximately 90% coverage regardless of whether the underlying predictions come from a parametric GLM, gradient boosting, or random forest ensemble. The interval widths vary slightly. LightGBM tends to produce marginally narrower intervals due to its superior point prediction accuracy, but the coverage levels remain stable. This stability confirms the theoretical property that conformal prediction calibrates on residuals rather than model structure, allowing the coverage guarantee to hold regardless of how the point predictions are generated, provided only that the calibration and test data are exchangeable.

The DGCP methods similarly maintain their coverage properties across learners, with the mild over-coverage observed for Poisson GLM persisting for the tree-based methods. Cluster Conditional ICP shows the most variation across learners, with coverage ranging from approximately 87% to 89%, reflecting the sensitivity of stratified calibration to the cluster assignments induced by different predictive models. However, even this variation remains within acceptable tolerances for practical applications.

[3]To narrow them further, one could use non-contiguous sets. On an individual level, however, such sets are difficult to explain. One could also consider larger values of the hyperparameter m but a more detailed and theoretical analysis of the optimal choice of m in relation to sample size and distributional characteristics is beyond the scope of this paper.

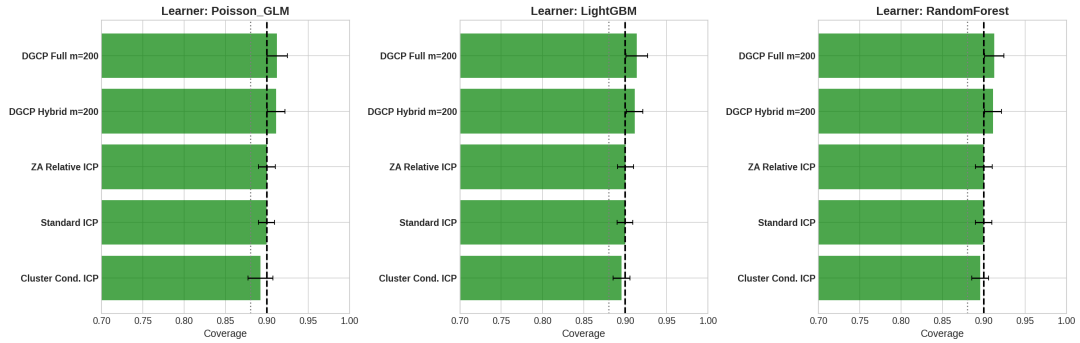

Figure 4.5: Coverage and interval width by base learner for conformal prediction methods. Left: Poisson GLM. Center: LightGBM. Right: Random Forest. The relative performance of methods remains consistent across learners, confirming that coverage properties are driven by the conformal methodology rather than the predictive model.

Robustness to Zero-Inflation Zero-inflation represents a particularly challenging scenario for prediction interval methods, as it introduces a bimodal structure to the conditional distribution that violates the assumptions underlying many parametric approaches. Figure 4.6 examines how coverage evolves as the zero-inflation parameter π_0 increases from 0 to 0.5 under both the ZIP and Hurdle DGPs.

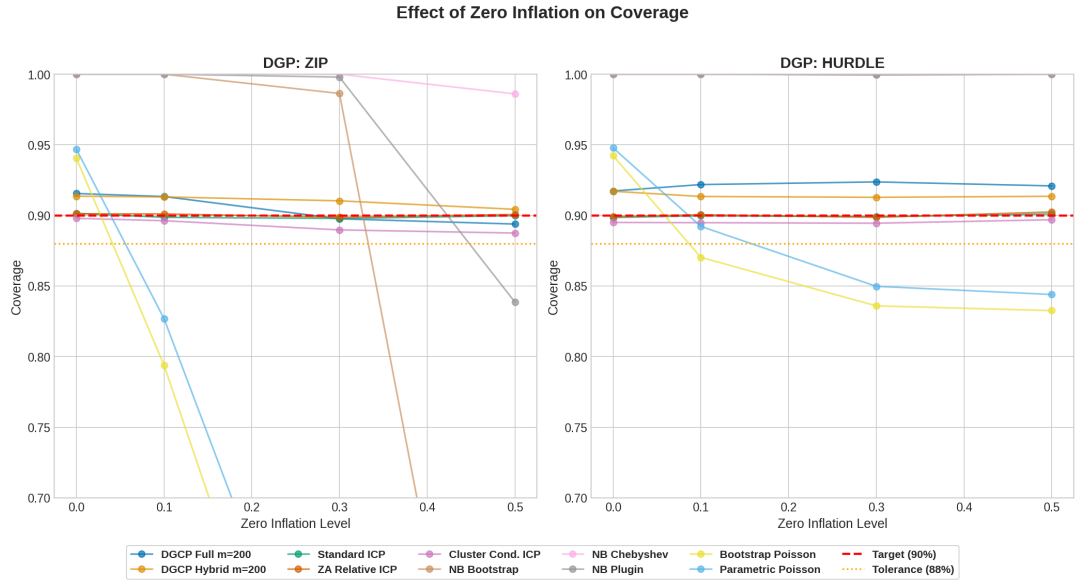

Figure 4.6: Impact of zero-inflation level on marginal coverage for ZIP (left) and Hurdle (right) DGPs. Cluster Conditional ICP exhibits increasing under-coverage as zero-inflation rises. DGCP Hybrid converges toward target coverage as zero-inflation increases, reflecting its design for this setting. Parametric methods show inconsistent behavior across zero-inflation levels.

The Cluster Conditional method exhibits a gradual decline in coverage as zero-inflation increases, dropping from approximately 89% at $\pi_0 = 0$ to around 85% at $\pi_0 = 0.5$ under the ZIP DGP. This degradation occurs because the cluster-based stratification does not align with the zero/positive structure of the data; observations with $Y = 0$ may be grouped with observations having positive counts if their covariates are similar, leading to miscalibrated thresholds for the zero class.

DGCP Full demonstrates interesting non-monotonic behavior: at low zero-inflation levels, it produces mild over-coverage of approximately 92%, but as zero-inflation increases toward extreme

values, coverage drops toward and slightly below the 90% target. This pattern reflects the tension between the method's symmetric treatment of outcome groups and the increasingly asymmetric structure of the data. When zeros dominate the distribution, the hypothetical score calculations for positive outcomes are based on limited calibration data, potentially leading to undercoverage for those outcomes.

DGCP Hybrid, by contrast, shows the behavior for which it was designed: as zero-inflation increases, its coverage converges toward the target 90% level from above. At low zero-inflation, the method over-covers by approximately 3 percentage points, but this over-coverage diminishes as the proportion of zeros increases. The specialized treatment of zero outcomes through probability-based scores, combined with the grouping mechanism that pools rare positive outcomes, allows DGCP Hybrid to maintain stable performance across the full range of zero-inflation levels. Under the Hurdle DGP, similar patterns emerge, though the magnitude of effects is somewhat attenuated due to the different generative mechanism for zeros.

Outcome-Conditional Coverage Analysis The aggregate coverage metrics examined above may mask important heterogeneity in performance across different regions of the outcome distribution. Figure 4.7 presents heatmaps of outcome-conditional coverage for each combination of DGP and method, with outcomes grouped into bins corresponding to the regions of the DGPs shown in Figure 4.2: $\{0\}$, $\{1, \ldots, 5\}$, $\{6, \ldots, 10\}$, $\{11, \ldots, 20\}$, and $\{21, \ldots, 50\}$. Standard ICP and Cluster Conditional ICP exhibit substantial heterogeneity in outcome-conditional coverage across both DGPs and outcome groups. Under the Poisson DGP, Standard ICP achieves near-zero coverage for $y = 0$, while for the ZIP, NegBin, and Hurdle DGPs, it exceeds 95% coverage for observations in the range $\{11, \ldots, 20\}$. This extreme imbalance arises because the calibration threshold is set by the empirical distribution of nonconformity scores, which in the Poisson case is dominated by observations with moderate counts. The resulting intervals systematically exclude zero for low-risk observations while being overly conservative for high-count observations. Under the Negative Binomial DGP, the pattern reverses somewhat due to the heavier tails and higher variance, but substantial heterogeneity persists.

The parametric benchmarks show different patterns of outcome-conditional failure. Parametric Poisson and Bootstrap Poisson achieve reasonable coverage for the $Y = 0$ group but systematically undercover positive counts, particularly in the $\{6, \ldots, 10\}$ and higher ranges where overdispersion effects are most pronounced. This reflects the fundamental misspecification: by assuming equidispersion, these methods construct intervals that are too narrow for the actual conditional variance at moderate to high predicted means.

DGCP Hybrid demonstrates notably more stable outcome-conditional coverage across all DGPs. The method maintains coverage between 85% and 95% for most DGP-outcome combinations, with the primary exception being mild over-coverage in the $\{6, \ldots, 10\}$ group under the Poisson, ZIP, and Hurdle DGPs. This over-coverage can be attributed to the neighbor-based grouping mechanism: when the minimum group size $m = 200$ is required, the relatively sparse observations in the 6–10 count range are pooled with neighboring outcomes, potentially including the more populous 1–5 group. The resulting calibration set draws heavily from the neighboring group, leading to conservative thresholds for the intermediate counts. Examining the outcome distributions in Figure 4.2 confirms that counts in this range represent a small fraction of observations under all DGPs except Negative Binomial, explaining why the grouping mechanism leads to over-coverage in this region. DGCP Full exhibits more variable performance, with under-coverage for $Y = 0$ in DGPs 1 and 3. In addition, we observe more severe under-coverage for the bin $Y \in [1, 5]$.

Summary of Simulation Findings The simulation study yields several conclusions that inform the practical application of prediction interval methods for count data. Standard

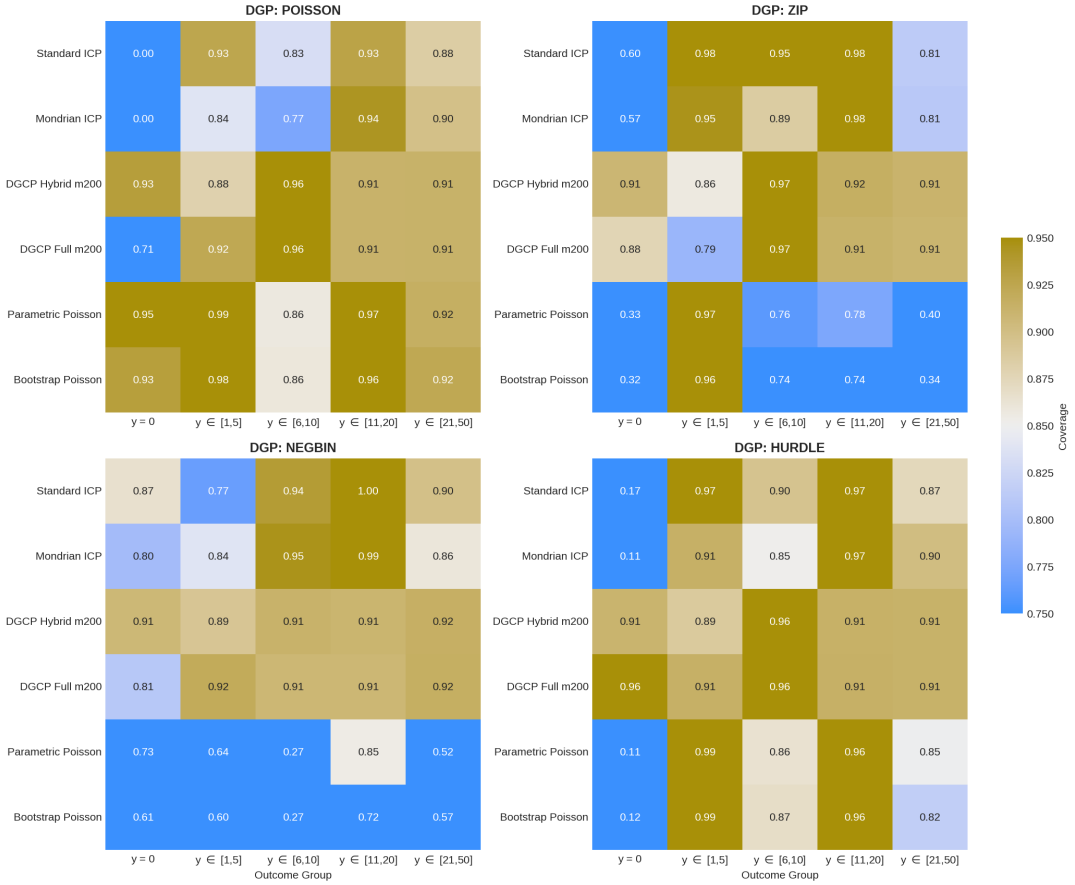

Figure 4.7: Outcome-conditional coverage heatmaps by DGP: Poisson (upper left), ZIP (upper right), Negative Binomial (lower left), Hurdle (lower right). Rows represent methods; columns represent outcome groups. Color scale indicates coverage from 0.75 (red) to 0.95 (green). Standard ICP and Cluster Conditional ICP show substantial variation across outcome groups, while DGCP Hybrid maintains more uniform coverage.

ICP methods achieve the most efficient intervals in terms of width but exhibit severe outcome-conditional coverage imbalance, systematically undercovering minority outcome classes while overcovering the majority. This imbalance renders them unsuitable for applications where coverage guarantees for specific outcome groups are required. Parametric benchmarks fail in predictable ways: Poisson-based methods undercover due to equidispersion misspecification, while Negative Binomial methods overcover through conservative variance estimates. Neither method adapts well to zero-inflation, which induces violations of their respective distributional assumptions.

DGCP Hybrid emerges as the method best suited for actuarial count data applications. It maintains approximately balanced outcome-conditional coverage across DGPs, adapts appropriately to varying levels of zero-inflation, and produces intervals of reasonable width given its stronger coverage objectives. The mild over-coverage for intermediate count ranges represents an acceptable price for the stability and balance achieved across the outcome distribution. The method's performance is robust to the choice of base learner, confirming that practitioners can combine DGCP calibration with their preferred predictive modeling framework without sacrificing coverage properties.

4.6 Empirical Application: German Motor Insurance

This section evaluates conformal prediction methods on a real-world insurance claims dataset from a German insurer, demonstrating the practical relevance of distribution-free prediction intervals for actuarial applications.

4.6.1 Data and Modeling Framework

The dataset comprises 579,456 policy-year observations from comprehensive motor insurance (*Vollkasko*) policies[4]. Table 4.3 shows the highly zero-inflated claim count distribution characteristic of insurance data: 89.7% of policies record no claims, 10.2% file exactly one claim, and fewer than 0.1% experience multiple claims within a single policy year. The observed claim frequency is approximately 12.16% when accounting for exposure. This extreme class imbalance, where the minority class of claimants represents only about one-tenth of observations, creates the fundamental challenge that motivates our methodological contributions. The dataset includes 11 categorical variables (`Brand`, `GaragePresent`, `Company`, `DrivAge`, `YoungDriv`, `VehAge`, `Mileage`, `ContractDuration`, `UserFixed`, `FedState`, `Rural`) and 7 numerical variables (`SF_Class`, `PopDensity`, `MedianIncome`, `Perc_Empl`, `Altitude`, `TypeClass_VK`, `SF_Class_KH`). Full variable descriptions are provided in Appendix 4.8.4.

Table 4.3: Distribution of claim counts

Claims	Count	Percentage
0	519,535	89.7%
1	59,387	10.2%
2	531	0.09%
3	3	<0.01%

Following the methodology of Wüthrich and Merz (2023) and Wüthrich et al. (2026), we fit Poisson and Negative Binomial GLMs with log link and exposure offset, alongside gradient boosting (LightGBM) and Random Forest benchmarks. We consider four nested Poisson GLM specifications: GLM1 uses full categorical dummy encoding for all rating factors without grouping rare categories; GLM2 groups infrequent categories (frequency $< 2\%$) into residual classes to reduce dimensionality while preserving main effects; GLM3 adds polynomial transformations of driver age (DrivAge^1, DrivAge^2, DrivAge^3, DrivAge^4, $\log$(DrivAge)) to capture nonlinear age-risk relationships; GLM4 further includes interaction terms between driver age and bonus-malus (DrivAge $\times$ SF_Class, DrivAge^2 $\times$ SF_Class) to model differential claim patterns across experience groups. Table 4.4 presents the complete model comparison results. Note, the AICs can only be compared for Poisson and NB separately, due to different outcome definitions.

The estimated Negative Binomial dispersion parameters $\hat{\alpha} \approx 1.0$ across all specifications indicate negligible overdispersion after accounting for covariates, suggesting the comprehensive feature set adequately captures policyholder heterogeneity. All four GLM specifications achieve comparable out-of-sample Poisson deviance (0.4545–0.4547), with the full categorical model (GLM1) marginally outperforming the parsimonious alternatives[5]. LightGBM achieves the lowest test deviance (0.4526), representing a modest 0.4% improvement over GLM1. The deviance reduction from the null model is approximately 3.8%, indicating that although the covariates

[4]For reasons of anonymization, we explicitly leave the time frame open.

[5]Wüthrich and Merz (2023) point out that dummy coding of categorical features with many levels can lead to near-collinearity in the design matrix and increased parameter uncertainty. One could also argue for a model with fewer parameters, such as GLM2–4, given the negligible difference in out-of-sample deviance. The trade-off between interpretability and marginal predictive gains favours parsimonious specifications in practice. More broadly, the differences across specifications are negligible and their ranking depends on the choice of deviance metric; the negative binomial extension primarily affects variance modeling rather than mean prediction accuracy at this frequency level.

Table 4.4: Comparison of predictive performance for the null models (Poisson and negative binomial) corresponding GLMs, Random Forest and LightGBM, including run times, number of parameters, AICs, in-sample and out-of-sample deviance losses, and test sample average frequencies (following Wüthrich and Merz (2023), Table 5.7).

Model	**# Params**	**Run Time**	**AIC**	$\textbf{Dev}_{\textbf{Train}}$	$\textbf{Dev}_{\textbf{Test}}$	$\hat{\mu}_{\textbf{Test}}$
Poisson Null	1	0.5s	314,592	0.4712	0.4725	12.1622%
Poisson GLM1	96	19.0s	306,419	0.4532	0.4545	12.1591%
Poisson GLM2	75	15.3s	306,430	0.4533	0.4546	12.1604%
Poisson GLM3	70	14.8s	306,431	0.4534	0.4547	12.1601%
Poisson GLM4	72	14.8s	306,431	0.4534	0.4547	12.1601%
NB Null	2	0.5s	318,987	0.4712	0.4725	12.2067%
NB GLM1	97	38.6s	311,667	0.4532	0.4546	12.2250%
NB GLM2	76	32.0s	311,675	0.4533	0.4547	12.2259%
NB GLM3	71	30.3s	311,675	0.4534	0.4547	12.2260%
NB GLM4	73	31.2s	311,675	0.4534	0.4547	12.2266%
LightGBM	–	4.7s	–	0.4461	0.4526	12.1416%
Random Forest	–	80.5s	–	0.4378	0.4541	12.1523%

Note: Dev = Poisson Deviance Loss. Observed freq: Train = 12.1622%, Test = 12.1786%

explain statistically significant variation, substantial residual heterogeneity remains. This is a common finding in insurance frequency modeling and underscores the importance of proper uncertainty quantification.

The variable importance analysis via likelihood ratio tests from drop-one analysis can be found in Appendix 4.8.4 Table 4.9 identifies annual mileage (LRT = 1,933) as the strongest predictor, followed by the regulatory vehicle type classification (LRT = 1,013), vehicle age (LRT = 451), manufacturer brand (LRT = 387), and federal state (LRT = 166). Several variables including rural/urban classification, altitude, median income, and company type show marginal contributions conditional on other predictors, suggesting their effects are captured by correlated variables[6].

For the prediction interval analysis, we select Poisson GLM1 as the primary base learner, as it represents the standard actuarial approach and facilitates interpretation. Notably, the choice of base learner has a negligible impact on the relative performance of conformal prediction methods[7]. The coverage properties and efficiency rankings remain consistent across all base learners, suggesting that coverage behavior is driven by the conformal methodology rather than the underlying predictive model.

4.6.2 Prediction Interval Analysis

Examining the fundamental coverage–width trade-off, Figure 4.8 reveals distinct patterns across method classes. The parametric benchmarks (Parametric Poisson, Bootstrap Poisson, NB Bootstrap) all exceed the 90% target coverage substantially, producing conservative but uninformative prediction intervals. This over-coverage reflects the inherent difficulty of calibrating parametric methods for zero-inflated count data without producing intervals wide enough to trivially contain all plausible outcomes. The Inductive Conformal Prediction methods (Standard ICP, ZA Relative ICP, Cluster Conditional ICP) achieve the smallest mean interval widths among methods meeting marginal coverage requirements. However, this apparent efficiency proves deceptive upon closer examination of class-conditional coverage. The DGCP Hybrid methods achieve target coverage

[6]For more details on nested tests for GLMs we refer to Fahrmeir and Tutz (1994).

[7]For a more detailed comparison, see Figure 4.37.

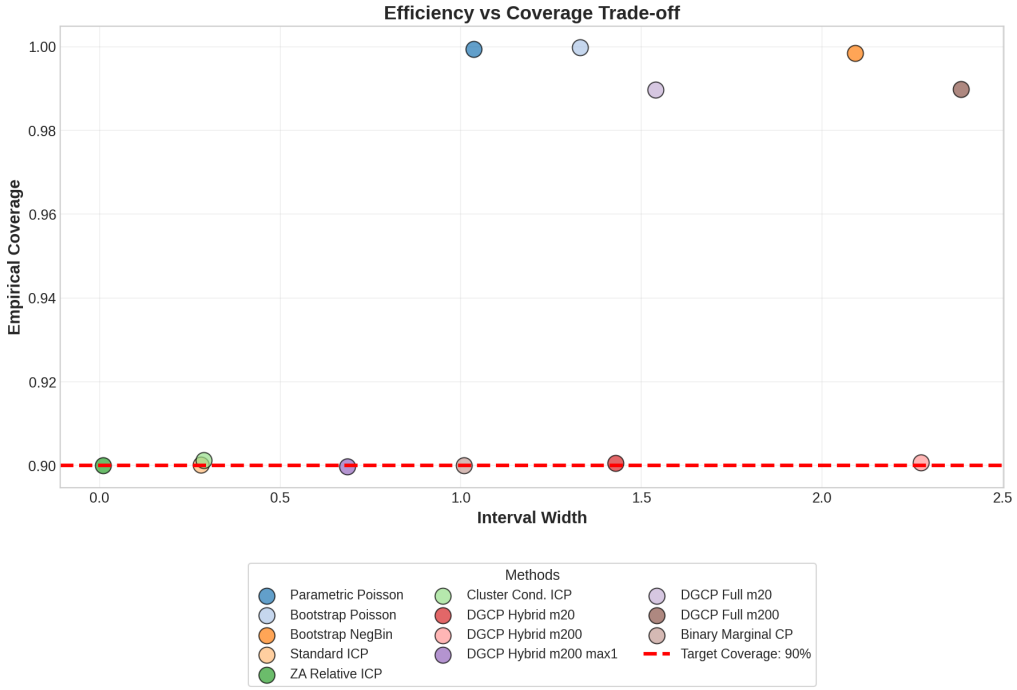

Figure 4.8: Coverage vs. mean interval width trade-off. The dashed red line indicates the target coverage of 90%.

with varying interval widths depending on the minimum group size parameter m: the variant with $m = 200$ and a maximum count restricted to 1 achieves the smallest width among DGCP methods by effectively reducing the problem to binary classification, while $m = 20$ produces slightly wider intervals but maintains granularity for multiple claims. The DGCP Full variants, by contrast, produce substantial over-coverage and fail to adapt to the extreme zero-inflation, treating all outcome groups symmetrically rather than exploiting the structure where nearly 90% of observations have zero claims.

Three metrics of direct actuarial relevance are summarized in Figure 4.9. The left panel confirms that DGCP methods achieve the target 90% coverage specifically for the claimant subpopulation, while simple ICP methods fall short at approximately 50%. This metric is crucial for actuarial applications where under-covering claims has direct financial implications through unanticipated loss experience. The center panel quantifies the mean degree of critical under-prediction, measuring the average shortfall when observed claims exceed the prediction interval upper bound (e.g., $y_i = 2$ but the interval is $[0, 0]$ or $[0, 1]$). Parametric benchmarks achieve values near zero through conservative wide intervals, while conformal DGCP variants exhibit values around 0.01, indicating occasional but limited under-prediction. The worst-performing methods here are Standard ICP and Cluster-Conditional ICP, with values around 0.05. This metric reveals that DGCP methods achieve efficiency gains relative to parametric benchmarks at the cost of occasional undercoverage for positive realizations, though the magnitude remains modest. The right panel displays the correlation between interval width and predicted claim probability; moderate correlation indicates that interval width reflects epistemic uncertainty about the prediction rather than simply tracking the aleatoric risk level[8].

The critical distinction between methods emerges when examining coverage separately for non-claimants ($Y = 0$), single-claim ($Y = 1$), and multiple-claim ($Y \geq 2$) groups, as shown in

[8]This metric provides only an indication; a more adaptive measure for this issue should be discussed.

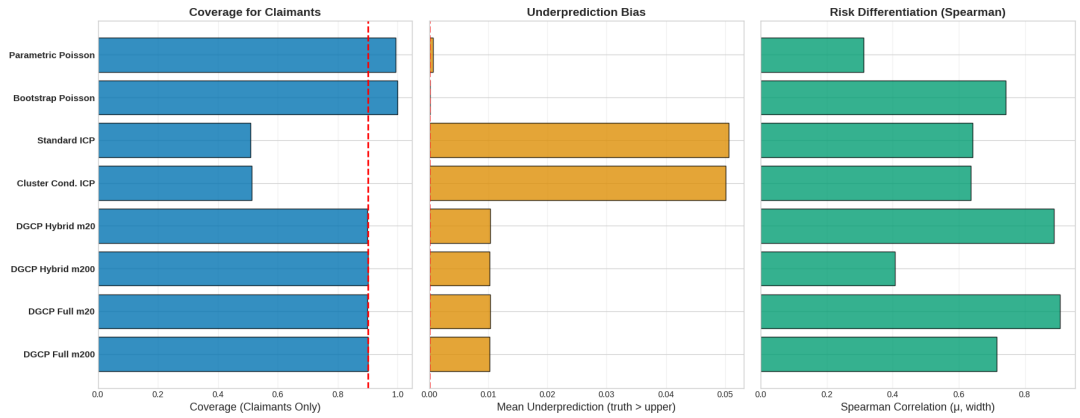

Figure 4.9: Actuarial performance metrics. **Left**: Coverage for claimants only ($Y > 0$). **Center**: Mean critical under-prediction, computed as the average shortfall when observed claims exceed the upper bound of the prediction interval. Values near zero indicate conservative intervals; higher values indicate more frequent and severe under-prediction events. **Right**: Spearman's rank correlation coefficient between interval width and predicted risk. High correlation (> 0.7) indicates intervals primarily reflect aleatoric risk rather than epistemic uncertainty.

Figure 4.10. This decomposition exposes the fundamental limitation of standard conformal methods for imbalanced count data. As predicted, standard ICP methods exhibit a stark coverage imbalance: they over-cover non-claimants (exceeding 95% coverage for $Y = 0$) while severely under-covering claimants (dropping to approximately 40–60% for $Y = 1$). This imbalance arises mechanically

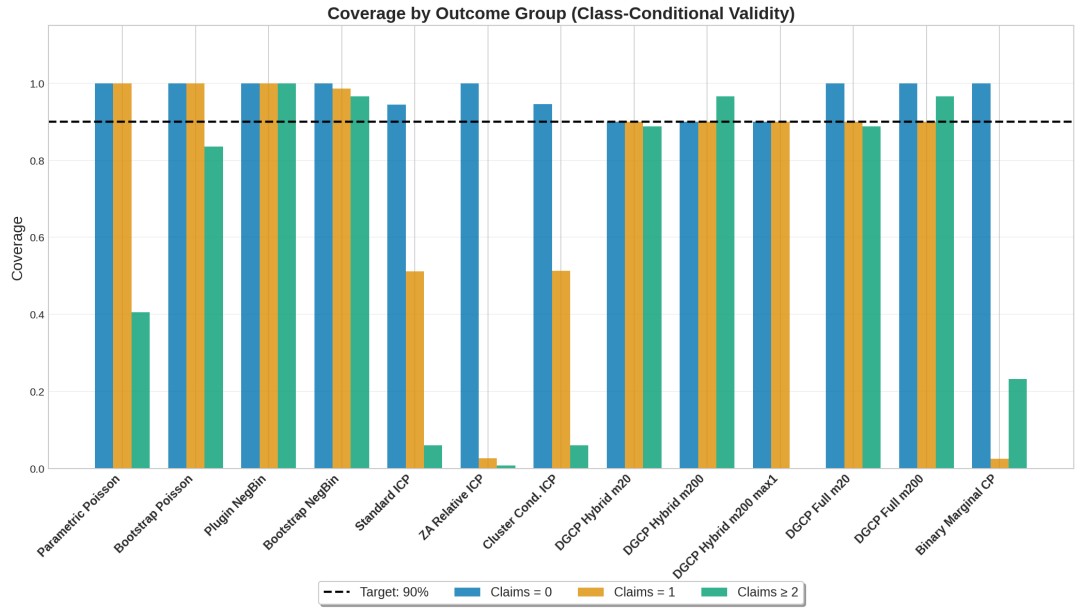

Figure 4.10: Coverage by outcome group. Standard ICP methods exhibit severe under-coverage for claimants despite achieving marginal coverage through over-coverage of non-claimants. DGCP Hybrid methods maintain balanced coverage across all outcome groups.

because the conformity score threshold is calibrated on data dominated by zero claims, producing intervals like $[0, 0]$ or $[0, 1]$ that satisfy the marginal coverage guarantee by correctly covering the majority class while systematically failing to cover actual claimants. The parametric benchmarks avoid this imbalance but only by producing intervals wide enough to trivially contain all plausible

outcomes—achieving near-perfect coverage for both groups through uninformative predictions like $[0, 2]$ or $[0, 3]$ for all observations regardless of risk profile.

The DGCP Hybrid methods resolve this tension by conditioning on discretized outcome groups. The $m = 20$ variant achieves correct coverage for all outcome groups simultaneously, while $m = 200$ slightly over-covers the rare $Y \geq 2$ group due to limited calibration data. The $m = 200$ variant with maximum count restricted to 1 achieves correct coverage for the two dominant groups by design, accepting that the rare multiple-claim cases will be grouped with single claims.

Beyond aggregate coverage statistics, the practical utility of prediction intervals depends on their composition – specifically, whether they provide actionable differentiation between risk profiles. Figure 4.11 shows the distribution of interval types $[l, u]$ produced by each method, revealing fundamental differences in how methods translate uncertainty into predictions.

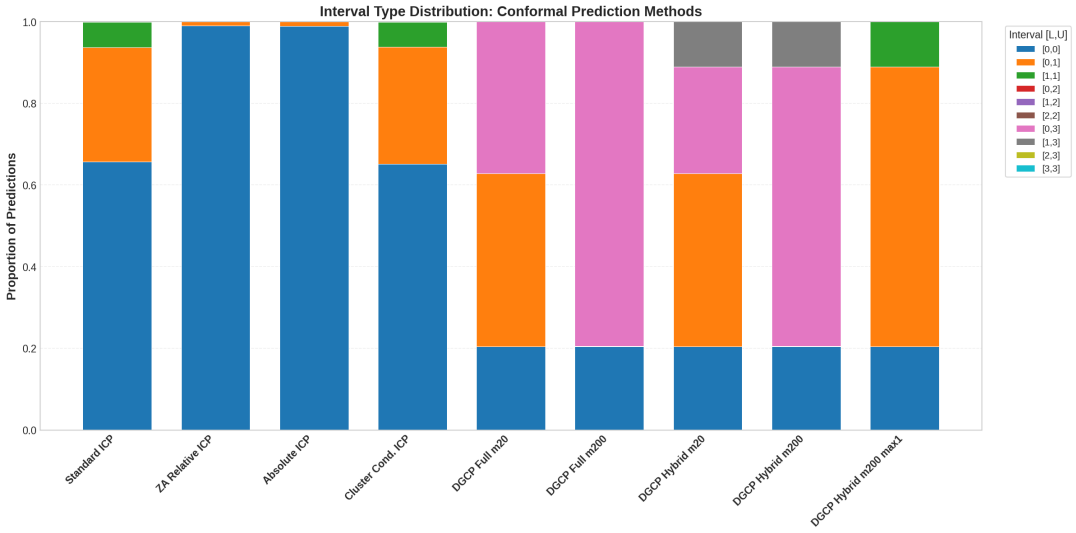

Figure 4.11: Distribution of prediction interval types by method. Standard ICP methods produce predominantly $[0, 0]$ and $[0, 1]$ intervals. DGCP Hybrid methods generate a richer set including $[1, 1]$, and $[1, 3]$ intervals that provide actionable differentiation.

Standard ICP methods produce almost exclusively $[0, 0]$ or $[0, 1]$ intervals, effectively reducing the count prediction problem to binary classification without the calibration benefits of explicit binary modeling. This explains both their efficiency (narrow intervals) and their coverage failures (the $[0, 0]$ intervals assigned to predicted low-risk policies fail whenever claims actually occur). The DGCP Hybrid methods produce a substantially richer distribution: the $m = 200$ variant with max$= 1$ includes a notable proportion of $[1, 1]$ singleton intervals, providing definitive claim predictions for high-risk policies, while the $m = 20$ and standard $m = 200$ variants include $[1, 2]$ and $[1, 3]$ intervals that explicitly exclude zero, signaling policies where claims are highly likely. This richer structure enables genuine risk differentiation rather than the implicit "predict zero for everyone" strategy of standard ICP.

4.6.3 Portfolio Segmentation via Clustering

Having established prediction intervals for individual policyholders, we now turn to understanding where in the covariate space uncertainty remains elevated. Rather than examining individual predictions in isolation, we segment the portfolio into homogeneous groups and evaluate prediction interval properties at the cluster level. This approach reveals systematic patterns of coverage violations and interval widths that may not be apparent from marginal analyses.

Motivation for Cluster-Based Evaluation Conformal prediction guarantees marginal coverage across the entire test population, but as demonstrated in Section 4.3.2, this guarantee does not extend to arbitrary subgroups. By partitioning policyholders into clusters based on their risk characteristics, the cluster-based evaluation serves three complementary purposes:

1. **Portfolio monitoring:** identify covariate regions where prediction intervals systematically undercover or overcover, flagging segments that may require revised pricing or enhanced reserving margins;
2. **Segment-wise coverage audit:** assess whether interval widths appropriately reflect local uncertainty, complementing the outcome-conditional analysis of Section 4.3.2 with a covariate-space perspective;
3. **Conditional validity stress test:** since distribution-free covariate-conditional coverage is unattainable in general (Vovk et al. 2022, Foygel Barber et al. 2020), uniform coverage across diverse clusters constitutes practical evidence that the intervals provide reliable uncertainty quantification beyond the marginal guarantee.

This is not anomaly detection per se; it is a diagnostic for conditional validity and model risk across portfolio segments. This philosophy aligns with classical actuarial applications of clustering for risk segmentation, as exemplified by the sports car classification problem of Ingenbleek and Lemaire (1988), where unsupervised learning methods were used to distinguish vehicle types based on technical characteristics – a problem revisited with modern techniques by Rentzmann and Wüthrich (2019).

Burt Distance-Based Clustering We now detail the Burt distance methodology following Jamotton et al. (2024), which extends classical correspondence analysis techniques (Burt 1950, Greenacre 1984) to the clustering context. A detailed summary of clustering approaches can be found in the Appendix 4.8.5.

Consider q_{cat} categorical rating factors with a total of $m = \sum_{b=1}^{q_{\text{cat}}} m_b$ modalities across all variables, where m_b denotes the number of categories for variable b. Each observation X_i can be represented as a row in the $n \times m$ super-indicator (disjunctive) matrix $\boldsymbol{D} = (d_{i,j})_{i=1,\dots,n;\, j=1,\dots,m}$, where $d_{i,j} = 1$ if observation i exhibits modality j, and $d_{i,j} = 0$ otherwise.

The symmetric $m \times m$ Burt matrix is defined as:

$$\boldsymbol{B} = \boldsymbol{D}^\top \boldsymbol{D}, \tag{4.23}$$

where entry $B_{r,c} = n_{r,c}$ counts the number of observations exhibiting both modalities r and c simultaneously. The Burt matrix captures co-occurrence patterns between categories: diagonal blocks $\boldsymbol{B}_{b,b}$ are diagonal matrices counting category frequencies within variable b, while off-diagonal blocks $\boldsymbol{B}_{b,b'}$ for $b \neq b'$ capture cross-tabulations between variables. By construction, $\boldsymbol{B}$ is composed of $q_{\text{cat}} \times q_{\text{cat}}$ blocks, and the sum of elements within any block $\boldsymbol{B}_{b,b'}$ equals n.

The χ^2 distance between modalities r and r' evaluates dissimilarity by comparing their joint frequency profiles across all other modalities (Burt 1950):

$$\chi^2(r, r') = \sum_{c=1}^{m} \frac{n}{n_{\cdot,c}} \left(\frac{n_{r,c}}{n_{r,\cdot}} - \frac{n_{r',c}}{n_{r',\cdot}} \right)^2, \tag{4.24}$$

where $n_{r,\cdot} = \sum_c n_{r,c} = q_{\text{cat}} \cdot n_{r,r}$ and $n_{\cdot,c} = \sum_r n_{r,c} = q_{\text{cat}} \cdot n_{c,c}$ denote marginal counts. Unlike Hamming distance which simply counts mismatches, this distance accounts for potential dependencies between categorical classes by weighting differences according to inverse marginal frequencies.

To enable use with Euclidean-based clustering algorithms, we construct the weighted Burt matrix:

$$\boldsymbol{B}^W = \frac{1}{q_{\text{cat}}} \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}, \quad \text{where } \boldsymbol{C} = \text{diag}(n_{1,1}^{-1/2}, \ldots, n_{m,m}^{-1/2}). \tag{4.25}$$

With this weighting, the χ^2 distance between rows r and r' of $\boldsymbol{B}^W$ simplifies to the squared Euclidean distance: $\chi^2(r, r') = \sum_{c=1}^{m} (B_{r,c}^W - B_{r',c}^W)^2$.

Each observation is then projected into Burt space by computing the center of gravity of its modalities:

$$X_i^{\text{Burt}} = \frac{1}{q_{\text{cat}}} \boldsymbol{D}_{i,\cdot} \boldsymbol{B}^W \in \mathbb{R}^m. \tag{4.26}$$

Geometrically, since each modality j is represented by row j of $\boldsymbol{B}^W$ (a vector in $\mathbb{R}^m$), an observation exhibiting modalities indicated by $\boldsymbol{D}_{i,\cdot}$ is mapped to the centroid of the corresponding rows in $\boldsymbol{B}^W$.

For numerical covariates, we apply univariate discretization via K-means to convert continuous variables into categorical form, then include them in the Burt matrix construction. As noted by Jamotton et al. (2024), this unsupervised discretization does not account for interactions between rating factors, but the subsequent projection into Burt space helps address this concern by capturing dependencies through joint frequency analysis. Alternatively, standardized numerical variables can be concatenated with the Burt-transformed categorical features before applying K-means clustering.

The Burt distance between observations i and i' is then:

$$L_{\text{Burt}}(i, i') = \left\| X_i^{\text{Burt}} - X_{i'}^{\text{Burt}} \right\|_2. \tag{4.27}$$

Clustering Algorithm. Given the Burt-space representations $\{X_1^{\text{Burt}}, \ldots, X_n^{\text{Burt}}\} \subset \mathbb{R}^m$, we apply the standard K-means algorithm with K-means++ initialization (Arthur and Vassilvitskii 2007). The algorithm iteratively:

1. Assigns each observation to the cluster with nearest centroid:

$$C_k^{(t)} = \{i : k = \arg\min_l \|X_i^{\text{Burt}} - \boldsymbol{\mu}_l^{(t-1)}\|_2^2\};$$

2. Updates centroids: $\boldsymbol{\mu}_k^{(t)} = |C_k^{(t)}|^{-1} \sum_{i \in C_k^{(t)}} X_i^{\text{Burt}}$.

Convergence is guaranteed since each step reduces the within-cluster sum of squares, and the algorithm terminates when cluster assignments stabilize.

Variables Used for Clustering For our German motor insurance portfolio (comprehensive coverage), we construct clusters using a subset of available rating factors selected based on their importance in the fitted Poisson GLM (Table 4.9 in Appendix 4.8.4). The clustering variables include driver age group (`DrivAge`), vehicle age (`VehAge`), annual mileage class (`Mileage`), federal state (`FedState`), rural/urban indicator (`Rural`), vehicle brand (`Brand`), bonus-malus level (`SF_Class`), and vehicle type class for comprehensive coverage (`TypeClass_VK`). These variables capture key dimensions of policyholder heterogeneity relevant to claims frequency: driver experience and age, vehicle characteristics, geographic location, and claims history. The resulting clusters provide interpretable segments for evaluating prediction interval performance across the covariate space.

Cluster Selection Criteria Selecting the number of clusters K requires balancing within-cluster homogeneity against model parsimony. We employ two complementary criteria, visualized in Figure 4.12. The total within-cluster inertia (sum of squared distances to cluster centroids) decreases monotonically with K:

$$I_{\text{within}}(K) = \sum_{k=1}^{K} \sum_{i \in C_k} \left\| X_i^{\text{Burt}} - \boldsymbol{\mu}_k^{\text{Burt}} \right\|_2^2, \tag{4.28}$$

where $\boldsymbol{\mu}_k^{\text{Burt}} = |C_k|^{-1} \sum_{i \in C_k} X_i^{\text{Burt}}$ denotes the centroid of cluster k in Burt space. The "elbow" in the inertia curve—where marginal reduction $[I_{\text{within}}(K-1) - I_{\text{within}}(K)]/I_{\text{within}}(K-1)$ stabilizes—indicates diminishing returns from additional clusters. Since our goal is claims frequency modeling,

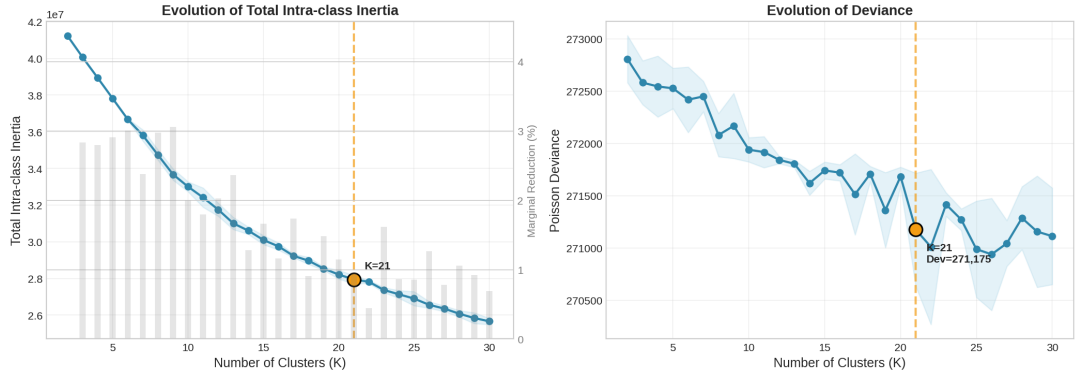

Figure 4.12: Cluster selection metrics for Burt distance-based K-means. **Left:** Total within-cluster inertia $I_{\text{within}}(K)$ (solid line) and marginal reduction percentage (bars). **Right:** Poisson deviance $D(K)$ of cluster-only model. Both metrics are averaged over 15 random seeds to account for initialization variability. The elbow criterion and deviance reduction suggest $K = 21$ clusters provide a good balance between homogeneity and parsimony.

we evaluate cluster quality using the Poisson deviance when fitting cluster-specific claim rates. For a partition into K clusters, we fit a Poisson GLM with cluster indicators as the sole predictors:

$$Y_i \mid X_i \sim \text{Poisson}(\hat{\lambda}_k \cdot v_i), \quad \text{where } i \in C_k, \tag{4.29}$$

with exposure v_i and cluster-specific rate $\hat{\lambda}_k = \sum_{i \in C_k} Y_i / \sum_{i \in C_k} v_i$. The resulting Poisson deviance:

$$D(K) = 2 \sum_{i=1}^{n} \left[Y_i \log \left(\frac{Y_i}{\hat{\lambda}_{k(i)} v_i} \right) - (Y_i - \hat{\lambda}_{k(i)} v_i) \right], \tag{4.30}$$

where $k(i)$ denotes the cluster containing observation i, measures how well the cluster partition captures heterogeneity in claim frequencies. This criterion directly connects to our prediction interval evaluation: clusters that explain substantial claim frequency variation should exhibit more homogeneous prediction interval behavior.

Based on the analysis shown in Figure 4.12, we select $K = 21$ clusters. The left panel shows that marginal inertia reduction stabilizes around this value, while the right panel demonstrates that the cluster-only Poisson model achieves deviance approaching that of a full GLM using all rating factors.

Cluster-Level Evaluation of Prediction Intervals Cluster-level coverage and interval width for our proposed DGCP Hybrid m200 are presented in Figure 4.13[9]. The left column

[9] A comparison with four alternatives is provided in Appendix Figure 4.29.

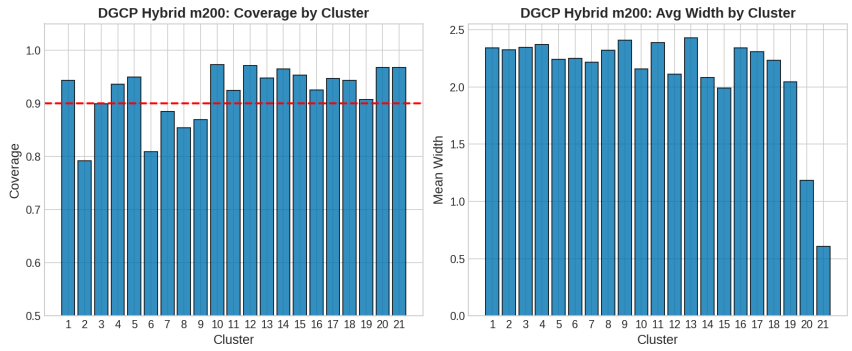

Figure 4.13: Coverage (left) and mean interval width (right) by cluster for DGCP Hybrid m200 with underlying Poisson GLM. The horizontal red line indicates the target coverage level $1 - \alpha = 0.90$.

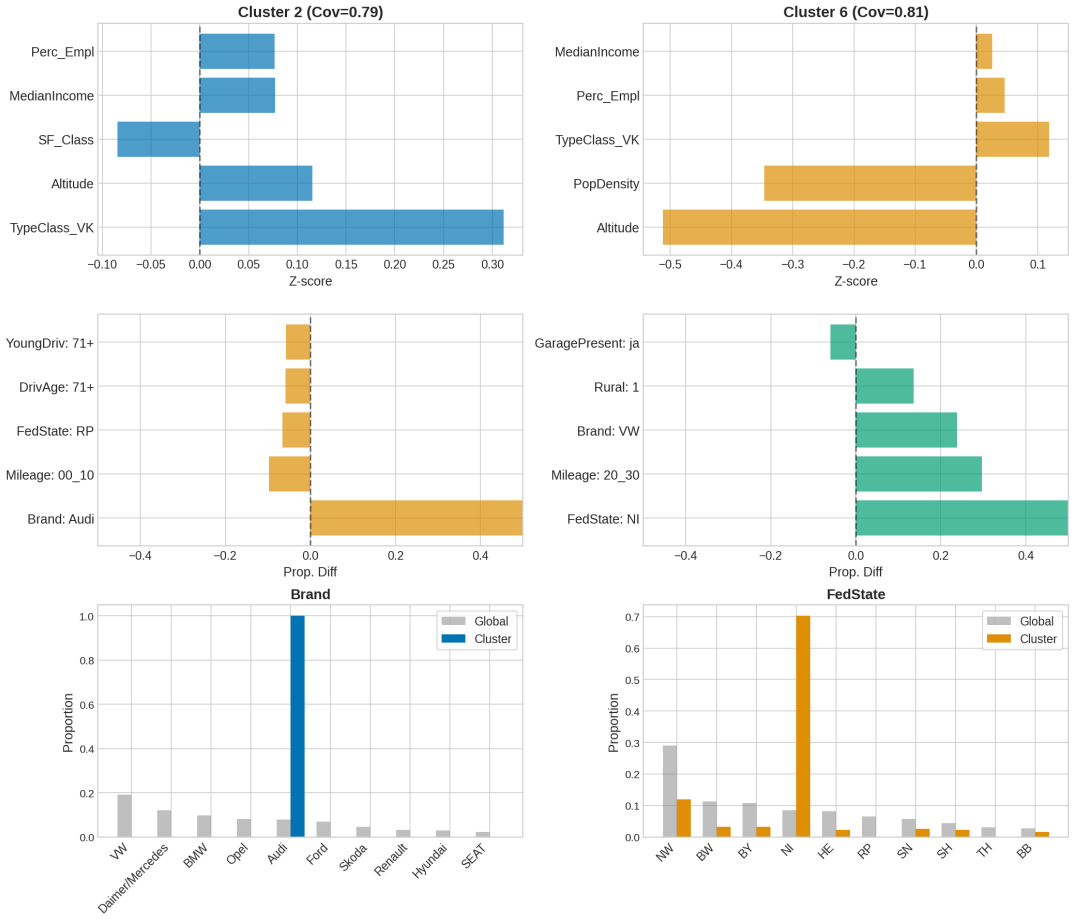

Figure 4.14: Characteristics of the two clusters with lowest coverage under DGCP Hybrid ($m = 200$). **First row:** Standardized deviation $((\bar{x}_{\text{cluster}} - \bar{x}_{\text{global}})/\sigma_{\text{global}})$ from population mean for top continuous variables. **Second row:** Largest categorical proportion differences (cluster vs. population). **Third row:** Full distribution of the most distinctive categorical variable comparing cluster (colored) to global (gray) proportions. Each column corresponds to one of the two worst-performing clusters, with coverage rate indicated in the subplot title.

shows empirical coverage by cluster, with the horizontal red line indicating the target coverage level $1 - \alpha = 0.90$. The right column displays mean interval width by cluster. To understand which policyholder segments exhibit the largest prediction interval violations, we examine the characteristics of clusters with the worst coverage performance under the DGCP Hybrid method.

Figure 4.14 presents this analysis for the two clusters with lowest empirical coverage. The cluster-based diagnostic reveals systematic patterns in prediction interval performance across portfolio segments. Table 4.15 summarizes the 21 identified clusters, with cluster sizes ranging from 0.8% to 27.8% of the portfolio. Clusters exhibit substantial heterogeneity in claims frequency (5.6% to 15.7%) and defining characteristics, with most clusters dominated by specific vehicle brands (e.g., Cluster 2: 100% Audi, Cluster 6: 43% VW) or geographic regions (e.g., Cluster 3: 100% Rhineland-Palatinate, Cluster 9: 100% Thuringia).

Note that the surprisingly small average widths for clusters 20 and 21 arise from the fact that these clusters have a lower share of policyholders with claims, namely 5.93% and 4.46%, and consist of older cars with low mileage; see Table 4.14 and Table 4.15. A closer inspection in Figure 4.32 shows that for these clusters, we have poor coverage for claimants and over-coverage (equal to 1) for non-claimants. In contrast, for the two clusters with the lowest overall coverage, clusters 2 and 6, we observe low coverage values of 77% to 79% for non-claimants in the DGCP Hybrid model.

4.7 Conclusion

This work provides the first comprehensive study of conformal prediction methods for modeling the frequency of insurance claims for count data. We developed a novel two-stage DGCP Hybrid framework that addresses the fundamental challenge of outcome-conditional coverage in zero-inflated count data. By combining binary Mondrian calibration for the claim/no-claim decision with hypothetical-score DGCP for positive counts, our approach achieves balanced coverage across outcome groups while maintaining interval efficiency.

Our simulation study across four data-generating processes demonstrates robust performance under challenging conditions, including overdispersion, zero-inflation, and model misspecification. The empirical application to German motor insurance data confirms these findings in practice: DGCP Hybrid produces informative prediction intervals that distinguish high-risk from low-risk policies, rather than the uninformative $[0, 1]$ intervals that dominate standard ICP output. The Burt distance-based clustering diagnostic further enables the identification of portfolio segments where prediction uncertainty is systematically elevated, providing actionable guidance for model refinement.

A few practical considerations are worth highlighting. The minimum group size m controls the granularity of outcome-conditional calibration, with $m = 200$ providing stable coverage approaching binary conditioning, while $m = 20$ offers finer-grained calibration when sufficient data exists. The choice depends on the specific portfolio structure, and a more detailed analysis of its impact is left for future work. A key advantage of conformal methods is their compatibility with established actuarial pipelines: production GLMs can be wrapped with CP to provide individual-level uncertainty quantification without modifying fitted parameters, informing reserving, pricing adjustments, and underwriting decisions through honest communication of prediction reliability.

This work focused on claims frequency, where the discrete zero-inflated structure poses particular challenges for uncertainty quantification. Claims severity, while seemingly simpler due to its continuous nature, presents its own difficulties. Severity often depends on factors external to the policyholder (e.g., the other vehicle involved), limiting the explanatory power of rating factors. Nonetheless, extending the DGCP framework to composite frequency-severity models for direct premium prediction represents a promising direction for future research. Additionally, the cluster-based diagnostic approach demonstrated here could be developed into a systematic tool for identifying subpopulations requiring targeted model improvements or enhanced underwriting scrutiny.

4.8 Appendix

4.8.1 Proof of Theorem 6

Proof. We establish marginal validity by showing that, for each possible outcome $y^* \in \mathbb{N}_0$, the true outcome is included in the hybrid prediction set with probability at least $1 - \alpha$. Marginal coverage then follows by the law of total probability.

Throughout, we condition on the fitted model $\hat{f}$, which is trained on $\mathcal{D}_{\text{train}}$ disjoint from both calibration and test data, and is therefore fixed with respect to the calibration and test points. The remaining source of randomness is the exchangeable sequence $\{(X_i, Y_i)\}_{i=1}^n \cup \{(X_{\text{test}}, Y_{\text{test}})\}$.

Stage 1: Validity for $Y_{\text{test}} = 0$. Stage 1 applies Mondrian conformal prediction calibrated on $\mathcal{C}_0 = \{i : Y_i = 0\}$. By Proposition 4.6 of Vovk et al. (2022), under exchangeability of $(X_{\text{test}}, 0)$ with the calibration points in $\mathcal{C}_0$:

$$\mathbb{P}\big(0 \in C^{(0)}_{\text{binary}}(X_{\text{test}}) \mid Y_{\text{test}} = 0\big) \geq 1 - \alpha. \tag{4.31}$$

Since $C^{(0)}_{\text{binary}} \subseteq C_{\text{hybrid}}$, this implies $\mathbb{P}(Y_{\text{test}} \in C_{\text{hybrid}} \mid Y_{\text{test}} = 0) \geq 1 - \alpha$.

Stage 2: Validity for $Y_{\text{test}} = y^*$ with $y^* \geq 1$. Fix any $y^* \geq 1$. We must show $\mathbb{P}(y^* \in C^{(+)}_{\text{count}}(X_{\text{test}}) \mid Y_{\text{test}} = y^*) \geq 1 - \alpha$.

Stage 2 constructs the calibration neighborhood $G_{y^*} \subseteq \{1, \ldots, y_{\max}\}$ by symmetric neighbor expansion (Algorithm 3), ensuring $y^* \in G_{y^*}$ and $\sum_{k \in G_{y^*}} |\mathcal{C}_k| \geq m$. Let $\mathcal{I}_{y^*} = \{i \in \{1, \ldots, n\} : Y_i \in G_{y^*},\ Y_i > 0\}$ denote the set of calibration indices whose outcomes fall in this neighborhood. We establish that the conformal p-value $p(y^*)$ is super-uniform through the following argument.

Step (i): G_{y^*} and $\mathcal{I}_{y^*}$ are symmetric functions of the data. The neighborhood G_{y^*} is determined by the multiset of calibration outcomes $\{Y_1, \ldots, Y_n\}$ and the fixed parameter m, through the deterministic group-construction rule. The index set $\mathcal{I}_{y^*} = \{i : Y_i \in G_{y^*}\}$ depends on the *values* of the Y_i's but not on their ordering or on the covariates X_i. Both are therefore symmetric functions of the exchangeable sequence.

Step (ii): Conditional exchangeability of scores. Condition on $Y_{\text{test}} = y^*$ and on $\mathcal{I}_{y^*}$ taking a particular value (i.e., a specific set of calibration indices falls in G_{y^*}). We distinguish two cases.

Case 1: Singleton group ($G_{y^*} = \{y^*\}$). All calibration points in $\mathcal{I}_{y^*}$ share the outcome y^* with the test point. Under the original exchangeability of the full sequence, conditioning on the outcome vector preserves exchangeability of covariates among indices with the same Y-value (see Aldous 1985, Section 1). Since $\hat{f}$ is fixed, the hypothetical scores $\{s(\hat{f}(X_i), y^*)\}_{i \in \mathcal{I}_{y^*}} \cup \{s(\hat{f}(X_{\text{test}}), y^*)\}$ are exchangeable, being identical measurable transformations of exchangeable random variables. Assumption 17 is therefore satisfied exactly.

Case 2: Non-singleton group ($|G_{y^*}| > 1$). The calibration set $\mathcal{I}_{y^*}$ may contain points with $Y_i \neq y^*$. Exchangeability of covariates across indices with different Y-values is not guaranteed by exchangeability of the original sequence alone. In this case, Assumption 17 provides the required exchangeability of the scores directly.

Step (iii): Super-uniformity of the p-value. Under Assumption 17, the scores $\{s(\hat{f}(X_i), y^*)\}_{i \in \mathcal{I}_{y^*}} \cup \{s(\hat{f}(X_{\text{test}}), y^*)\}$ are exchangeable conditional on $Y_{\text{test}} = y^*$ and $\mathcal{I}_{y^*}$. By the fundamental property of conformal p-values computed from exchangeable scores (cf. Lemma 1 of Tibshirani et al. 2019), the conformal p-value

$$p(y^*) = \frac{|\{i \in \mathcal{I}_{y^*} : s(\hat{\mu}_i, y^*) \geq s(\hat{\mu}_{\text{test}}, y^*)\}| + 1}{|\mathcal{I}_{y^*}| + 1} \tag{4.32}$$

satisfies:

$$\mathbb{P}\big(p(y^*) \leq \alpha \mid Y_{\text{test}} = y^*, \mathcal{I}_{y^*}\big) \leq \alpha. \tag{4.33}$$

Since this bound holds for every realization of $\mathcal{I}_{y^*}$, integrating over $\mathcal{I}_{y^*}$ yields:

$$\mathbb{P}\big(p(y^*) \leq \alpha \mid Y_{\text{test}} = y^*\big) \leq \alpha. \tag{4.34}$$

Therefore $\mathbb{P}(y^* \in C^{(+)}_{\text{count}} \mid Y_{\text{test}} = y^*) \geq 1 - \alpha$, and since $C^{(+)}_{\text{count}} \subseteq C_{\text{hybrid}}$:

$$\mathbb{P}\big(Y_{\text{test}} \in C_{\text{hybrid}} \mid Y_{\text{test}} = y^*\big) \geq 1 - \alpha, \quad \forall\, y^* \geq 1. \tag{4.35}$$

Marginal Coverage. Combining Stages 1 and 2, we have $\mathbb{P}(Y_{\text{test}} \in C_{\text{hybrid}} \mid Y_{\text{test}} = y) \geq 1 - \alpha$ for all $y \in \mathbb{N}_0$. By the law of total probability over the natural partition $\{0\}, \{1\}, \{2\}, \ldots$ of $\mathbb{N}_0$:

$$\mathbb{P}\big(Y_{\text{test}} \in C_{\text{hybrid}}\big) = \sum_{y=0}^{\infty} \mathbb{P}(Y_{\text{test}} = y) \cdot \mathbb{P}\big(Y_{\text{test}} \in C_{\text{hybrid}} \mid Y_{\text{test}} = y\big) \tag{4.36}$$

$$\geq \sum_{y=0}^{\infty} \mathbb{P}(Y_{\text{test}} = y) \cdot (1 - \alpha) = 1 - \alpha. \tag{4.37}$$

This completes the proof. We emphasize that this argument does *not* require the calibration neighborhoods $\{G_y\}_{y \geq 1}$ to form a partition of $\mathbb{N}_{\geq 1}$, nor does it require that all calibration points in G_{y^*} share the outcome y^*. For singleton groups, validity follows from exchangeability alone; for non-singleton groups, it relies on Assumption 17. □

4.8.2 Description of the Benchmarks

To contextualize the performance of conformal prediction methods, we compare against classical parametric prediction intervals for Poisson and Negative Binomial regression models. These methods serve as important benchmarks under correct model specification. For comprehensive treatments, we refer to Kim et al. (2022) and Meeker et al. (2017), Chapter 7.

Background: IID Poisson Prediction Intervals

We briefly summarize foundational methods for the IID setting, where X events are observed in n units of exposure and the goal is to predict Y events in m future units, assuming a common rate λ. These methods underpin many regression extensions.

Conservative (Exact) Method. Based on the conditional distribution $X \mid (X + Y) \sim \text{Binomial}(X + Y, \pi)$ where $\pi = n/(n + m)$, Weiss (1955) and later Bain and Patel (1993) derive exact prediction intervals by solving:

$$1 - F_{\text{Binom}}(x - 1; x + \underset{\sim}{Y}, \pi) > \frac{\alpha}{2} \quad \text{and} \quad F_{\text{Binom}}(x; x + \tilde{Y}, \pi) > \frac{\alpha}{2} \tag{4.38}$$

This method guarantees at least nominal coverage but can be overly conservative for small $n\lambda$ or $m\lambda$.

Joint-Sample Approximate Method. Krishnamoorthy and Peng (2011) introduced an improved normal approximation with better finite-sample properties:

$$\Gamma^{\alpha}_{\text{JS}} = m\hat{\lambda} + \frac{m z^2_{1-\alpha/2}}{2n} \mp z_{1-\alpha/2} \left[m\hat{\lambda} \left(\frac{1}{n} + \frac{1}{m} \right) + \left(\frac{m z_{1-\alpha/2}}{2n} \right)^2 \right]^{1/2} \tag{4.39}$$

This method maintains coverage probabilities close to nominal when $n\lambda$ and $m\lambda$ both exceed 10.

Jeffreys Prediction Interval. Derived from Bayesian prediction with the Jeffreys prior $\pi(\lambda) \propto \lambda^{-1/2}$, this yields a negative binomial predictive distribution (Bejleri and Nandram 2018):

$$\Gamma^{\alpha}_{\text{Jeff}} = \left[F^{-1}_{\text{NB}}\left(\frac{\alpha}{2}; x + 0.5, \frac{n}{n+m}\right), F^{-1}_{\text{NB}}\left(1 - \frac{\alpha}{2}; x + 0.5, \frac{n}{n+m}\right)\right] \tag{4.40}$$

where $F^{-1}_{\text{NB}}(\cdot; k, p)$ is the negative binomial quantile function.

Methods for Poisson Regression Models

The following methods extend to regression settings where $Y_i \mid X_i \sim \text{Poisson}(\lambda(X_i))$ with $\lambda(X_i) = \exp(X_i\beta)$.

Normal Approximation Method ($\tilde{\Gamma}_1$). Building on standard GLM theory, Myers and Montgomery (1997) present a plug-in prediction interval based on the asymptotic normality of the MLE. Given a test covariate x_0 with predicted mean $\hat{\lambda}_0 = \exp(x_0\hat{\beta})$:

$$\tilde{\Gamma}_1 = \left[\hat{\lambda}_0 \pm z_{1-\alpha/2}\sqrt{\hat{\lambda}_0\hat{V}_0}\right] \tag{4.41}$$

where $\hat{V}_0 = 1 + \hat{\psi}_0^2\hat{\lambda}_0 \cdot \text{tr}[I(\hat{\beta})^{-1}x_0^T x_0]$ incorporates estimation uncertainty, with

$$\hat{\psi}_0 = d/dx \log \rho(x_0\hat{\beta})$$

and $I(\hat{\beta})$ the observed Fisher information. This method provides shorter intervals but may undercover for small sample sizes. This is the first Poisson benchmark used in our experiments (`Parametric_Poisson`).

Residual Bootstrap Method ($\tilde{\Gamma}_2$). Following Davison and Hinkley (1997) (Chapter 7.2), this approach resamples standardized Pearson residuals $r_i = (Y_i - \hat{\mu}_i)/\sqrt{\hat{\phi}\hat{\mu}_i}$, where $\hat{\phi} = 1$ for Poisson. Bootstrap predictions are generated as:

$$Y^*_{0,b} = \hat{\lambda}_0 + r^*_b\sqrt{\hat{\phi}\hat{\lambda}_0}, \quad b = 1, \ldots, B \tag{4.42}$$

The prediction interval uses empirical quantiles:

$$\tilde{\Gamma}_2 = \left[\lfloor Q_{\alpha/2}(\{Y^*_{0,b}\})\rfloor, \lceil Q_{1-\alpha/2}(\{Y^*_{0,b}\})\rceil\right] \tag{4.43}$$

This naturally handles estimation uncertainty and can accommodate overdispersion through $\hat{\phi} > 1$. This is the second Poisson benchmark used in our experiments (`Bootstrap_Poisson`).

Methods for Negative Binomial Regression Models

When overdispersion is present ($\text{Var}[Y] > \text{E}[Y]$), the Negative Binomial (NB) distribution provides a natural extension. Following Kim et al. (2022), we consider the NB2 parametrization where $Y_i \mid x_i \sim \text{NB}(r, p_i)$ with $\text{E}[Y_i] = \lambda_i$ and $\text{Var}[Y_i] = \lambda_i + \alpha\lambda_i^2$, where $\alpha = 1/r$ is the overdispersion parameter.

Plug-in Normal Approximation ($\tilde{\Gamma}_3$). Extending the Poisson normal approximation to NB2, Kim et al. (2020) provide:

$$\tilde{\Gamma}_3 = \left[\hat{\lambda}_0 \pm z_{1-\alpha/2}\sqrt{\hat{\lambda}_0(1+\hat{\lambda}_0)/\hat{\xi} + \hat{\lambda}_0 + n^{-1}\hat{\lambda}_0^2\hat{\psi}^T\hat{\Xi}_{11}\hat{\psi}}\right] \tag{4.44}$$

where $\hat{\Xi}$ is the estimated asymptotic covariance matrix of $(\hat{\beta}, \hat{\xi})$ with $\xi = 1/\alpha$. Note that as $\xi \to \infty$ (no overdispersion), this approaches the Poisson normal approximation $\tilde{\Gamma}_1$. This is our third benchmark (`NB_PlugIn`).

Chebyshev Inequality Method ($\tilde{\Gamma}_4$). Wood (2005) proposed a conservative approach requiring only the first two moments. Ash et al. (2021) extended this to various Poisson mixture models including the Poisson-Gamma mixture (NB). Using the one-sided Chebyshev inequality $\Pr(Y_0 - \mu_0 \geq t\sigma_0) \leq 1/(1+t^2)$:

$$\tilde{\Gamma}_4 = \left[0, \hat{\lambda}_0 + \sqrt{\alpha^{-1} - 1} \cdot \sqrt{\hat{\lambda}_0 + \hat{\sigma}_0^2}\right] \tag{4.45}$$

where for the Gamma mixture (NB) case with shape parameter $r = 1/\alpha$:

$$\hat{\sigma}_0^2 = \hat{\lambda}_0^2 \left\{\widehat{\text{Var}}[\hat{\eta}_0] + (1 + \widehat{\text{Var}}[\hat{\eta}_0])/r\right\} \tag{4.46}$$

This method produces conservative (one-sided) intervals but is versatile across different mixture specifications. This is our fourth benchmark (`NB_Chebyshev`).

Parametric Bootstrap Method ($\tilde{\Gamma}_5$). Extending the bootstrap approach of Davison and Hinkley (1997) to NB regression, and following the nonparametric-parametric hybrid of Olive et al. (2022), bootstrap samples $Y_1^*, \ldots, Y_B^*$ are generated from $\text{NB}(r, r/(r + \hat{\lambda}_0))$. To incorporate parameter estimation uncertainty, we use residual-based resampling where standardized Pearson residuals are computed using the NB variance:

$$r_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + \hat{\alpha}\hat{\mu}_i^2}} \tag{4.47}$$

Bootstrap predictions are then:

$$Y_{0,b}^* = \hat{\lambda}_0 + r_b^*\sqrt{\hat{\lambda}_0 + \hat{\alpha}\hat{\lambda}_0^2} \tag{4.48}$$

The prediction interval uses empirical quantiles. This is our fifth benchmark (`NB_Bootstrap`).

4.8.3 Details on the Simulation Study

Data Generating Process Details

We consider four Data Generating Processes (DGP) that represent progressively more complex departures from the standard Poisson assumption. Each DGP generates claim counts $Y_i \in \mathcal{Y} = \mathbb{N}_0$ conditional on a d-dimensional covariate vector $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and an exposure measure $v_i > 0$. The covariate effects enter through a linear predictor $\eta_i = X_i^\top \beta$, where the coefficient vector $\beta \in \mathbb{R}^p$ is constructed to produce heterogeneous effects across covariates. Specifically, the first $k = \min(5, p)$ coefficients corresponding to key actuarial variables are drawn as $\beta_j \sim \mathcal{N}(0.5, 0.04)$

for $j = 1, \ldots, k$, while the remaining coefficients follow $\beta_j \sim \mathcal{N}(0, 0.01)$ for $j = k+1, \ldots, p$. To ensure numerical stability, the linear predictor is bounded as $\eta_i \in [-10, 3]$, yielding base rates $\mu_i^{(0)} = v_i \cdot \exp(\eta_i)$ in the range approximately $[10^{-6}, 50]$.

Poisson DGP. The baseline Poisson DGP generates counts according to

$$Y_i \mid X_i \sim \text{Poisson}(\mu_i), \quad \text{where} \quad \mu_i = v_i \cdot \exp(X_i^\top \beta). \tag{4.49}$$

Under this specification, the conditional mean and variance satisfy $\mathbb{E}[Y_i \mid X_i] = \text{Var}(Y_i \mid X_i) = \mu_i$, the classical equidispersion property. However, the heterogeneity induced by the covariate structure and optional interaction effects produces marginal overdispersion when averaging over the covariate distribution, with the marginal variance exceeding the marginal mean.

Zero-Inflated Poisson (ZIP) DGP. The ZIP DGP introduces structural zeros through a two-component mixture:

$$Y_i \mid X_i \sim \begin{cases} 0 & \text{with probability } \pi_i, \\ \text{Poisson}(\mu_i^{(0)}) & \text{with probability } 1 - \pi_i, \end{cases} \tag{4.50}$$

where the zero-inflation probability depends on covariates through the logistic specification

$$\pi_i = \frac{\pi_0 \cdot \text{logit}^{-1}(X_i^\top \gamma)}{\mathbb{E}[\text{logit}^{-1}(X^\top \gamma)]}, \tag{4.51}$$

with $\gamma_j \sim \mathcal{N}(0.5, 0.01)$ for $j = 1, \ldots, \min(3, p)$ and $\pi_0 \in [0, 1]$ controlling the overall zero-inflation level. The normalizing denominator ensures that $\mathbb{E}[\pi_i] \approx \pi_0$. The conditional mean under the ZIP model is $\mathbb{E}[Y_i \mid X_i] = (1 - \pi_i)\mu_i^{(0)}$, and the conditional variance is

$$\text{Var}(Y_i \mid X_i) = (1 - \pi_i)\mu_i^{(0)} \left(1 + \pi_i \mu_i^{(0)}\right), \tag{4.52}$$

which exceeds the mean whenever $\pi_i > 0$, inducing overdispersion.

Negative Binomial DGP. The Negative Binomial DGP accommodates overdispersion through a gamma-mixed Poisson formulation. Letting $\lambda_i \mid X_i \sim \text{Gamma}(\theta, \theta/\mu_i^{(0)})$ with shape $\theta > 0$ and rate $\theta/\mu_i^{(0)}$, we have $Y_i \mid \lambda_i \sim \text{Poisson}(\lambda_i)$. Marginalizing over λ_i yields the negative binomial distribution:

$$Y_i \mid X_i \sim \text{NegBin}\left(\theta, \frac{\theta}{\theta + \mu_i^{(0)}}\right), \tag{4.53}$$

with probability mass function

$$\mathbb{P}(Y_i = y \mid X_i) = \binom{y + \theta - 1}{y} \left(\frac{\theta}{\theta + \mu_i^{(0)}}\right)^\theta \left(\frac{\mu_i^{(0)}}{\theta + \mu_i^{(0)}}\right)^y, \quad y \in \mathbb{N}_0. \tag{4.54}$$

The conditional moments are $\mathbb{E}[Y_i \mid X_i] = \mu_i^{(0)}$ and

$$\text{Var}(Y_i \mid X_i) = \mu_i^{(0)} + \frac{(\mu_i^{(0)})^2}{\theta}, \tag{4.55}$$

so the dispersion ratio $\text{Var}(Y_i \mid X_i)/\mathbb{E}[Y_i \mid X_i] = 1 + \mu_i^{(0)}/\theta$ exceeds unity and increases with the mean.

Hurdle DGP. The Hurdle DGP separates the zero-generating process from the positive count process:

$$Y_i \mid X_i \sim \begin{cases} 0 & \text{with probability } \pi_i, \\ \text{ZT-Poisson}(\mu_i^{(0)}) & \text{with probability } 1 - \pi_i, \end{cases} \tag{4.56}$$

where ZT-Poisson(μ) denotes the zero-truncated Poisson distribution with probability mass function

$$\mathbb{P}(Y = y \mid Y > 0) = \frac{\mu^y e^{-\mu}}{y!(1 - e^{-\mu})}, \quad y \in \mathbb{N}. \tag{4.57}$$

The hurdle probability follows a logistic specification analogous to the ZIP model, with $\gamma_j \sim \mathcal{N}(-0.5, 0.04)$ for $j = 1, \ldots, \min(2, p)$. The conditional mean is

$$\mathbb{E}[Y_i \mid X_i] = (1 - \pi_i) \cdot \frac{\mu_i^{(0)}}{1 - e^{-\mu_i^{(0)}}}, \tag{4.58}$$

reflecting the truncation adjustment for positive counts.

Interaction and Spatial Effects. When interaction effects are enabled, the linear predictor is augmented with a product term between the first two continuous covariates (policyholder age and vehicle age):

$$\eta_i = X_i^\top \beta + \beta_{\text{int}} \cdot X_{i,1} \cdot X_{i,2}, \tag{4.59}$$

where $\beta_{\text{int}} \sim \mathcal{N}(0.2, 0.0025)$. When spatial effects are enabled, region-specific random effects $\zeta_r \sim \mathcal{N}(0, 0.09)$ are added, with additional Poisson-distributed counts $Y_i^{(s)} \sim \text{Poisson}(\max(0, e^{\zeta_{r(i)}} - 1))$ contributing to the total.

Covariate Structure and Configuration

The covariate vector $X_i \in \mathbb{R}^d$ comprises both continuous and categorical variables designed to mimic realistic insurance rating factors. Let $d_c = \lfloor d \cdot (1 - \rho_{\text{cat}}) \rfloor$ denote the number of continuous covariates and $d - d_c$ the number of categorical covariates, where $\rho_{\text{cat}} \in [0, 1]$ is the categorical ratio.

The first three continuous covariates follow actuarially motivated distributions:

$$X_{i,1} \sim \text{TruncNormal}(45, 15^2; 18, 90) \quad \text{(policyholder age)}, \tag{4.60}$$
$$X_{i,2} \sim \text{Exponential}(1/8) \quad \text{(vehicle age)}, \tag{4.61}$$
$$X_{i,3} \sim \text{Uniform}(0, 20) \quad \text{(policy duration)}. \tag{4.62}$$

The remaining $d_c - 3$ continuous covariates are drawn from a multivariate normal distribution with Toeplitz correlation structure:

$$(X_{i,4}, \ldots, X_{i,d_c})^\top \sim \mathcal{N}(\mu_X, \Sigma_X), \quad \text{where} \quad (\Sigma_X)_{jk} = \rho^{|j-k|} \tag{4.63}$$

for correlation strength parameter $\rho \in [0, 1)$ and mean vector μ_X with components drawn uniformly from $[-1, 1]$.

Categorical covariates include geographic region with five levels drawn according to probabilities $(0.30, 0.25, 0.20, 0.15, 0.10)$, coverage type with three levels drawn according to $(0.50, 0.30, 0.20)$,

and additional factors with randomly determined numbers of levels between 2 and 4.

Exposure v_i follows a realistic mixture distribution:

$$v_i = \begin{cases} 1 & \text{with probability } 0.8, \\ \text{Beta}(2, 5) & \text{with probability } 0.2, \end{cases} \tag{4.64}$$

with the result clipped to $[0.1, 1.0]$, reflecting policy inception and termination patterns observed in practice.

Baseline Configuration and Sensitivity Analysis

Table 4.5: Simulation configuration: baseline values and sensitivity analysis ranges.

Parameter	Symbol	Baseline	Variation Range
Sample size	N	10,000	$\{1000, 2000, 4000, 6000, 10000\}$
Covariate dimension	d	10	$\{10, 20, 50\}$
Correlation strength	ρ	0.3	$\{0.0, 0.3, 0.5\}$
Categorical ratio	ρ_{cat}	0.3	$\{0.1, 0.3, 0.5\}$
Zero-inflation (ZIP/Hurdle)	π_0	0.3	$\{0.0, 0.1, 0.3, 0.5\}$
Overdispersion (NegBin)	θ	3.0	$\{2.0, 3.0, 4.0, 5.0\}$
Interaction effects	—	True	$\{\text{True}, \text{False}\}$
Spatial effects	—	False	$\{\text{True}, \text{False}\}$

The baseline configuration specifies $N = 10{,}000$ observations with $d = 10$ covariates, of which 30% are categorical ($\rho_{\text{cat}} = 0.3$). The correlation strength is set to $\rho = 0.3$, and interaction effects between policyholder age and vehicle age are included. Spatial effects are excluded in the baseline. For DGP-specific parameters, the ZIP and Hurdle models use a zero-inflation level of $\pi_0 = 0.3$, while the Negative Binomial model uses an overdispersion parameter of $\theta = 3.0$.

Table 4.5 summarizes the parameter variations examined in the sensitivity analysis. Each parameter is varied systematically while holding others at baseline values, enabling isolation of individual effects on method performance.

Computational Efficiency

All methods described are computationally efficient once the underlying prediction model is fitted:

- **Standard ICP**: $O(n \log n)$ for quantile computation, $O(n_{\text{test}} \cdot y_{\max})$ for prediction set construction.
- **DGCP variants**: Additional $O(y_{\max})$ factor for group construction, amortized across test points. The group structure can be precomputed once from the calibration set.

For typical insurance applications with $n \approx 10^4$–10^6 calibration points, $n_{\text{test}} \approx 10^4$–$10^5$ test points, and $y_{\max} \leq 20$, all conformal methods complete in seconds on standard hardware.

The simulation study was conducted on a high-performance computing cluster using AMD EPYC 9654 96-Core processors with 375 GiB available memory per node. The simulations were partitioned into two components based on method complexity:

- **Baseline and standard CP methods** (Poisson Deviance ICP, ZA Relative ICP, Cluster Conditional CP, Binary Marginal CP, and all parametric/bootstrap benchmarks): Total runtime of approximately 9.2 node-hours, with 95.9% CPU utilization and peak memory usage of 54.3 GiB. The longest run times were observed for the bootstrap benchmarks.
- **DGCP variants**: Runtime varied with the minimum group size parameter m. DGCP Hybrid with $m \in \{10, 20, 50\}$ required 4.4 hours; DGCP Hybrid with $m \in \{100, 200\}$ required 3.1 hours. DGCP Full variants exhibited similar runtimes of 4.5 hours across all m values. All DGCP runs achieved near-complete CPU utilization (100% of 96 cores) with modest memory requirements (25–29 GiB, representing 7–8% of available node memory).

The reduced runtime for larger values of m in the Hybrid variant reflects the computational savings from fewer distinct calibration groups, as more outcome values are pooled together. The Full variant shows less sensitivity to m since it applies uniform grouping across all count values including zero, resulting in more consistent group structures regardless of the minimum size threshold.

Sensitivity Analysis and Additional Results

This appendix presents additional simulation results including sensitivity analyses across key parameters. All results are based on the baseline configuration unless otherwise noted.

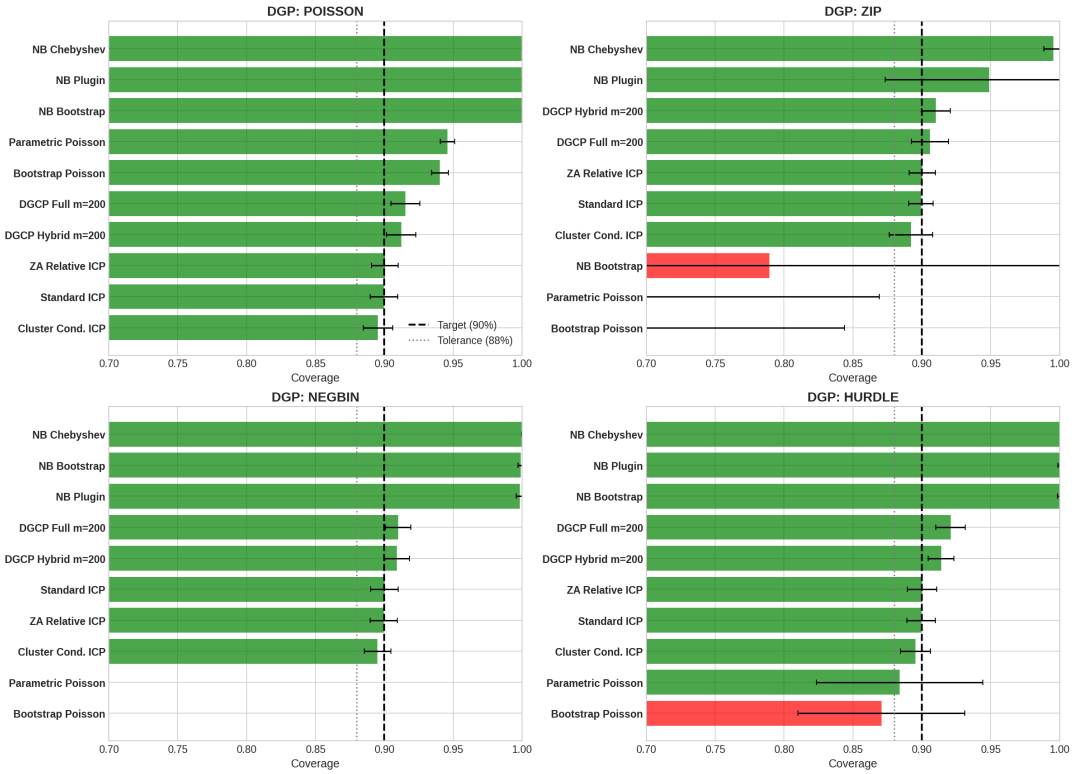

Figure 4.15: Coverage by DGP type across all methods. Each panel displays empirical coverage for a different data-generating process: Poisson (upper left), ZIP (upper right), Negative Binomial (lower left), and Hurdle (lower right). Green bars indicate methods achieving the 88% tolerance threshold; red bars indicate under-coverage. Error bars show standard deviation across repetitions. DGCP Hybrid m200 consistently achieves target coverage across all DGPs.

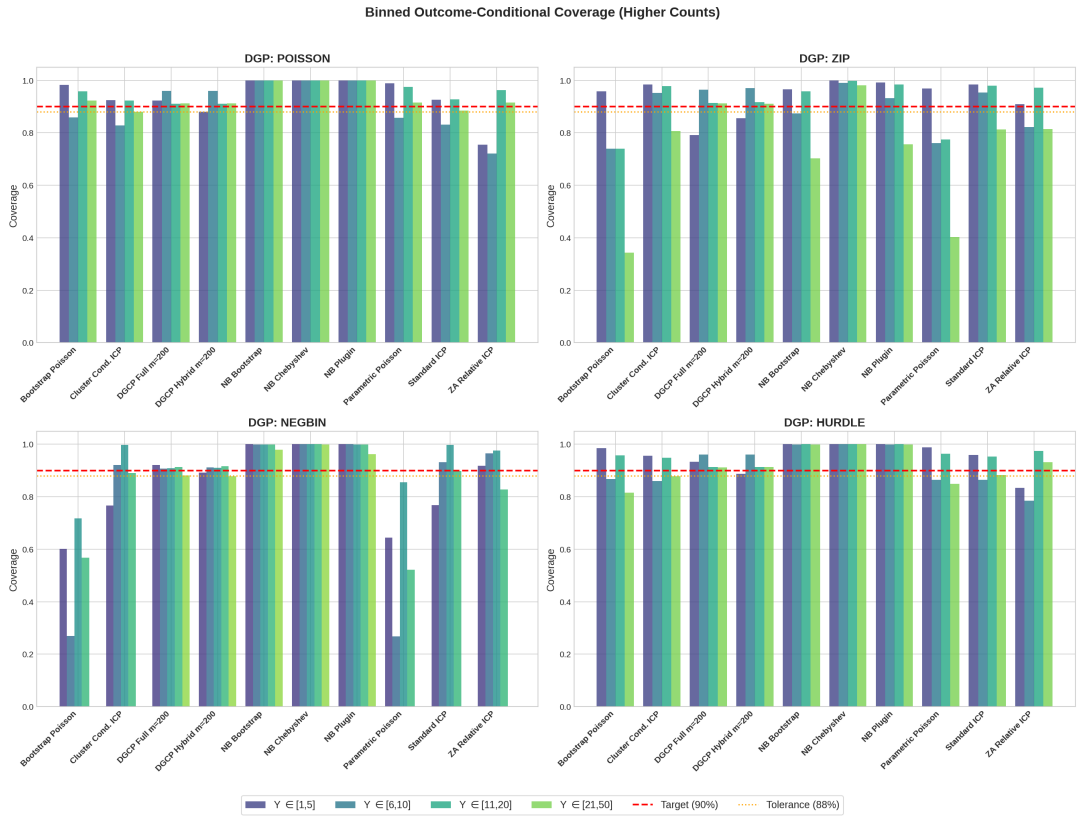

Figure 4.16: Binned outcome-conditional coverage for higher count values. Bars represent coverage for count ranges: $Y \in [1, 5]$, $Y \in [6, 10]$, $Y \in [11, 20]$, and $Y \in [21, 50]$. Standard ICP methods exhibit declining coverage for higher counts, while DGCP methods maintain more stable coverage across the outcome distribution. This demonstrates the importance of outcome-conditional calibration for count data with heavy tails.

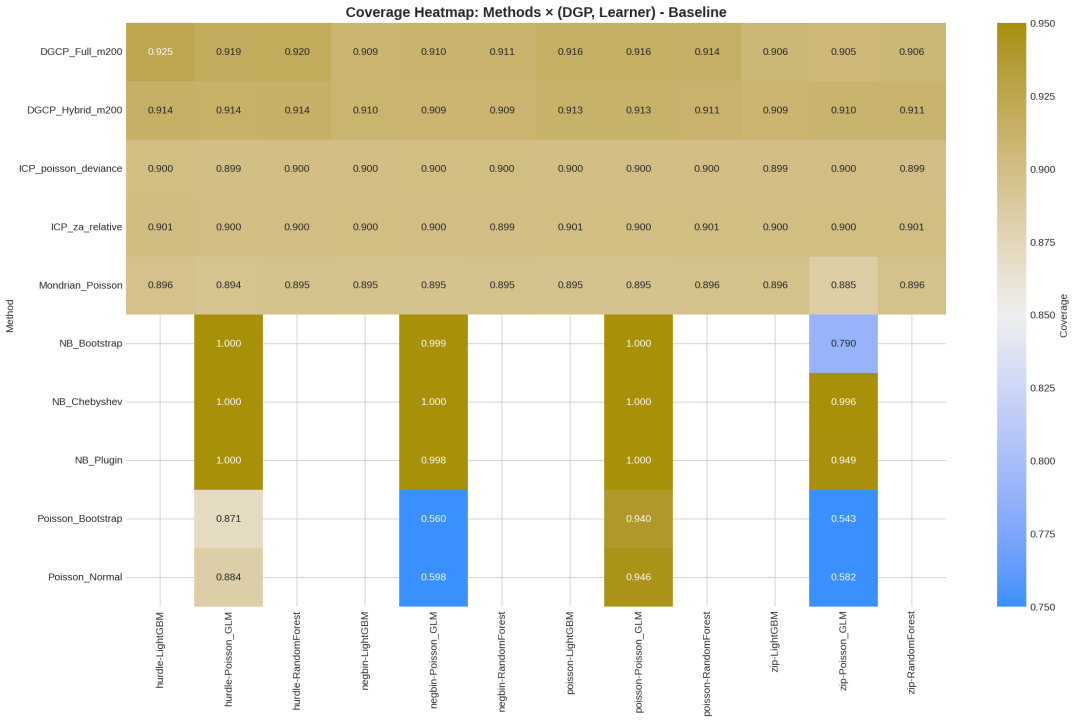

Figure 4.17: Coverage heatmap across all method, DGP, and learner combinations. Rows represent prediction interval methods; columns represent (DGP, Learner) pairs. Color scale from 0.75 to 0.95 indicates empirical coverage. Darker cells indicate under-coverage relative to the 90% target. DGCP Hybrid methods show consistent coverage across configurations, while parametric benchmarks exhibit more variation.

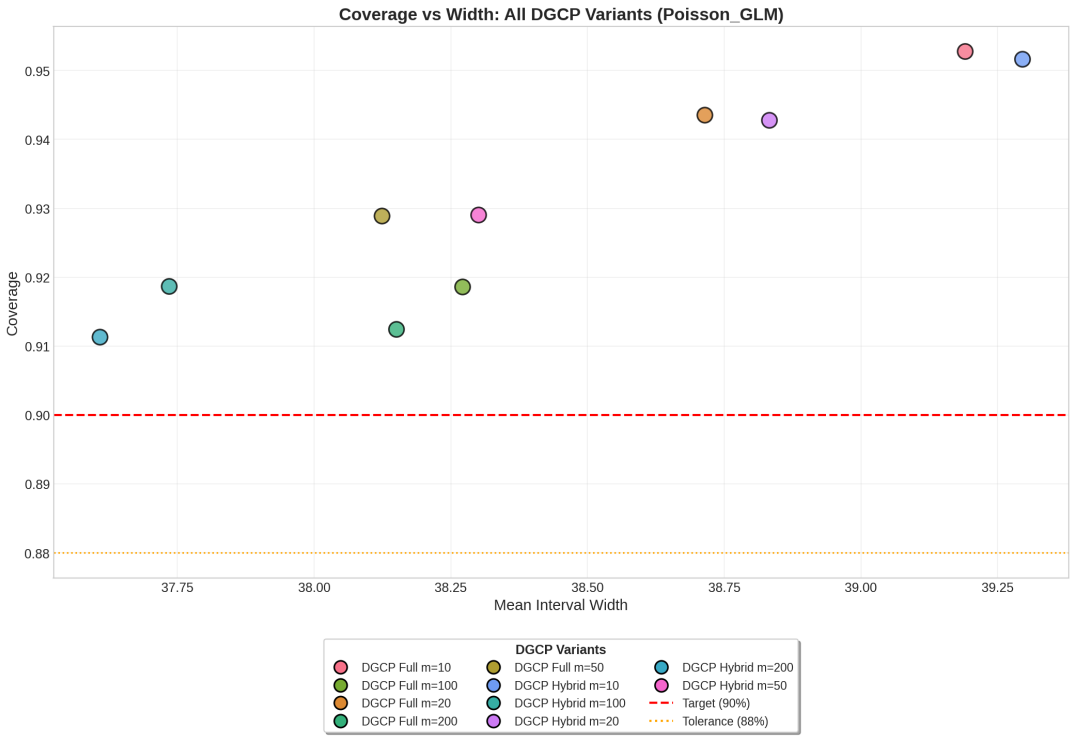

Figure 4.18: Coverage versus width trade-off for all DGCP variants using the Poisson GLM learner. Each point represents a DGCP configuration with different minimum group sizes ($m \in \{10, 20, 50, 100, 200\}$) and variants (Hybrid vs. Full). DGCP Hybrid methods cluster near the target coverage line with narrower intervals, while DGCP Full methods tend toward over-coverage with wider intervals. The $m = 200$ setting provides the best balance between coverage validity and interval efficiency.

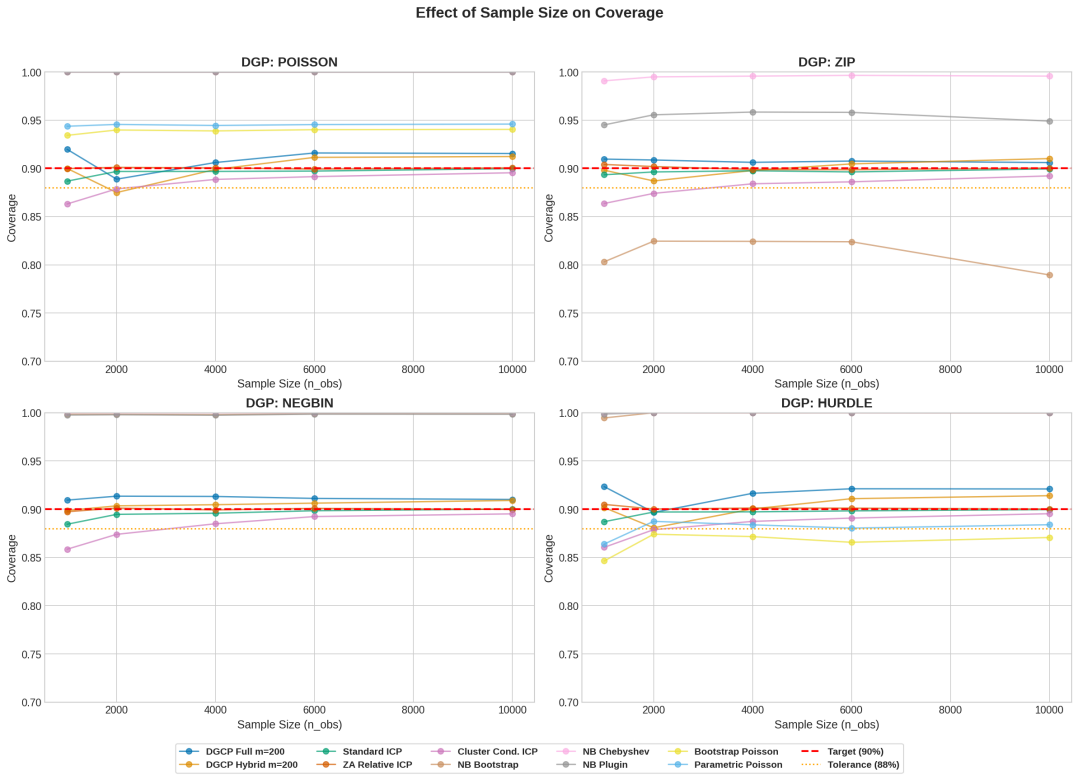

Figure 4.19: Effect of sample size on coverage. Each panel shows coverage trajectories as sample size increases from 1,000 to 10,000 observations. Conformal methods maintain stable coverage across sample sizes due to their finite-sample guarantees, while parametric methods (Poisson Normal, NB Plugin) show slight improvement with larger samples. DGCP methods benefit from larger calibration sets through improved group formation.

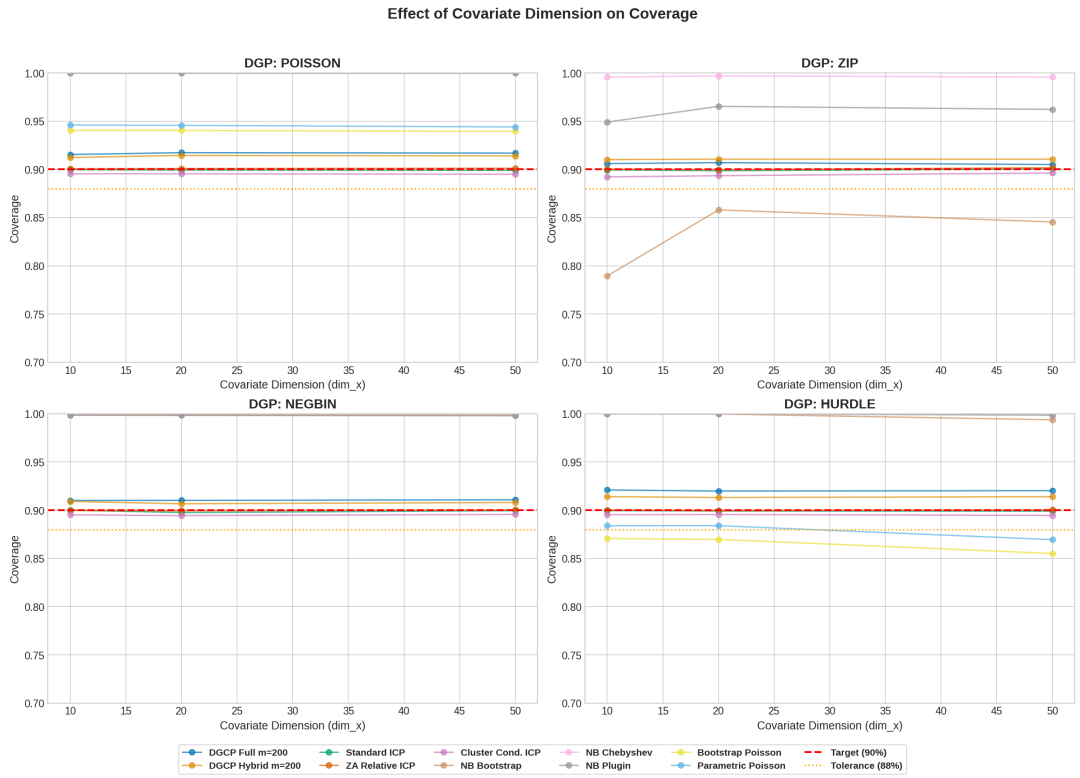

Figure 4.20: Effect of covariate dimension on coverage. Coverage is plotted against increasing covariate dimension ($d \in \{10, 20, 50\}$). All conformal methods maintain valid coverage regardless of dimensionality, demonstrating robustness to the curse of dimensionality. Parametric benchmarks show slight degradation in high dimensions due to increased model misspecification.

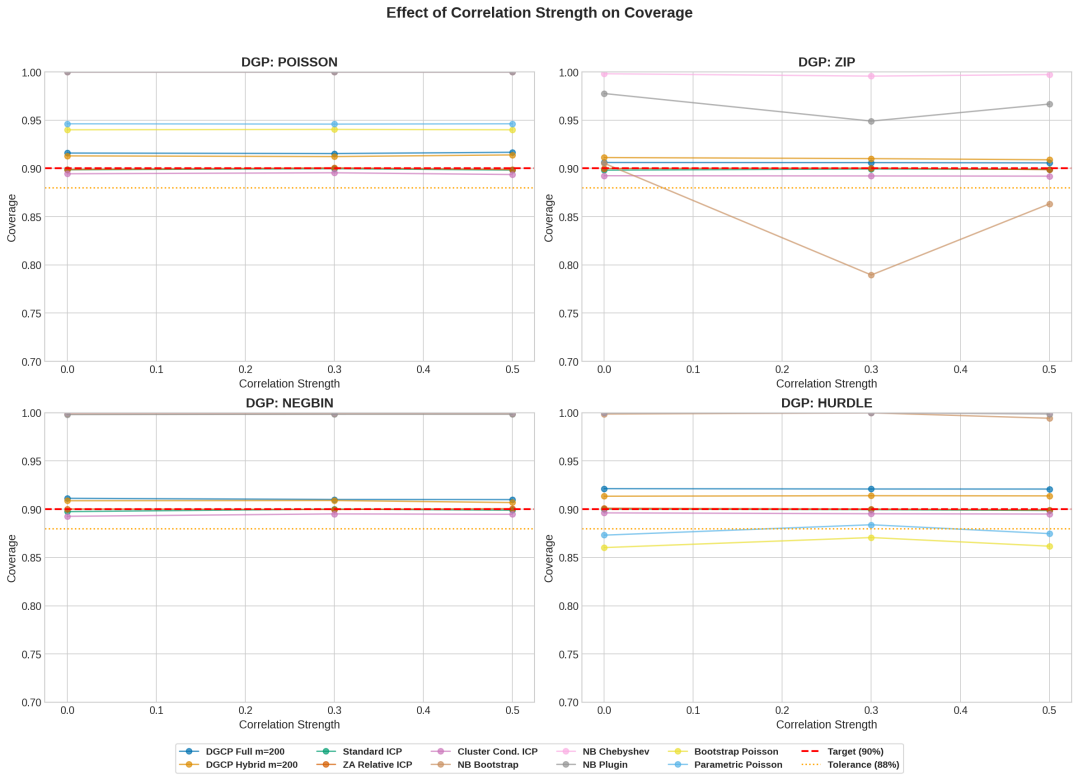

Figure 4.21: Effect of covariate correlation strength on coverage. Coverage is shown for correlation strengths $\rho \in \{0.0, 0.3, 0.5\}$ in the Toeplitz covariance structure. Conformal methods are largely insensitive to correlation structure, maintaining valid coverage across all settings. This robustness is a key advantage of distribution-free methods over parametric approaches that may be affected by multicollinearity.

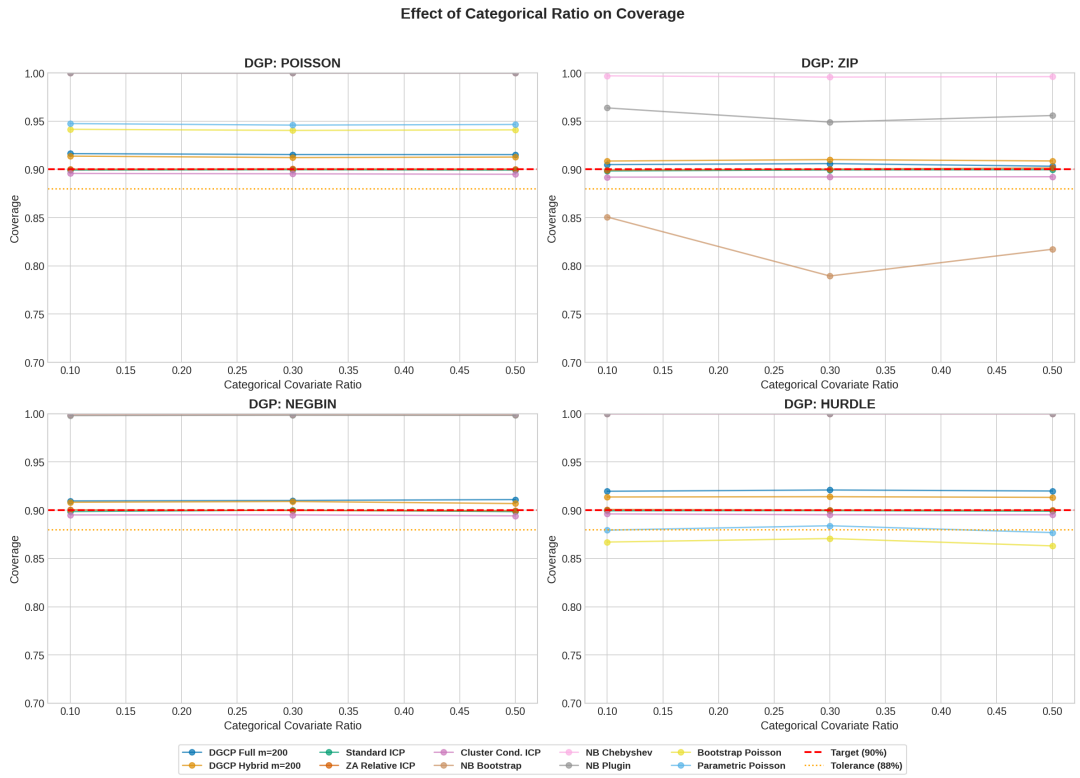

Figure 4.22: Effect of categorical covariate ratio on coverage. Coverage is plotted against the proportion of categorical covariates ($\rho_{\text{cat}} \in \{0.1, 0.3, 0.5\}$). All methods maintain stable coverage regardless of the mix of continuous and categorical predictors, indicating that the conformal framework handles mixed covariate types without special accommodation.

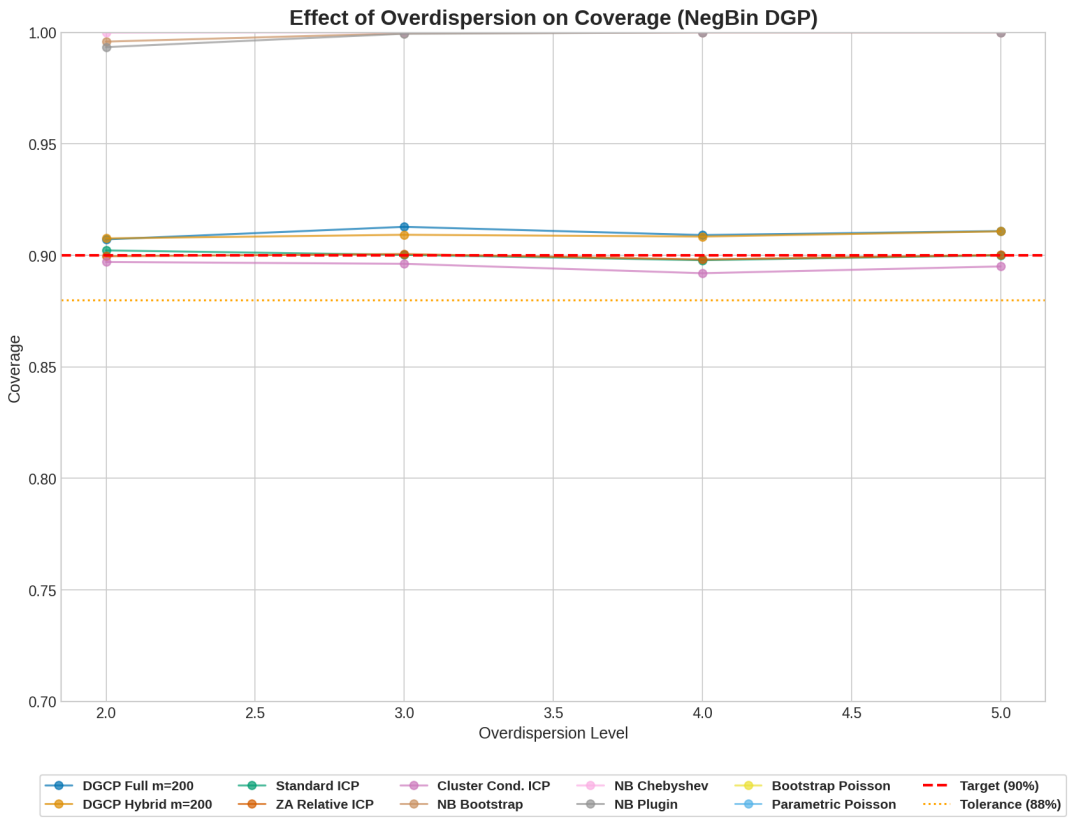

Figure 4.23: Effect of overdispersion on coverage (Negative Binomial DGP only). Coverage is shown for overdispersion parameters $\theta \in \{2.0, 3.0, 4.0, 5.0\}$, where lower θ indicates greater overdispersion. Standard ICP methods show declining coverage under severe overdispersion, while NB-based benchmarks and DGCP methods maintain validity. This demonstrates the importance of appropriate distributional assumptions or distribution-free methods when overdispersion is present.

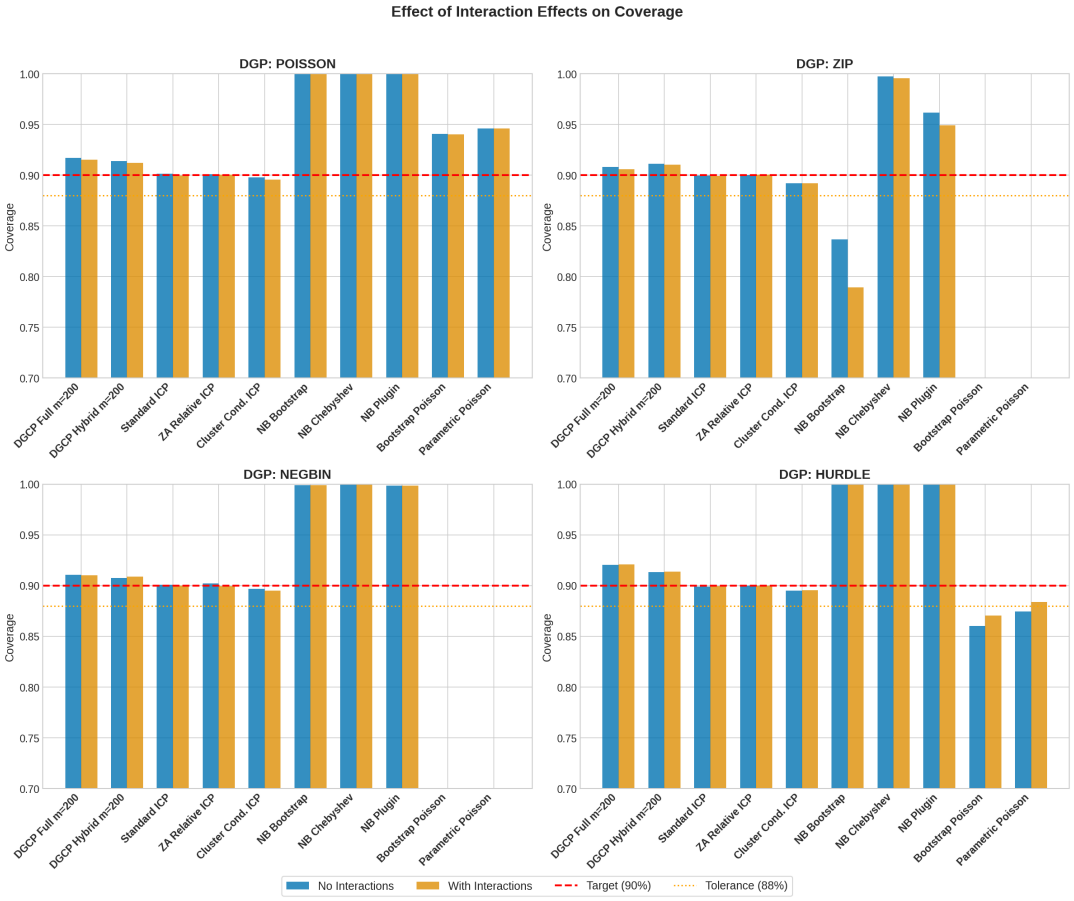

Figure 4.24: Comparison of coverage with and without covariate interaction effects. Grouped bars show coverage for each method under both settings. Conformal methods maintain valid coverage regardless of whether interaction effects are present in the DGP, demonstrating robustness to model misspecification when the base learner omits interaction terms. Parametric benchmarks show slight degradation when interactions are present but not modeled.

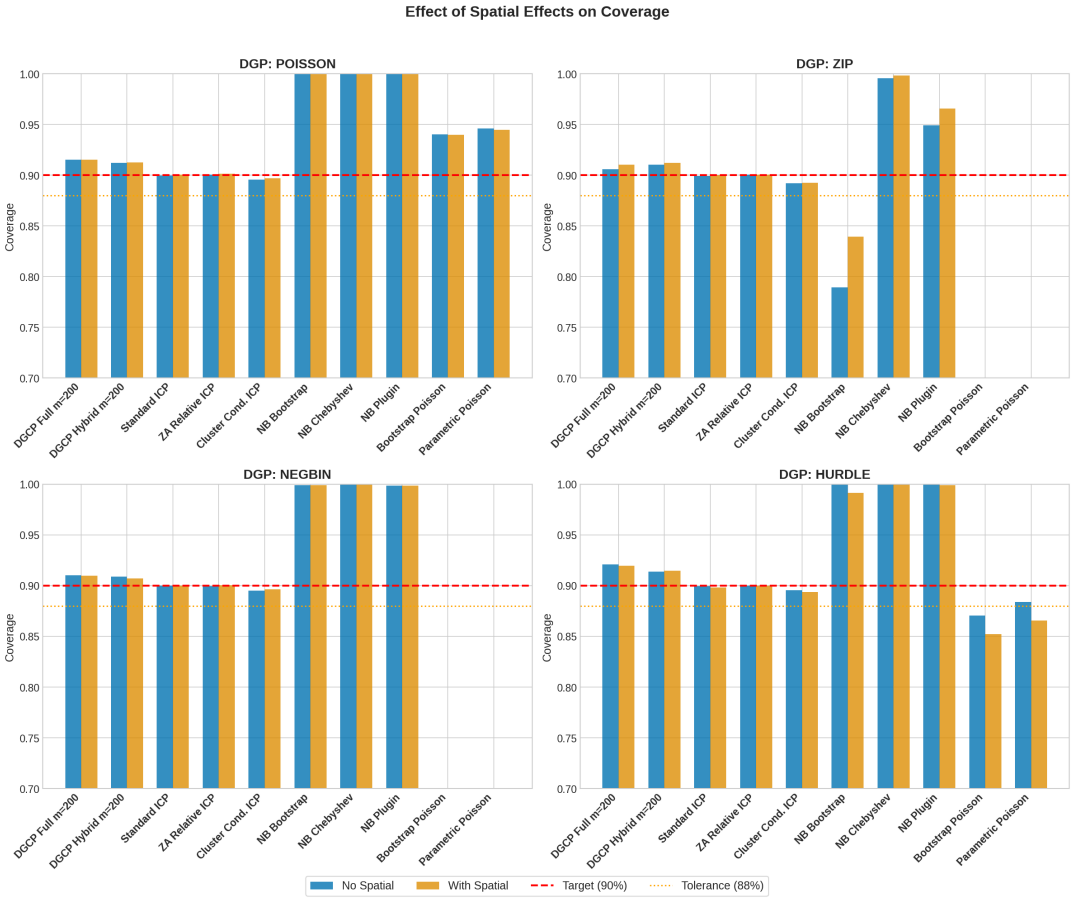

Figure 4.25: Comparison of coverage with and without spatial random effects. Grouped bars show coverage for each method under both settings. The inclusion of unmodeled spatial heterogeneity represents a form of omitted variable bias. Conformal methods maintain valid marginal coverage even when spatial effects are present but not captured by the base learner, while parametric methods show more sensitivity to this form of model misspecification.

4.8.4 Details on the German Car Insurance Application

Variable Overview

This appendix presents the variable overview for German motor insurance dataset.

Table 4.6: Variable Dictionary: Insurance Claims Dataset

Variable	Original	Description
Identifier		
PolicyholderID	vn_id	Anonymized policyholder ID
ContractID	vt_id	Anonymized contract ID
ContractState	VSTNR	Contract state number
Target		
Claims	sanz	Number of claims in the observation year
ClaimAmount	aufwand	Total claim amount in the observation year (EUR)
Exposure	je	Policy-years (exposure units)
Policyholder		
DrivAge	Alter_VN	Age of policyholder at year-end (categorical)
YoungDriv	Differenziertes_Nutzeralter	Young driver differentiation based on youngest user age (GDV definition)
AccompanyingDriving	Begleitendes_Fahren	Accompanied driving indicator (young drivers only)
Company	Firma	Policyholder type: Private or Company
Vehicle		
VehAge	Fahrzeugalter	Age of vehicle (categorical)
Brand	Hersteller	Vehicle manufacturer/brand
TypeClass_VK	TKL_VK	Type class for comprehensive insurance (VK)
Contract		
Mileage	Jahresfahrleistung	Agreed annual mileage (categorical)
ContractDuration	Laufzeit_Vertrag	Duration of the contract
UserFixed	Nutzergruppe	User group: Fixed drivers / Any driver
GaragePresent	Garage	Garage availability indicator (yes/no)
Bonus-Malus		
SF_Class_TK	SF_Klasse	No-claims bonus class for comprehensive ins. (M=Malus, TK ohne SF=No bonus)
SF_Class_KH	SF_Klasse_KH	No-claims bonus class for liability ins. (KH) (M=Malus)
Geographic		
DistrictNumber	Kreisnummer	District number of policyholder residence
FedState	Bundesland	Federal state (Bundesland)
Rural	—	Rural/Urban classification (Federal Statistical Office)
PopDensity	—	Population density (inhabitants/km²) (Federal Statistical Office)
Altitude	—	Mean altitude of district (meters) (Federal Statistical Office)
Socioeconomic		
MedianIncome	—	Median income in district (EUR) (Federal Employment Agency, 2019)
Perc_Empl	—	Percentage employed in district (Federal Statistical Office)

Note: Variables marked with "—" in Original Name were added from external sources (German Federal Statistical Office, Federal Employment Agency).
VK = Vollkasko (Comprehensive Insurance), KH = Kraftfahrzeug-Haftpflicht (Motor Liability Insurance).
SF_Class = Schadenfreiheitsklasse (No-Claims Bonus Class).

Table 4.7: Descriptive Statistics for Numeric Variables after Preprocessing

Variable	N	Mean	Std	Min	Q1	Median	Q3	Max	Missing
Claims	579456	0.10	0.31	0.00	0.00	0.00	0.00	3.00	0
Exposure	579456	0.86	0.27	0.01	0.87	1.00	1.00	1.00	0
ClaimAmount	579456	119.67	767.57	0.00	0.00	0.00	0.00	165276.00	0
SF_Class	579456	20.77	13.75	0.00	11.00	20.00	30.00	217.00	0
SF_Class_KH	579456	19.18	13.72	0.00	9.00	18.00	29.00	217.00	0
TypeClass_VK	579456	22.26	9.74	10.00	17.00	19.00	24.00	186.00	0
PopDensity	579456	872.49	1232.89	36.00	160.00	324.00	1043.00	19160.00	0
MedianIncome	579456	3874.01	1435.39	2494.00	3285.00	3450.00	3677.00	23940.00	0
Perc_Empl	579456	0.46	0.16	0.31	0.39	0.41	0.43	2.62	0
Altitude	579456	194.06	200.26	1.00	55.00	120.00	267.00	3032.00	0

Table 4.8: Descriptive Statistics for Categorical Variables after Preprocessing

Variable	N	Unique	Mode	Mode Freq	Mode %	Missing
Brand	579456	26	VW	111814	19.3%	0
GaragePresent	579456	2	ja	343763	59.3%	0
Company	579456	2	Privat	570349	98.4%	0
DrivAge	579456	12	51_55	95982	16.6%	0
YoungDriv	579456	11	51_55	81887	14.1%	0
VehAge	579456	13	02_03	124576	21.5%	0
Mileage	579456	4	00_10	269784	46.6%	0
ContractDuration	579456	11	00_01	149016	25.7%	0
UserFixed	579456	2	Festgelegte Fah	518816	89.5%	0
FedState	579456	16	NW	168004	29.0%	0
Rural	579456	2	1	433057	74.7%	0

Details on the Claims Frequency Modeling

This appendix presents supplementary tables from the claims frequency modeling on the German motor insurance dataset.

Table 4.9: Variable Importance: Ranked by Likelihood Ratio Test (LRT)

Rank	Variable	LRT	Pr(>Chi)	Sig.
1	Mileage	1932.72	0.00e+00	***
2	TypeClass_VK	1012.94	0.00e+00	***
3	VehAge	450.80	0.00e+00	***
4	Brand	386.65	0.00e+00	***
5	FedState	165.92	0.00e+00	***
6	YoungDriv	85.12	4.93e-14	***
7	SF_Class	58.22	2.34e-14	***
8	ContractDuration	56.05	2.01e-08	***
9	DrivAge	36.66	1.31e-04	***
10	GaragePresent	13.48	2.41e-04	***
11	Perc_Empl	11.67	6.34e-04	***
12	SF_Class_KH	6.26	1.23e-02	*
13	UserFixed	3.82	5.08e-02	.
14	PopDensity	3.56	5.92e-02	.
15	Altitude	1.16	2.81e-01	
16	MedianIncome	0.37	5.44e-01	
17	Company	0.25	6.15e-01	
18	Rural	0.20	6.58e-01	

Variables ranked by decrease in deviance when removed from full model.
Higher LRT indicates greater importance for model fit.

Table 4.10: Poisson GLM Sequential ANOVA: Terms Added Sequentially (First to Last)

Variable	**Df**	**Deviance**	**Resid.Df**	**Resid.Dev**	**Pr(>Chi)**	**Sig.**
NULL	–	–	579,455	273203.52	–	
Brand	25	1409.82	579,430	271793.70	0.00e+00	***
GaragePresent	1	158.03	579,429	271635.67	0.00e+00	***
Company	1	22.77	579,428	271612.90	1.83e-06	***
DrivAge	11	1191.68	579,417	270421.22	0.00e+00	***
YoungDriv	10	81.31	579,407	270339.92	2.78e-13	***
VehAge	12	1973.58	579,395	268366.33	0.00e+00	***
Mileage	3	2552.69	579,392	265813.64	0.00e+00	***
ContractDuration	10	107.33	579,382	265706.31	0.00e+00	***
UserFixed	1	2.42	579,381	265703.90	1.20e-01	
FedState	15	253.78	579,366	265450.11	0.00e+00	***
Rural	1	0.05	579,365	265450.06	8.15e-01	
SF_Class	1	120.32	579,364	265329.74	0.00e+00	***
PopDensity	1	224.50	579,363	265105.24	0.00e+00	***
MedianIncome	1	1310.95	579,362	263794.29	0.00e+00	***
Perc_Empl	1	32.76	579,361	263761.54	1.04e-08	***
Altitude	1	0.22	579,360	263761.31	6.39e-01	
TypeClass_VK	1	1009.17	579,359	262752.15	0.00e+00	***
SF_Class_KH	1	6.26	579,358	262745.89	1.23e-02	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 4.11: Poisson GLM Drop1 Analysis: Single Term Deletions from Full Model

Variable	Df	Deviance	AIC	LRT	Pr(>Chi)
Mileage	3	264678.61	385039.5	1932.72	0.00e+00 ***
TypeClass_VK	1	263758.83	384123.7	1012.94	0.00e+00 ***
VehAge	12	263196.69	383539.5	450.80	0.00e+00 ***
Brand	25	263132.54	383449.4	386.65	0.00e+00 ***
FedState	15	262911.81	383248.7	165.92	0.00e+00 ***
YoungDriv	10	262831.01	383177.9	85.12	4.93e-14 ***
SF_Class	1	262804.10	383169.0	58.22	2.34e-14 ***
ContractDuration	10	262801.93	383148.8	56.05	2.01e-08 ***
DrivAge	11	262782.54	383127.4	36.66	1.31e-04 ***
GaragePresent	1	262759.37	383124.2	13.48	2.41e-04 ***
Perc_Empl	1	262757.56	383122.4	11.67	6.34e-04 ***
SF_Class_KH	1	262752.15	383117.0	6.26	1.23e-02 *
UserFixed	1	262749.70	383114.6	3.82	5.08e-02 .
PopDensity	1	262749.45	383114.3	3.56	5.92e-02 .
Altitude	1	262747.05	383111.9	1.16	2.81e-01
MedianIncome	1	262746.25	383111.1	0.37	5.44e-01
Company	1	262746.14	383111.0	0.25	6.15e-01
Rural	1	262746.08	383110.9	0.20	6.58e-01

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.12: Forecast Dominance: Out-of-sample Deviance Losses (following Wüthrich and Merz (2023)Table 5.8)

Model	Poisson Deviance	NegBin Deviance
Poisson Null	0.4725	0.4016
Poisson GLM1	0.4545	0.3855
Poisson GLM2	0.4546	0.3855
Poisson GLM3	0.4547	0.3856
Poisson GLM4	0.4547	0.3856
NB Null	0.4725	0.4016
NB GLM1	0.4546	0.3854
NB GLM2	0.4547	0.3855
NB GLM3	0.4547	0.3856
NB GLM4	0.4547	0.3856
LightGBM	0.4526	0.3838
Random Forest	0.4541	0.3851

Table 4.13: Contingency Table: Observed vs Predicted Policy Counts (following Wüthrich and Merz (2023) Table 5.9)

Model	**0**	**1**	**2**	**3**	**4**	**5+**
Observed	103,862	11,890	102	0	0	0
Poisson Null	104,443	10,771	616	24	1	0
Poisson GLM1	104,566	10,548	696	40	3	0
Poisson GLM2	104,564	10,551	696	40	3	0
Poisson GLM3	104,564	10,551	696	40	3	0
Poisson GLM4	104,564	10,551	696	40	3	0
NB Null	104,980	9,757	999	105	11	1
NB GLM1	105,160	9,461	1,060	142	23	7
NB GLM2	105,158	9,464	1,060	142	23	7
NB GLM3	105,158	9,465	1,059	142	23	7
NB GLM4	105,157	9,465	1,060	142	23	7
LightGBM	104,573	10,546	696	37	2	0
Random Forest	104,563	10,556	696	37	2	0

Details on the Cluster Analysis

This appendix presents supplementary tables and figures from the empirical cluster analysis on the German motor insurance dataset.

Table 4.14: Cluster summary statistics

Cluster	N	Pct	Total_Claims	Claims_Freq	Pct_With_Claims
1	9808	1.693	927	0.109	9.400
2	39582	6.831	5219	0.157	13.039
3	30159	5.205	3202	0.123	10.514
4	10575	1.825	1033	0.113	9.674
5	36954	6.377	3266	0.101	8.789
6	42833	7.392	5197	0.142	12.033
7	19455	3.357	1995	0.123	10.172
8	83537	14.416	9806	0.138	11.621
9	14353	2.477	1643	0.131	11.308
10	7894	1.362	645	0.094	8.082
11	23568	4.067	2449	0.120	10.298
12	5839	1.008	479	0.095	8.203
13	11285	1.948	1076	0.108	9.490
14	12117	2.091	933	0.090	7.659
15	11096	1.915	919	0.095	8.246
16	161055	27.794	16511	0.119	10.159
17	34173	5.897	3220	0.109	9.341
18	5526	0.954	495	0.104	8.958
19	8065	1.392	816	0.128	9.994
20	7070	1.220	424	0.070	5.926
21	4512	0.779	203	0.056	4.455

Table 4.15: Cluster Characteristics: Size, Claims Distribution, and Dominant Features

Cluster	N	% Total	Frequency	Claims=0	Claims=1	Claims=2	Claims≥3	Dominant Characteristics
1	9,808	1.7	0.1088	8,886 (90.6%)	917 (9.3%)	5 (0.1%)	0 (0.0%)	Brand=Citroen (100%); Mileage=00_10 (47%)
2	39,582	6.8	0.1565	34,421 (87.0%)	5,104 (12.9%)	56 (0.1%)	1 (0.0%)	Brand=Audi (100%); Mileage=10_20 (51%)
3	30,159	5.2	0.1234	26,988 (89.5%)	3,140 (10.4%)	31 (0.1%)	0 (0.0%)	FedState=RP (100%); Mileage=10_20 (48%)
4	10,575	1.8	0.1127	9,552 (90.3%)	1,013 (9.6%)	10 (0.1%)	0 (0.0%)	Brand=KIA (100%); Mileage=10_20 (46%)
5	36,954	6.4	0.1012	33,706 (91.2%)	3,231 (8.7%)	16 (0.0%)	1 (0.0%)	Brand=Opel (100%); Mileage=00_10 (52%)
6	42,833	7.4	0.1416	37,679 (88.0%)	5,111 (11.9%)	43 (0.1%)	0 (0.0%)	FedState=NI (71%); Brand=VW (43%)
7	19,455	3.4	0.1234	17,476 (89.8%)	1,963 (10.1%)	16 (0.1%)	0 (0.0%)	FedState=B (53%); Mileage=00_10 (51%)
8	83,537	14.4	0.1378	73,829 (88.4%)	9,610 (11.5%)	98 (0.1%)	0 (0.0%)	Brand=BMW (53%); Mileage=10_20 (47%)
9	14,353	2.5	0.1308	12,730 (88.7%)	1,603 (11.2%)	20 (0.1%)	0 (0.0%)	FedState=TH (100%); Mileage=00_10 (45%)
10	7,894	1.4	0.0942	7,256 (91.9%)	631 (8.0%)	7 (0.1%)	0 (0.0%)	Brand=Dacia (100%); Mileage=00_10 (51%)
11	23,568	4.1	0.1198	21,141 (89.7%)	2,405 (10.2%)	22 (0.1%)	0 (0.0%)	Brand=Skoda (100%); Mileage=10_20 (49%)
12	5,839	1.0	0.0948	5,360 (91.8%)	479 (8.2%)	0 (0.0%)	0 (0.0%)	Brand=Suzuki (100%); Mileage=00_10 (59%)
13	11,285	1.9	0.1083	10,214 (90.5%)	1,066 (9.4%)	5 (0.0%)	0 (0.0%)	Brand=Nissan (100%); Mileage=00_10 (50%)
14	12,117	2.1	0.0900	11,189 (92.3%)	923 (7.6%)	5 (0.0%)	0 (0.0%)	Brand=Fiat (100%); Mileage=00_10 (60%)
15	11,096	1.9	0.0953	10,181 (91.8%)	911 (8.2%)	4 (0.0%)	0 (0.0%)	Brand=Peugeot (100%); Mileage=00_10 (51%)
16	161,055	27.8	0.1193	144,694 (89.8%)	16,212 (10.1%)	148 (0.1%)	1 (0.0%)	Brand=VW (49%); Mileage=00_10 (50%)
17	34,173	5.9	0.1088	30,981 (90.7%)	3,164 (9.3%)	28 (0.1%)	0 (0.0%)	Brand=Ford (100%); Mileage=10_20 (46%)
18	5,526	1.0	0.1041	5,031 (91.0%)	495 (9.0%)	0 (0.0%)	0 (0.0%)	Brand=Mitsubishi (100%); Mileage=00_10 (48%)
19	8,065	1.4	0.1275	7,259 (90.0%)	796 (9.9%)	10 (0.1%)	0 (0.0%)	Brand=Porsche (100%); Mileage=00_10 (60%)
20	7,070	1.2	0.0704	6,651 (94.1%)	414 (5.9%)	5 (0.1%)	0 (0.0%)	VehAge=16_17 (100%); Mileage=00_10 (68%)
21	4,512	0.8	0.0556	4,311 (95.5%)	199 (4.4%)	2 (0.0%)	0 (0.0%)	VehAge=20_24 (100%); Mileage=00_10 (81%)
Overall	579,456	100.0	0.1217	519,535 (89.7%)	59,387 (10.2%)	531 (0.1%)	3 (0.0%)	-

Note: Frequency is computed as total claims divided by total exposure within each cluster. Dominant characteristics show categorical variable modes with concentration $\geq 40\%$.

Table 4.16: Cluster descriptives, numerical variables

Cluster	N	Pct	SF_Class Mean	SF_Class Std	TypeClass_VK Mean	TypeClass_VK Std	Claims Freq	Pct Claimants
1	9808	1.7%	20.311	13.440	19.677	8.011	0.109	9.4%
2	39582	6.8%	19.881	13.913	25.282	10.877	0.157	13.0%
3	30159	5.2%	21.243	13.835	22.565	9.778	0.123	10.5%
4	10575	1.8%	20.600	13.621	21.452	8.670	0.113	9.7%
5	36954	6.4%	21.113	13.441	19.639	7.734	0.101	8.8%
6	42833	7.4%	20.787	14.323	23.407	10.408	0.142	12.0%
7	19455	3.4%	18.623	14.237	23.445	10.581	0.123	10.2%
8	83537	14.4%	20.520	14.058	24.162	10.524	0.138	11.6%
9	14353	2.5%	21.165	12.917	22.059	9.125	0.131	11.3%
10	7894	1.4%	20.625	12.587	18.005	6.099	0.094	8.1%
11	23568	4.1%	20.045	13.366	19.152	7.538	0.120	10.3%
12	5839	1.0%	21.361	13.586	21.055	7.880	0.095	8.2%
13	11285	1.9%	22.356	13.683	21.535	8.625	0.108	9.5%
14	12117	2.1%	18.378	13.515	18.307	7.660	0.090	7.7%
15	11096	1.9%	21.281	13.225	20.323	8.144	0.095	8.2%
16	161055	27.8%	20.961	13.648	22.252	9.559	0.119	10.2%
17	34173	5.9%	20.702	13.489	21.009	8.159	0.109	9.3%
18	5526	1.0%	21.723	13.266	23.864	8.935	0.104	9.0%
19	8065	1.4%	21.109	14.824	31.450	12.524	0.128	10.0%
20	7070	1.2%	25.682	11.904	16.134	7.496	0.070	5.9%
21	4512	0.8%	25.521	12.396	15.895	8.099	0.056	4.5%

Table 4.17: Cluster Characteristics: Categorical Variable Distributions

Cluster	DrivAge	VehAge	FedState	Brand
1	51_55(17%), 56_60(15%), 46_50(14%)	02_03(21%), 04_05(19%), 00_01(18%)	NW(30%), BW(11%), SN(10%)	Citroen(100%), Audi(0%), BMW(0%)
2	51_55(17%), 56_60(14%), 46_50(14%)	02_03(22%), 04_05(19%), 00_01(17%)	NW(32%), BY(15%), BW(13%)	Audi(100%), BMW(0%), Chrysler(0%)
3	51_55(17%), 56_60(16%), 61_65(12%)	02_03(21%), 04_05(19%), 00_01(16%)	RP(100%), B(0%), BW(0%)	VW(20%), Daimer/Mercedes(14%), BMW(12%)
4	51_55(17%), 56_60(15%), 46_50(13%)	02_03(25%), 04_05(22%), 00_01(22%)	NW(36%), BY(10%), NI(8%)	KIA(100%), Audi(0%), Chrysler(0%)
5	51_55(16%), 71+(16%), 56_60(15%)	02_03(22%), 04_05(19%), 06_07(15%)	NW(35%), BY(13%), BW(13%)	Opel(100%), Audi(0%), Chrysler(0%)
6	51_55(18%), 56_60(15%), 46_50(14%)	02_03(22%), 04_05(19%), 00_01(18%)	NI(71%), NW(11%), BW(3%)	VW(43%), Daimer/Mercedes(22%), BMW(9%)
7	51_55(16%), 71+(15%), 56_60(13%)	02_03(25%), 00_01(23%), 04_05(18%)	B(53%), HH(47%), BW(0%)	Daimer/Mercedes(20%), VW(19%), BMW(14%)
8	51_55(18%), 56_60(15%), 46_50(13%)	02_03(23%), 00_01(20%), 04_05(19%)	NW(38%), BY(16%), BW(13%)	BMW(53%), Hyundai(16%), SEAT(13%)
9	56_60(16%), 51_55(15%), 46_50(13%)	02_03(23%), 04_05(21%), 00_01(17%)	TH(100%), B(0%), BW(0%)	VW(24%), Skoda(10%), Daimer/Mercedes(10%)
10	56_60(18%), 51_55(17%), 61_65(14%)	02_03(27%), 00_01(27%), 04_05(21%)	NW(27%), BW(10%), RP(9%)	Dacia(100%), Audi(0%), BMW(0%)
11	51_55(15%), 56_60(14%), 61_65(12%)	02_03(25%), 00_01(22%), 04_05(20%)	NW(26%), BY(12%), SN(12%)	Skoda(100%), Audi(0%), Chrysler(0%)
12	56_60(17%), 51_55(16%), 71+(16%)	02_03(23%), 00_01(20%), 04_05(20%)	NW(25%), BY(12%), BW(11%)	Suzuki(100%), Audi(0%), Chrysler(0%)
13	51_55(17%), 56_60(16%), 71+(14%)	02_03(24%), 04_05(21%), 00_01(17%)	NW(30%), BW(11%), BY(10%)	Nissan(100%), Audi(0%), Chrysler(0%)
14	51_55(21%), 56_60(16%), 46_50(15%)	02_03(21%), 04_05(19%), 06_07(18%)	NW(29%), BY(16%), BW(14%)	Fiat(100%), Audi(0%), BMW(0%)
15	56_60(16%), 51_55(16%), 71+(14%)	02_03(19%), 04_05(17%), 00_01(16%)	NW(27%), BW(12%), BY(10%)	Peugeot(100%), Audi(0%), Chrysler(0%)
16	51_55(16%), 71+(15%), 56_60(14%)	02_03(20%), 04_05(19%), 06_07(16%)	NW(37%), BW(15%), BY(13%)	VW(49%), Daimer/Mercedes(31%), Renault(8%)
17	51_55(17%), 56_60(16%), 46_50(12%)	02_03(23%), 04_05(20%), 00_01(18%)	NW(36%), BW(13%), BY(11%)	Ford(100%), Audi(0%), BMW(0%)
18	51_55(18%), 56_60(18%), 61_65(14%)	00_01(26%), 02_03(22%), 04_05(17%)	NW(21%), BW(12%), BY(12%)	Mitsubishi(100%), Audi(0%), Chrysler(0%)
19	51_55(21%), 46_50(17%), 56_60(16%)	02_03(23%), 00_01(20%), 04_05(16%)	NW(28%), BY(15%), HE(14%)	Porsche(100%), Audi(0%), Chrysler(0%)
20	71+(27%), 56_60(15%), 51_55(14%)	16_17(100%), 02_03(0%), 04_05(0%)	NW(30%), BW(12%), NI(10%)	VW(21%), Daimer/Mercedes(20%), BMW(12%)
21	71+(35%), 56_60(13%), 51_55(13%)	20_24(100%), 02_03(0%), 04_05(0%)	NW(33%), BW(13%), BY(9%)	Daimer/Mercedes(26%), VW(17%), BMW(15%)

Figure 4.26: Covariate characteristics of the four clusters with lowest coverage under DGCP Hybrid m200. Left column: standardized deviations of continuous variables from global means. Middle column: categorical variable differences showing over/under-representation. Right column: distribution of the most distinctive categorical variable compared to global proportions. These clusters typically represent high-risk segments with atypical covariate combinations.

4.8.5 Alternative Clustering Approaches

Before detailing our chosen methodology, we briefly survey alternative clustering paradigms applicable to insurance portfolio segmentation. The literature distinguishes several major families of clustering methods (Kaufman and Rousseeuw 1990, Hastie et al. 2009, Wüthrich et al. 2026).

Hierarchical methods construct tree-like structures (dendrograms) of nested clusters through either agglomerative (bottom-up) or divisive (top-down) procedures. Given n observations $X_1, \ldots, X_n \in \mathcal{X} \subseteq \mathbb{R}^d$ and a dissimilarity function $L : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, agglomerative methods iteratively merge the two most similar clusters according to a linkage criterion such as single linkage $L(C_k, C_l) = \min_{X \in C_k, X' \in C_l} L(X, X')$, complete linkage using the maximum, or average linkage $L(C_k, C_l) = (|C_k| \cdot |C_l|)^{-1} \sum_{X \in C_k, X' \in C_l} L(X, X')$. While these approaches do not require pre-specification of the number of clusters K, the resulting dendrograms become difficult to interpret for large insurance portfolios with $n > 10^5$ policies, and computational complexity scales as $\mathcal{O}(n^2)$ for storage (Wüthrich et al. 2026). The classical centroid-based K-means algorithm (Hastie et al. 2009) partitions observations into K disjoint clusters $\mathscr{C} = \{C_1, \ldots, C_K\}$ by minimizing the within-cluster sum of squares $W(\mathscr{C}) = \sum_{k=1}^{K} \sum_{X_i \in C_k} \|X_i - \boldsymbol{\mu}_k\|_2^2$, where $\boldsymbol{\mu}_k = |C_k|^{-1} \sum_{X_i \in C_k} X_i$ denotes the cluster centroid; the algorithm alternates between assigning observations to their nearest centroid and updating centroids until convergence to a local minimum, but is restricted to continuous variables with squared Euclidean distance. For mixed-type data common in insurance applications, Huang (1998) proposed K-prototypes, which combines Euclidean distance for numerical variables with Hamming distance for categorical variables through a weighted sum $L(X_i, \boldsymbol{c}_k) = \sum_{j \in \mathcal{J}_{\text{num}}} (x_{i,j} - c_{k,j})^2 + \gamma \sum_{j \in \mathcal{J}_{\text{cat}}} \mathbf{1}\{x_{i,j} \neq c_{k,j}\}$, where $\gamma > 0$ controls relative importance; however, as demonstrated by Jamotton et al. (2024), K-prototypes requires careful tuning of this weighting parameter. Alternatively, K-medoids clustering with Gower distance (Gower 1971) selects actual observations as cluster centers (medoids) rather than computed means, solving $\arg\min_{(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K) \subset \mathcal{X}} \sum_{k=1}^{K} \sum_{X_i \in C_k} L(\boldsymbol{c}_k, X_i)$ via the partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw 1990), but computing pairwise Gower distances requires $\mathcal{O}(n^2)$ distance calculations which becomes prohibitive for large portfolios. Distribution-based approaches such as Gaussian Mixture Models assume observations arise from $f(X) = \sum_{k=1}^{K} p_k \cdot \mathcal{N}(X; \boldsymbol{\mu}_k, \Sigma_k)$ with parameters estimated via Expectation-Maximization, providing soft cluster assignments and capturing elliptical shapes through component-specific covariance matrices, but they assume continuous covariates and require additional preprocessing for categorical variables. Density-based methods like DBSCAN (Ester et al. 1996) identify clusters as dense regions separated by sparse areas, naturally handling non-convex shapes and detecting outliers, though the required hyperparameters (neighborhood radius ε and minimum density M) can be difficult to tune for insurance data with varying local densities. Finally, spectral clustering (Von Luxburg 2007) embeds data into a lower-dimensional space derived from the graph Laplacian, enabling detection of non-convex structures; Jamotton et al. (2024) demonstrate that spectral clustering can be combined with Burt distance for mixed-type insurance data, though the $\mathcal{O}(n^3)$ eigendecomposition requires prior data reduction for large portfolios.

We select the Burt distance-based K-means approach for several reasons: (i) it provides a principled treatment of categorical variables through the χ^2 distance framework, accounting for dependencies between modalities rather than treating categories as independent binary features; (ii) it scales efficiently to large insurance portfolios; (iii) it produces interpretable cluster assignments; and (iv) empirical comparisons in Jamotton et al. (2024) demonstrate superior goodness-of-fit (measured by Poisson deviance) compared to K-prototypes and K-medoids on motor insurance data.

Additional Prediction Interval Results

This appendix presents supplementary visualizations from the empirical analysis on the German motor insurance dataset.

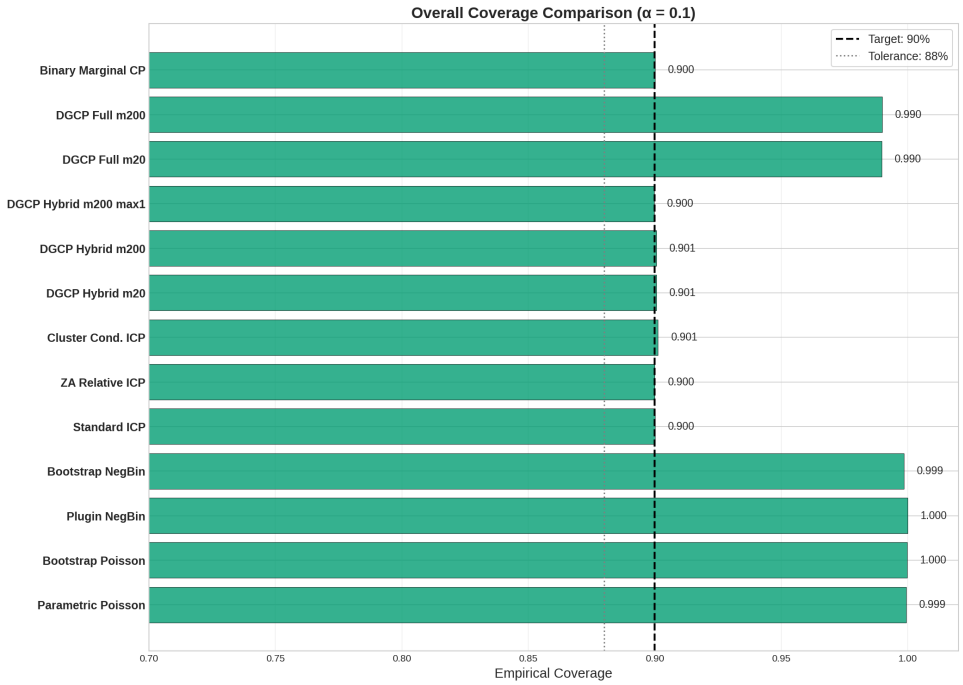

Figure 4.27: Overall empirical coverage by method. Green bars indicate methods achieving the 90% target coverage (within 2% tolerance); orange bars indicate methods falling below this threshold. Value labels show exact coverage rates. DGCP Hybrid variants and parametric benchmarks achieve target coverage, while standard ICP methods show marginal validity.

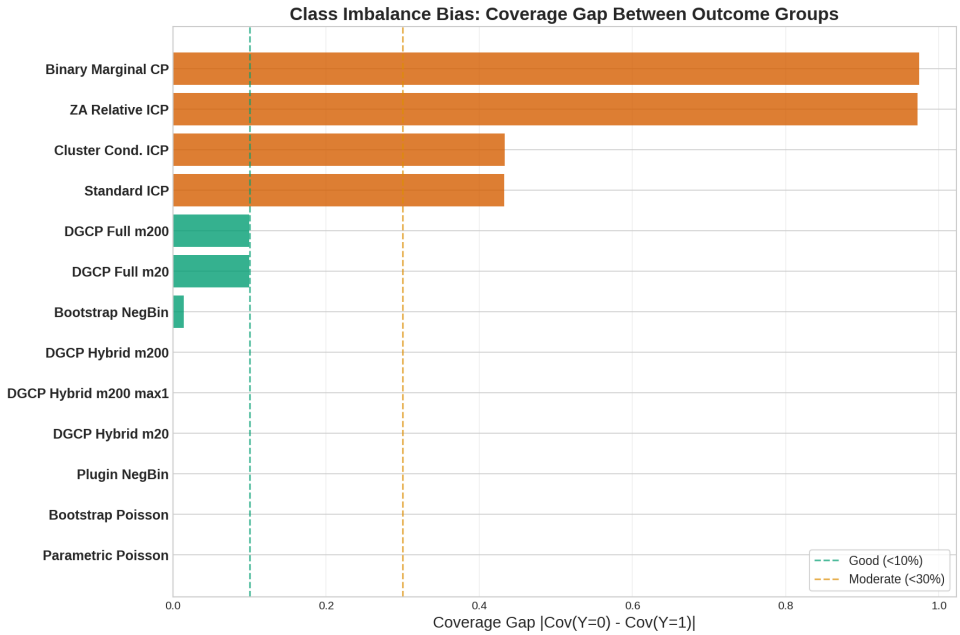

Figure 4.28: Coverage gap between outcome groups, measured as $|\text{Cov}(Y = 0) - \text{Cov}(Y = 1)|$. Methods are sorted by increasing gap. DGCP Hybrid methods exhibit minimal class imbalance bias (gap $< 10\%$), while standard ICP methods show substantial gaps exceeding 30%, indicating systematic under-coverage of the minority claimant class.

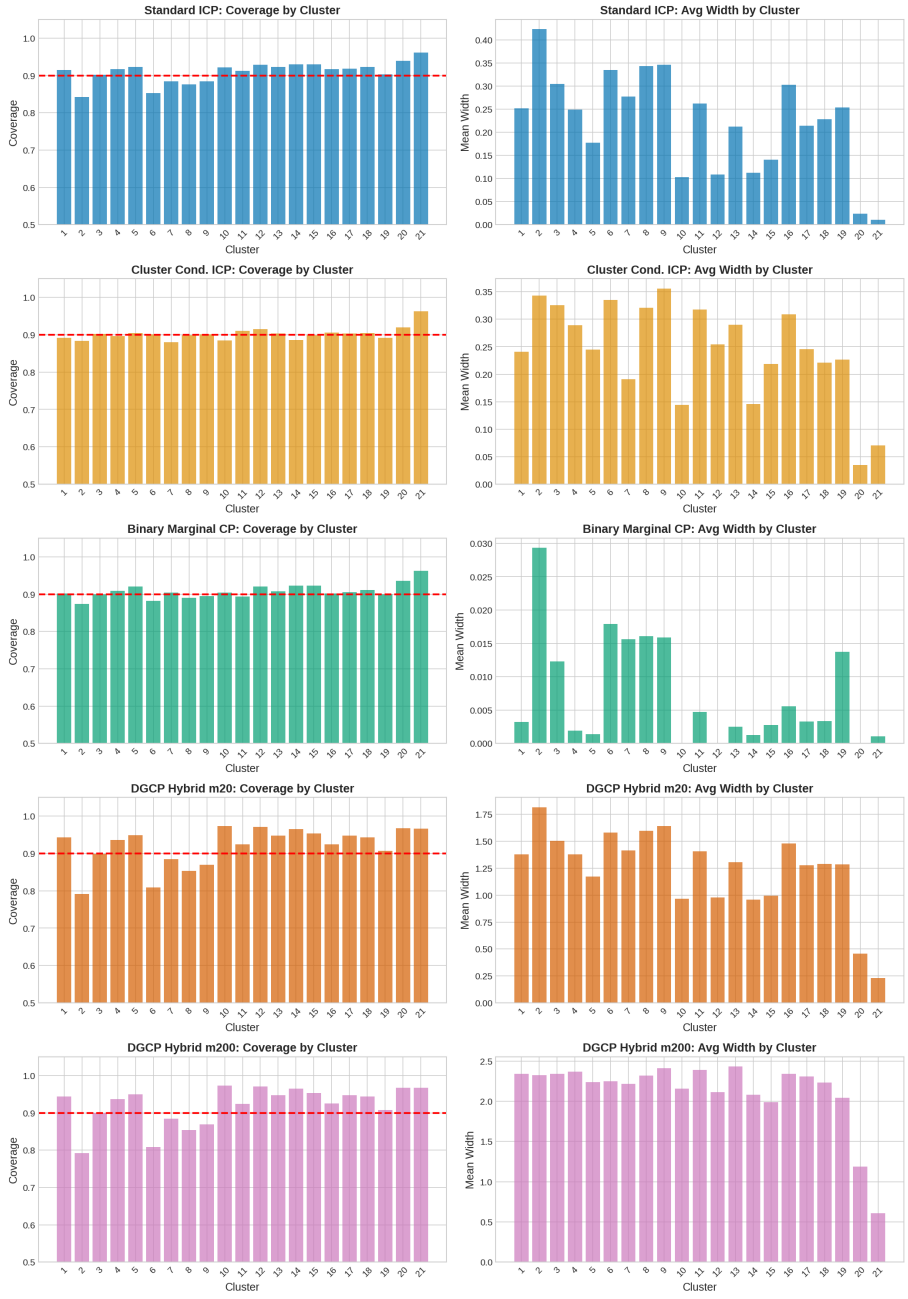

Figure 4.29: Coverage and mean interval width by actuarial risk cluster for five representative methods. Each row shows a different method; left panels display coverage by cluster with the 90% target line, right panels show corresponding mean widths. Cluster Conditional ICP achieves uniform coverage across clusters by design, while DGCP Hybrid methods maintain coverage with narrower intervals.

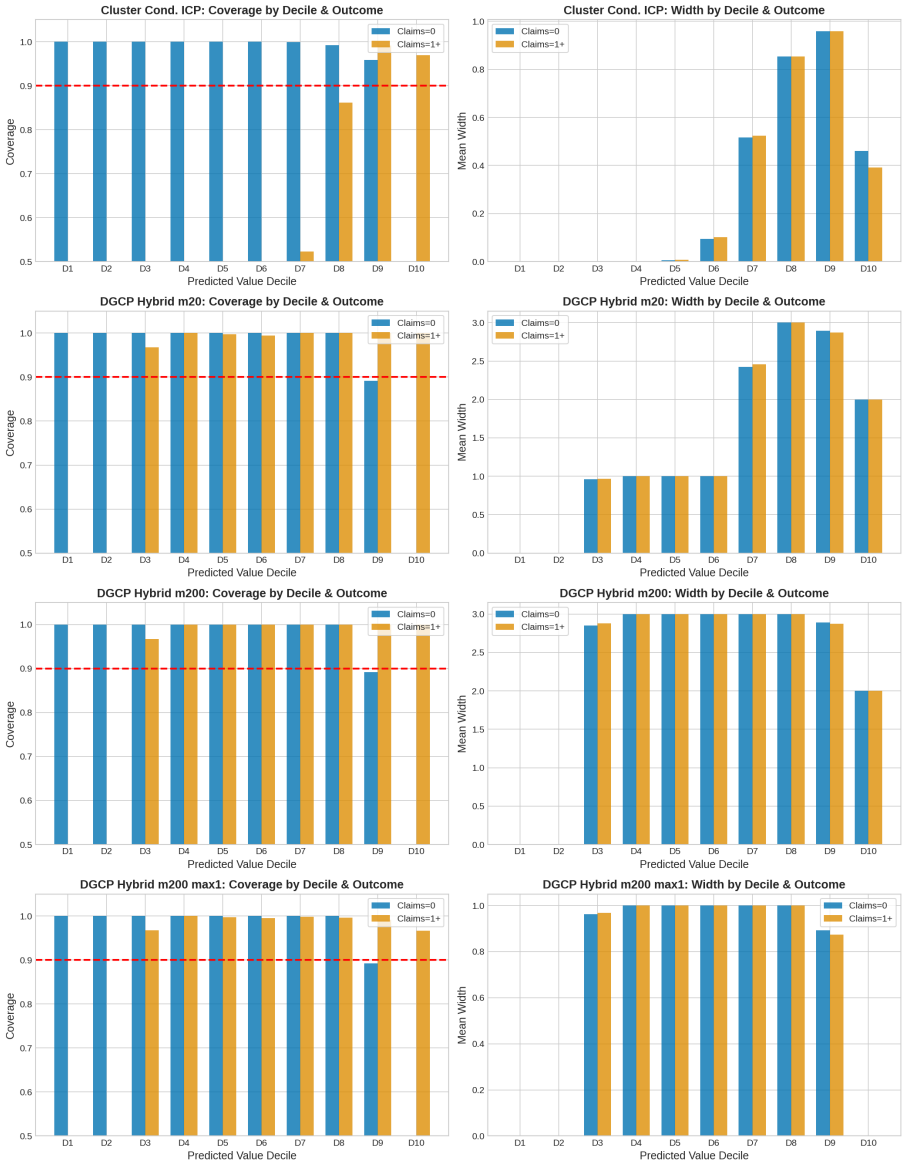

Figure 4.30: Coverage and width by predicted value decile, stratified by outcome group (Claims=0 vs. Claims=1+). Each row corresponds to a DGCP variant. Left panels show coverage rates; right panels show mean interval widths. DGCP Hybrid methods maintain balanced coverage across deciles for both outcome groups, unlike standard methods which exhibit systematic patterns.

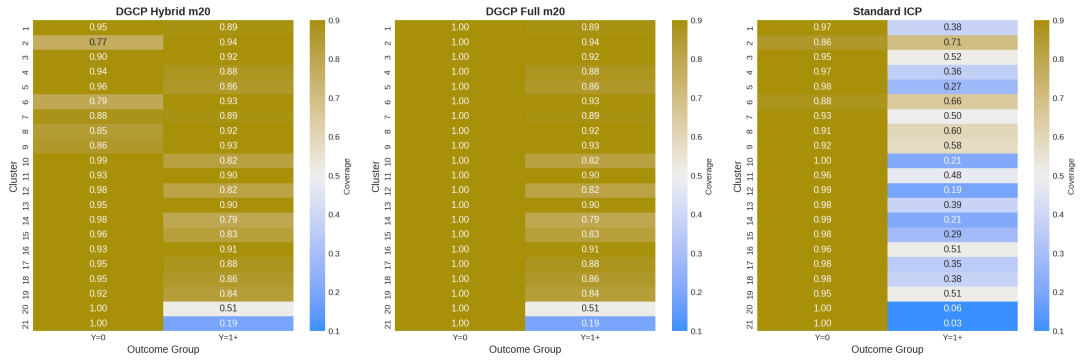

Figure 4.31: Coverage heatmaps by cluster and outcome group for DGCP m20 variants and Standard ICP. Cell values indicate empirical coverage rates; color scale ranges from red (under-coverage) through yellow to green (target coverage). Standard ICP shows pronounced under-coverage for claimants (Y=1+) across all clusters, while DGCP Hybrid m20 achieves balanced coverage.

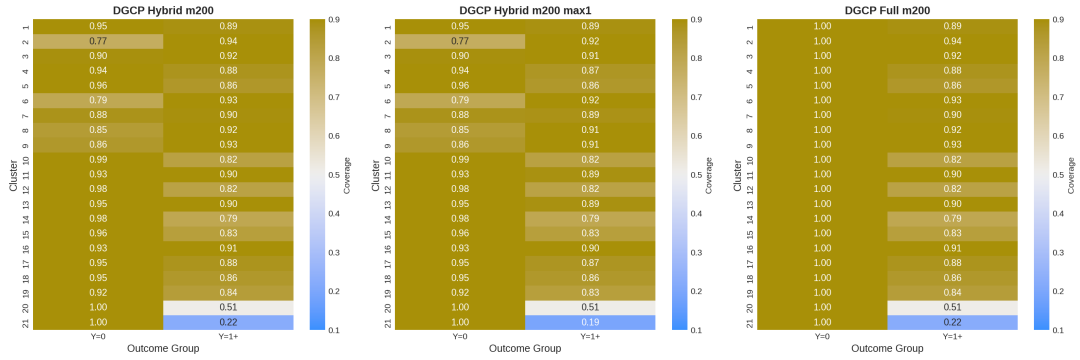

Figure 4.32: Heatmap of Outcome x Cluster Coverage - DGCP m200 and Deviance ICP - Coverage Heatmaps)

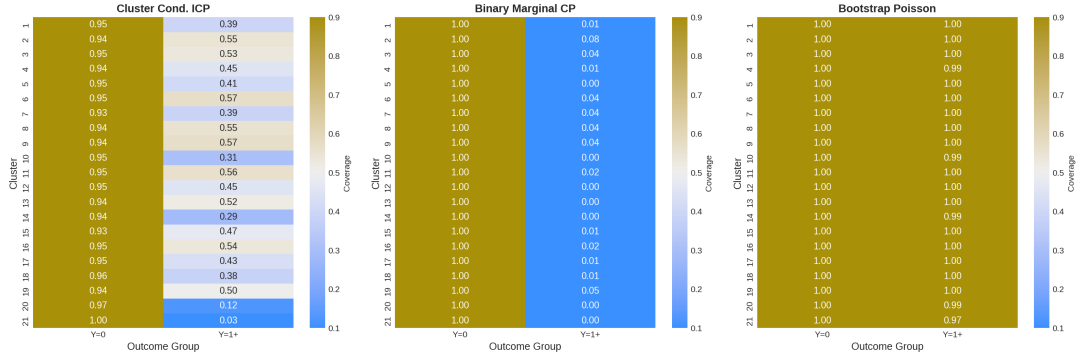

Figure 4.33: Coverage heatmaps for Cluster Conditional ICP, Binary Marginal CP, and Bootstrap Poisson. Cluster Conditional ICP achieves uniform coverage across clusters but exhibits outcome-conditional bias similar to standard ICP. Binary Marginal CP shows balanced binary coverage. Bootstrap Poisson achieves high coverage uniformly through conservative wide intervals.

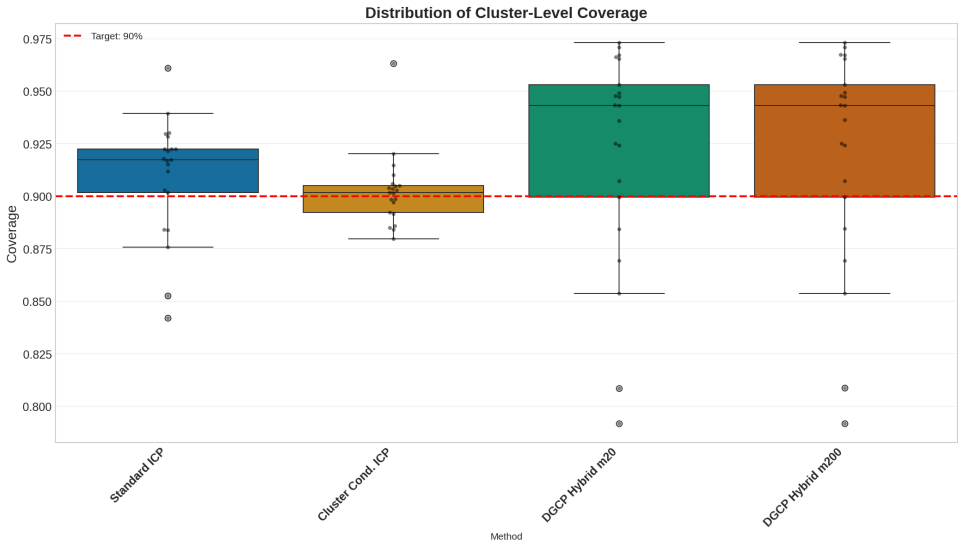

Figure 4.34: Distribution of cluster-level coverage rates across methods. Box plots show the spread of coverage across actuarial risk clusters; individual points represent specific clusters. DGCP Hybrid m200 exhibits the tightest distribution around the 90% target, indicating consistent performance across heterogeneous risk segments.

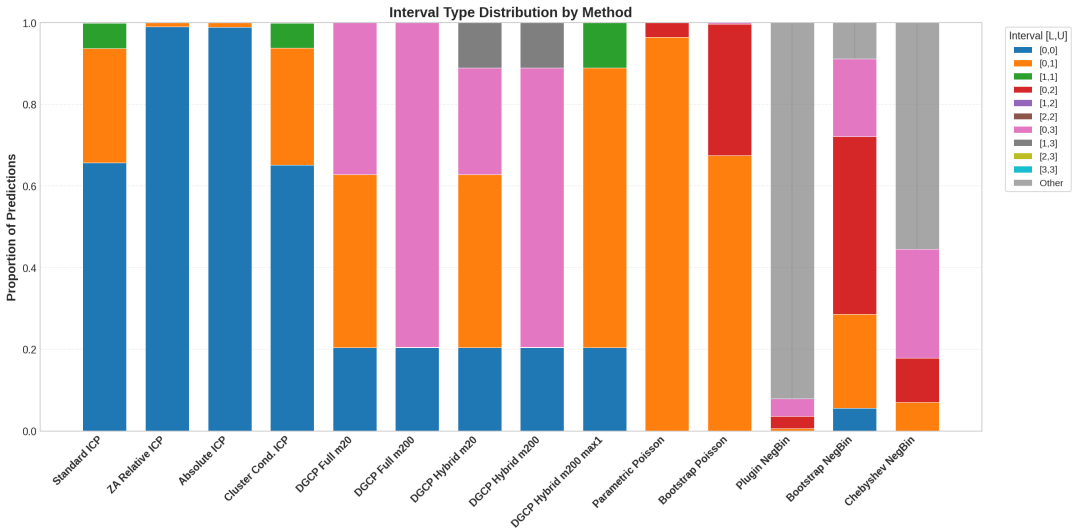

Figure 4.35: Distribution of prediction interval types across all evaluated methods. Stacked bars show the proportion of each interval type $[L, U]$ produced by each method. Methods producing predominantly $[0, 0]$ or $[0, 1]$ intervals offer limited risk differentiation, while DGCP Hybrid variants generate diverse interval types including $[1, 1]$ singletons and $[1, 2]$, $[1, 3]$ ranges that exclude zero.

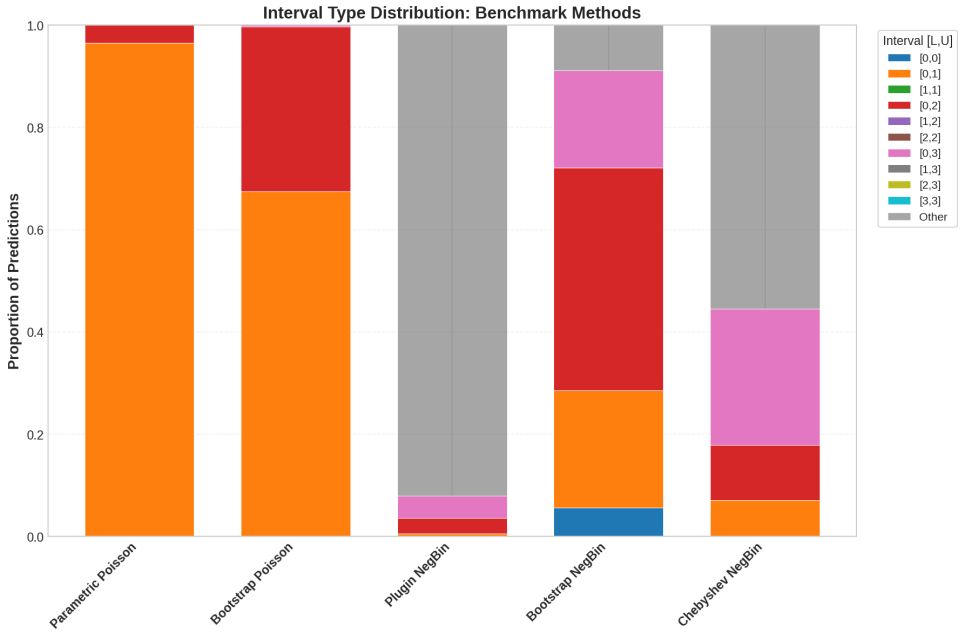

Figure 4.36: Interval type distribution for parametric benchmark methods. All benchmarks produce intervals dominated by $[0, 1]$, $[0, 2]$, or $[0, 3]$ types, reflecting their conservative approach of including zero in virtually all predictions. This limits their utility for identifying high-risk policies where claims are likely.

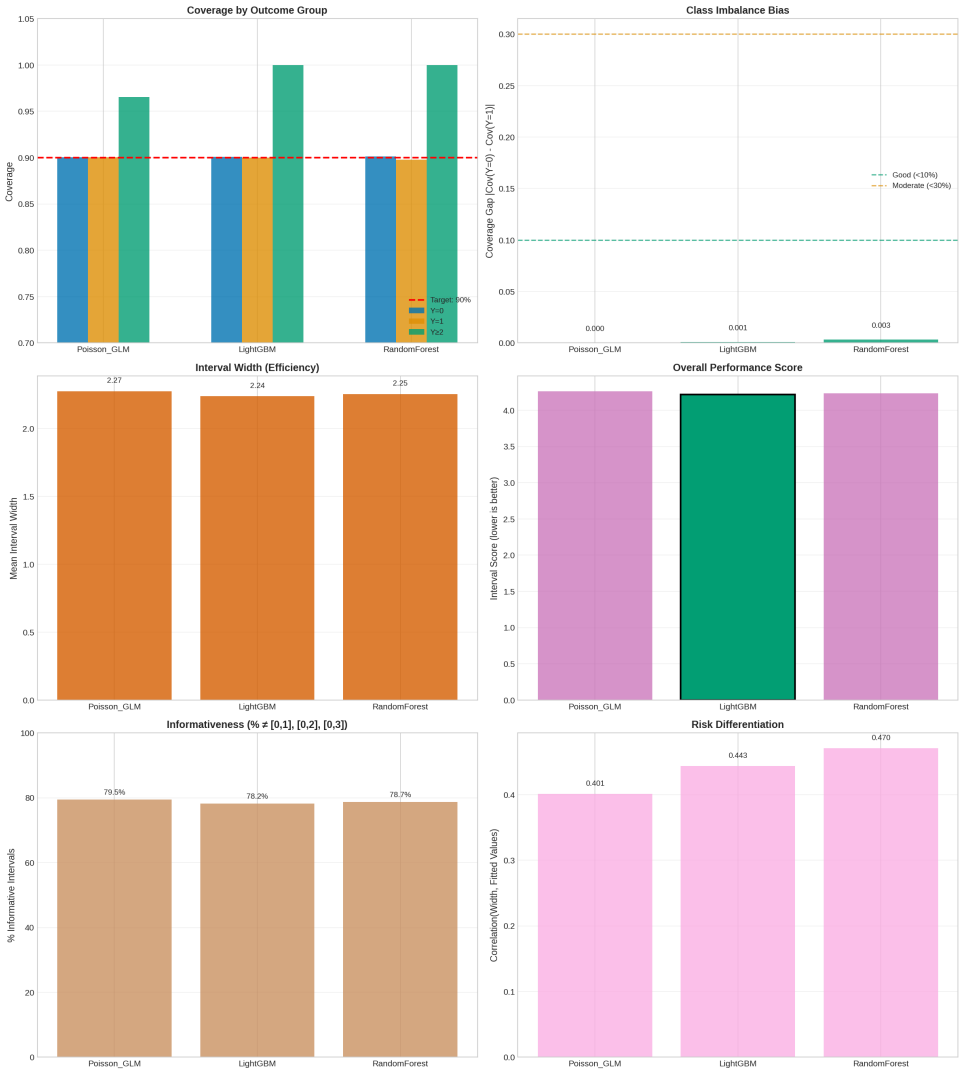

Figure 4.37: Detailed Comparison of DGCP Hybrid m200 results for the different underlying frequency models: Poisson GML, LightGBM, and Random Forest.

Chapter 5

General Conclusion and Outlook

This dissertation has addressed estimation robustness across causal inference and predictive uncertainty. For a detailed discussion of the individual contributions, I refer to the conclusions at the end of each chapter. Here, a high-level overview is provided, and directions for future research are outlined.

Chapter 2 extends the double machine learning framework to combinatorial treatment regimes. By formalizing marginal attribution through a regularized tempered policy and integrating simulation-based inference, the proposed estimator enables valid inference on marginal treatment effects in settings where the 2^d treatment configurations render standard approaches intractable. The empirical application to NHANES data identifies obesity and smoking as the strongest modifiable determinants of blood pressure. Chapter 3 investigates propensity score calibration within the DML framework. We establish that isotonic calibration preserves the theoretical guarantees of DML and provide systematic guidance on the interaction between calibration methods, sample-splitting schemes, and base learners. The key practical finding is that full-sample calibration using cross-fitted propensity scores consistently stabilizes inverse-propensity-weighted estimators without degrading performance when calibration is unnecessary. Chapter 4 develops conformal prediction methods for zero-inflated count data. The proposed DGCP Hybrid framework resolves the coverage imbalance of standard conformal methods on imbalanced data, and the Burt distance-based clustering diagnostic identifies portfolio segments with systematically elevated prediction uncertainty.

Several directions for future research emerge from this work. The double machine learning framework, while theoretically mature, is still relatively young as a tool in applied work. We are observing a transition from primarily academic use to industrial applications, which introduces new requirements: non-standard data types, higher-dimensional treatment spaces, and demands for automation. The multiple treatment framework of Chapter 2 could benefit from techniques developed in the recommender systems literature, where sparse user-item interaction matrices present analogous challenges to high-dimensional treatment configurations and could provide scalable solutions for even larger treatment dimensions. More broadly, the combination of tabular foundation models with the DML framework represents a promising direction for automating nuisance estimation. The emerging interest in causal reasoning within large language models further suggests that the demand for robust, extensible causal estimation frameworks will continue to grow.

For conformal prediction in actuarial science, a natural extension of Chapter 4 is composite frequency-severity models, where the DGCP framework would need to accommodate the continuous severity component alongside the discrete, zero-inflated frequency structure. More generally, actuarial applications present domain-specific challenges – regulatory constraints, long-tailed distributions, and the need for interpretable uncertainty communication – that require tailored solutions rather than off-the-shelf methods. There is also a natural intersection with the causal perspective: insurers increasingly seek to understand not only the predicted risk of a policyholder but also the causal effect of interventions such as pricing changes or risk prevention programs. Bridging

the causal and predictive uncertainty frameworks developed in this dissertation could provide a foundation for such analyzes.

With this dissertation, I hope to contribute to making modern statistical learning frameworks more applicable in practice, and thus help to avoid unreliable conclusions that may arise when general-purpose methods are applied without accounting for the specific structure of the problem at hand.

Bibliography

David J. Aldous. Exchangeability and related topics. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg. ISBN 978-3-540-39316-0.

David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. SIAM, 2007.

James E. Ash, Yajie Zou, Dominique Lord, and Yinhai Wang. Comparison of confidence and prediction intervals for different mixed-Poisson regression models. *Journal of Transportation Safety & Security*, 13(3):357–379, 2021.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, January 2021.

Peter C. Austin and Elizabeth A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679, 2015.

Philipp Bach, Victor Chernozhukov, Malte S. Kurz, and Martin Spindler. DoubleML – An object-oriented implementation of double machine learning in Python. *Journal of Machine Learning Research*, 23(53):1–6, 2022.

Philipp Bach, Malte S. Kurz, Victor Chernozhukov, Martin Spindler, and Sven Klaassen. DoubleML: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108(3):1–56, 2024a.

Philipp Bach, Oliver Schacht, Victor Chernozhukov, Sven Klaassen, and Martin Spindler. Hyperparameter tuning for causal inference with double machine learning: A simulation study. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 1065–1117. PMLR, 2024b.

Lee J. Bain and Jagdish K. Patel. Prediction intervals based on partial observations for some discrete distributions. *IEEE Transactions on Reliability*, 42(3):459–463, 1993.

Daniele Ballinari. Calibrating doubly-robust estimators with unbalanced treatment assignment, 2024.

Daniele Ballinari and Nora Bearth. Improving the finite sample estimation of average treatment effects using double/debiased machine learning with propensity score calibration, 2025.

R. E. Barlow and H. D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972. ISSN 01621459, 1537274X.

Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. Econml: A python package for ml-based heterogeneous treatment effects estimation. https://github.com/py-why/EconML, 2019. Version 0.16.

Valbona Bejleri and Balgobin Nandram. Bayesian and frequentist prediction limits for the Poisson distribution. *Communications in Statistics – Theory and Methods*, 47(17):4254–4271, 2018.

Alexandre Belloni, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Ying Wei. Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Annals of statistics*, 46(6B):3643, 2018.

Michela Bia, Martin Huber, and Lukáš Lafférs. Double machine learning for sample selection models. *Journal of Business & Economic Statistics*, 42(3):958–969, 2024.

M. Sh. Birman and M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes w_p^α. *Matematicheskii Sbornik*, 73 (115)(3):295–317, 1967.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *Proceedings of the 32nd International Conference on Machine Learning*, pages 1613–1622, 2015.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Cyril Burt. The factorial analysis of qualitative data. *British Journal of Statistical Psychology*, 3(3): 166–185, 1950.

Matias Busso, John DiNardo, and Justin McCrary. New evidence on the finite sample properties of propensity score reweighting and matching estimators. *The Review of Economics and Statistics*, 96(5):885–897, 2014. ISSN 00346535, 15309142.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration?, 2023.

Emmanuel J. Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society: Series B*, 85(1):24–45, 2023.

Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS). National health and nutrition examination survey (nhanes). `https://www.cdc.gov/nchs/nhanes/index.html`, 2024. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Nian-Cih Chang. Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, 23(2):177–191, 2020.

Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. Causalml: Python package for causal machine learning, 2020.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564 – 1597, 2014.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. *National Bureau of Economic Research Working Paper*, (w30302), 2022a.

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, May 2022b.

Victor Chernozhukov, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via riesz regression, 2024.

Victor Chernozhukov, Whitney K Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of riesz representers. *Journal of the American Statistical Association*, (just-accepted): 1–23, 2025.

Harold D. Chiang, Kengo Kato, Yukun Ma, and Yuya Sasaki. Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics*, 40(3):1046–1056, 2022.

Paul S. Clarke and Alessio Polselli. Double machine learning for static panel models with fixed effects. *Econometrics Journal*, 29(1):1–20, 2026.

Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *Journal of Econometrics*, 232(2):468–496, 2023.

D. R. Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3-4):562–565, 12 1958. ISSN 0006-3444.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap Methods and Their Application*, volume 1. Cambridge University Press, 1997.

A. P. Dawid. *Probability Forecasting*. John Wiley & Sons, Ltd, 2014. ISBN 9781118445112.

Michel Denuit, Donatien Hainaut, and Julien Trufin. *Effective Statistical Learning Methods for Actuaries II: Tree-Based Methods and Extensions*. Springer, New York, 2020.

Shachi Deshpande and Volodymyr Kuleshov. Calibrated propensity scores for causal effect estimation. *arXiv preprint arXiv:2306.00382*, 2023.

Ricardo Diaz-Rincon, Muxuan Liang, Adolfo Ramirez-Zamora, and Benjamin Shickel. Uncertainty-aware prediction of parkinson's disease medication needs: A two-stage conformal prediction approach. In *Proceedings of Machine Learning for Health (ML4H)*, 2024.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, volume 36, pages 64555–64576. Curran Associates, Inc., 2023.

Bradley Efron. *Bootstrap methods: another look at the jackknife*, volume 7. Institute of Mathematical Statistics, 1979.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.

Ludwig Fahrmeir and Gerhard Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, 1994.

Ping Feng, Xiao-Hua Zhou, Qing-Ming Zou, Ming-Yu Fan, and Xiao-Song Li. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, 31(7):681–697, 2012.

Andrea Ferrario and Roger Hämmerli. On boosting: Theory and applications. SSRN Electronic Journal, 2019. SSRN Manuscript ID 3402687, Version June 11, 2019.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: A unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 08 2020.

Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

David Gamarnik. Efficient learning of monotone concepts via quadratic optimization. In *Proceedings of the eleventh annual conference on computational learning theory*, pages 134–143, 1998.

A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 148–155. Morgan Kaufmann Publishers Inc., 1998.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 edition, 2013.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

T Gneiting. Calibration of medium-range weather forecasts, 03/2014 2014.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions. *Electronic Journal of Statistics*, 7, 06 2011.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

John C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4): 857–871, 1971.

Bryan S. Graham, Cristine Campos De Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies*, 79(3): 1053–1079, 04 2012. ISSN 0034-6527.

Helton Graziadei, Paulo C. Marques F., Eduardo F. L. de Melo, and Rodrigo S. Targino. Conformal prediction for frequency-severity modeling. *arXiv preprint*, July 2024.

Michael J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, London, 1984.

Noah Greifer. *WeightIt: Weighting for Covariate Balance in Observational Studies*, 2025. URL `https://ngreifer.github.io/WeightIt/`. R package version 1.4.0, https://github.com/ngreifer/WeightIt.

Justin Grimmer, Dean Knox, and Brandon Stewart. Naive regression requires weaker assumptions than factor models to adjust for multiple cause confounding. *Journal of Machine Learning Research*, 24(182):1–70, 2023.

Susan Gruber and Mark Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6:26–26, 08 2010.

Chirag Gupta and Aaditya K. Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting, 2021.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *Advances in Neural Information Processing Systems*, volume 33, pages 3711–3723. Curran Associates, Inc., 2020.

Rom Gutman, Ehud Karavani, and Yishai Shimoni. Improving inverse probability weighting by post-calibrating its propensity scores. *Epidemiology*, 35(4):473–480, 2024. ISSN 1044-3983.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Vassilis A Hajivassiliou and Daniel L McFadden. The method of simulated scores for the estimation of ldv models. *Econometrica*, pages 863–896, 1998.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.

Caleb Hays and Manish Raghavan. Double machine learning for causal inference under shared-state interference. *arXiv preprint arXiv:2504.08836*, 2025.

Liangyuan Hu and Chenyang Gu. Estimation of causal effects of multiple treatments in healthcare database studies with rare outcomes. *Health Services and Outcomes Research Methodology*, 21:1–22, 09 2021.

Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B*, 76(1):243–263, 2014.

Kosuke Imai and David van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99:854–866, 02 2004.

Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86:4–29, 02 2004.

Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

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

Jean-François Ingenbleek and Jean Lemaire. What is a sports car? *ASTIN Bulletin*, 18(2):175–187, 1988.

Charlotte Jamotton, Donatien Hainaut, and Thomas Hames. Insurance analytics with clustering techniques. *Risks*, 12(9):141, 2024.

Marshall M. Joffe and Paul R. Rosenbaum. Invited commentary: Propensity scores. *American Journal of Epidemiology*, 150(4):327–333, 08 1999.

Ulf Johansson, Tuwe Löfström, and Cecilia Sönströd. Well-calibrated probabilistic predictive maintenance using venn-abers, 2023.

Nathan Kallus. Treatment effect risk: Bounds and inference. *Manage. Sci.*, 69(8):4579–4590, 2023. ISSN 0025-1909.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167): 1–63, 2020.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Localized debiased machine learning: efficient inference on quantile treatment effects and beyond. *J. Mach. Learn. Res.*, 25(1), 2024. ISSN 1532-4435.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

Edward H Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

Taeho Kim, Benjamin Lieberman, George Luta, and Edsel A. Peña. Prediction regions for Poisson and over-dispersed Poisson regression models with applications to forecasting number of deaths during the COVID-19 pandemic. *arXiv preprint arXiv:2007.02105*, 2020.

Taeho Kim, Benjamin Lieberman, George Luta, and Edsel A. Peña. Prediction intervals for Poisson-based regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(5): e1568, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2014.

Steven Klaassen, Julian Teichert-Kluge, Philipp Bach, and Johannes Kueck. Doublemldeep: Estimation of causal effects with multimodal data. *arXiv preprint arXiv:2407.13912*, 2024.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

K. Krishnamoorthy and Jie Peng. Improved closed-form prediction intervals for binomial and Poisson distributions. *Journal of Statistical Planning and Inference*, 141(5):1709–1718, 2011.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, February 2019. ISSN 1091-6490.

Antonis Lambrou, Harris Papadopoulos, Ilia Nouretdinov, and Alexander Gammerman. Reliable probability estimates based on support vector machines for large multiclass datasets. *IFIP Advances in Information and Communication Technology*, 382:182–191, 09 2012.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0):191–246, 2006.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Greg Lewis and Vasilis Syrgkanis. Double/debiased machine learning for dynamic treatment effects. In *Advances in Neural Information Processing Systems*, volume 34, pages 22695–22707. Curran Associates, Inc., 2021.

Fan Li and Fan Li. Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389–2415, 2019.

Yufan Li and Pragya Sur. Optimal and provable calibration in high-dimensional binary classification: Angular calibration and platt scaling, 2025.

Molei Liu, Yi Zhang, and Doudou Zhou. Double/debiased machine learning for logistic partially linear model. *The Econometrics Journal*, 24(3):559–588, 2021.

Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.

Enno Mammen and Kyusang Yu. Additive isotone regression. *Lecture Notes-Monograph Series*, pages 179–195, 2007.

Daniel F. McCaffrey, Beth Ann Griffin, Daniel Almirall, Mary Ellen Slaughter, Rajeev Ramchand, and Lane F. Burgette. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414, 2013.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, 2nd edition, 1989.

William Q. Meeker, Gerald J. Hahn, and Luis A. Escobar. *Statistical Intervals: A Guide for Practitioners and Researchers*, volume 541. John Wiley & Sons, Hoboken, NJ, 2 edition, 2017.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7: 983–999, 2006.

Iván Díaz Muñoz and Mark Van Der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

Raymond H. Myers and Douglas C. Montgomery. A tutorial on generalized linear models. *Journal of Quality Technology*, 29(3):274–291, 1997.

Pakdaman Mahdi Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907, 04 2015.

John A. Nelder and Robert W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015.

X Nie and S Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, 2020. ISSN 0006-3444.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning, 2020.

Christoph Noack, Tomasz Olma, and Christoph Rothe. Automated bias-corrected inference for regression discontinuity designs. *arXiv preprint arXiv:2302.11452*, 2024.

Ilia Nouretdinov, Denis Volkhonskiy, Pitt Lim, Paolo Toccaceli, and Alexander Gammerman. Inductive Venn-Abers predictive distribution. In *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pages 15–36. PMLR, 11–13 Jun 2018.

David J. Olive, Ruwan C. Rathnayake, and Michael G. Haile. Prediction intervals for GLMs, GAMs, and some survival regression models. *Communications in Statistics – Theory and Methods*, 51(22):7818–7832, 2022.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*. InTech, 2008.

Jens-Michalis Papaioannou, Sebastian Jäger, Alexei Figueroa, David Stutz, Betty van Aken, Keno Bressem, Wolfgang Nejdl, Felix Gers, Alexander Löser, and Felix Biessmann. Robust conformal prediction for infrequent classes. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. NCoRE: Neural Counterfactual Representation Learning for Combinations of Treatments. *arXiv preprint arXiv:2103.11175*, 2021.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python, 2018.

Ivan Petej. venn-abers. `https://github.com/ip200/venn-abers`, 2024.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. Estimating multi-cause treatment effects via single-cause perturbation. In *Advances in Neural Information Processing Systems*, volume 34, pages 23754–23767. Curran Associates, Inc., 2021.

David Randahl, Jonathan P. Williams, and Håvard Hegre. Bin-conditional conformal prediction of fatalities from armed conflict. *Political Analysis*, 34(1):96–108, 2026.

Simon Rentzmann and Mario V. Wüthrich. Unsupervised learning: What is a sports car? Technical Report 3439358, SSRN, 2019.

Ronald Richman and Mario V Wüthrich. A neural network extension of the lee–carter model to multiple populations. *Annals of Actuarial Science*, 15(2):346–366, 2021.

Ronald Richman and Mario V. Wüthrich. LASSO regularization within the LocalGLMnet architecture. *Advances in Data Analysis and Classification*, 17(4):951–981, 2023a.

Ronald Richman and Mario V. Wüthrich. LocalGLMnet: Interpretable deep learning for tabular data. *Scandinavian Actuarial Journal*, 2023(1):71–95, 2023b.

Peter M. Robinson. Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Pedro H. C. Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122, 2020.

Jürg Schelldorfer and Mario V. Wüthrich. Nesting classical actuarial models into neural networks. Technical Report 3320525, SSRN, 2019. Version January 25, 2019.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1670–1679, New York, New York, USA, 20–22 Jun 2016. PMLR.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR, 06–11 Aug 2017.

Lloyd S Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

Yan Sun, Pratik Chaudhari, Ian J. Barnett, and Edgar Dobriban. A confidence interval for the ℓ_2 expected calibration error, 2024.

Ichiro Takeuchi, Quoc V Le, Timothy D Sears, and Alexander J Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

Zhiqiang Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107, 10 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 12 1996.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Zoltan Toth, Olivier Talagrand, and Yuejian Zhu. *The attributes of forecast systems: a general framework for the evaluation and calibration of weather forecasts*, page 584–595. Cambridge University Press, 2006.

Konstantinos Tsoumas and Harris Papadopoulos. Insurance claim prediction using unbiased confidence guarantees. In *Proceedings of AIAI 2024*, 2024.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

AW van der Vaart and Jon A Wellner. Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 127–384. Springer, 2023.

Sara A van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Lars van der Laan and Ahmed M. Alaa. Self-calibrating conformal prediction, 2024.

Lars van der Laan, Ernesto Ulloa-Perez, Marco Carone, and Alex Luedtke. Causal isotonic calibration for heterogeneous treatment effects. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34831–34854. PMLR, 23–29 Jul 2023.

Lars van der Laan, Ziming Lin, Marco Carone, and Alex Luedtke. Stabilized inverse probability weighting via isotonic calibration, 2024a.

Lars van der Laan, Alex Luedtke, and Marco Carone. Automatic doubly robust inference for linear functionals via calibrated debiased machine learning, 2024b.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Vladimir Vovk and Ivan Petej. Venn-abers predictors, 2014.

Vladimir Vovk, Glenn Shafer, and Ilia Nouretdinov. Self-calibrating probability forecasting. In *Advances in Neural Information Processing Systems 16 - Proceedings of the 2003 Conference, NIPS 2003*, Advances in Neural Information Processing Systems. Neural information processing systems foundation, 2004. ISBN 0262201526. 17th Annual Conference on Neural Information Processing Systems, NIPS 2003 ; Conference date: 08-12-2003 Through 13-12-2003.

Vladimir Vovk, Ivan Petej, and Valentina Fedorova. Large-scale probabilistic predictors with and without guarantees of validity. *CoRR*, abs/1511.00213, 2015.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2nd edition, 2022.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6638–6647. PMLR, 09–15 Jun 2019.

Xuekui Wang, Yong Liu, Guoyou Qin, and Yishi Yu. Robust double machine learning model with application to omics data. *BMC Bioinformatics*, 25(1):1–13, 2024.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

Lionel Weiss. A note on confidence sets for random variables. *The Annals of Mathematical Statistics*, 26(1):142–144, 1955.

Robert L Winkler. A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972.

Geoffrey R. Wood. Confidence and prediction intervals for generalised linear accident models. *Accident Analysis & Prevention*, 37(2):267–273, 2005.

Mario V. Wüthrich and Michael Merz. Editorial: Yes, we CANN! *ASTIN Bulletin - The Journal of the IAA*, 49(1):1–3, 2019.

Mario V. Wüthrich and Michael Merz. *Statistical Foundations of Actuarial Learning and its Applications*. Springer, Cham, 2023.

Mario V Wüthrich and Johanna Ziegel. Isotonic recalibration under a low signal-to-noise ratio. *arXiv preprint arXiv:2301.02692*, 2023.

Mario V. Wüthrich, Ronald Richman, Benjamin Avanzi, Mathias Lindholm, Marco Maggi, Michael Mayer, Jürg Schelldorfer, and Salvatore Scognamiglio. Ai tools for actuaries. Technical Report 5162304, SSRN, 2026.

Qingyan Xiang, Yubai Yuan, Dongyuan Song, Usman J. Wudil, Muktar H. Aliyu, C. William Wester, and Bryan E. Shepherd. Double machine learning to estimate the effects of multiple treatments and their interactions, 2025.

Gang Xu, Xian Zhou, Meng Wang, Bing Zhang, Wei Jiang, Brent Coull, Daniel Mork, and Joel Schwartz. Causal inference with double/debiased machine learning for evaluating the health effects of multiple mismeasured pollutants. *arXiv preprint arXiv:2407.06308*, 2024.

Fan Yang and Rina Foygel Barber. Contraction and uniform convergence of isotonic regression, 2019.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *ICML*, 1, 05 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 694–699, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 158113567X.

Cun-Hui Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528 – 555, 2002.

Kexin Zhao, Bo Wang, Cuiying Zhao, and Tongyao Wan. Multi-treatment-dml: Causal estimation for multi-dimensional continuous treatments with monotonicity constraints in personal loan risk optimization, 2025.

Min Zheng, Matteo Bonvini, and Zijun Guo. Perturbed double machine learning: Nonstandard inference beyond the parametric length. *arXiv preprint arXiv:2511.01222*, 2025.

Guanglin Zhou, Lina Yao, Xiwei Xu, Chen Wang, and Liming Zhu. Meta-learning for estimating multiple treatment effects with imbalance. In *Advances in Neural Information Processing Systems*, pages 886–895, 10 2023.

José R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

Appendix

A.1 Statement of Personal Contribution Pursuant to §6(4) PromO

Inference on Multiple Treatment Effects with an Application to Health Economics (Chapter 2)
Cathy Yi-Hsuan Chen, Victor Chernozhukov, Jan Rabenseifner, Martin Spindler

- Performed literature review
- Determined methodological choices
- Implemented algorithm and simulation study
- Constructed application data
- Visualized and interpreted results
- Prepared manuscript
- Presented project at the AI & Digital Assets (AIDA) Workshop, Edinburgh 2025

Calibration Strategies for Robust Causal Estimation: Theoretical and Empirical Insights on Propensity Score-Based Estimators (Chapter 3)
Sven Klaassen, Jan Rabenseifner, Philipp Bach, Jannis Kück

- Performed literature review
- Determined methodological choices
- Conceptualized statistical models
- Implemented algorithms and simulation study
- Visualized and interpreted results
- Prepared and revised manuscript
- Presented project at the Causal Data Science Meeting (online), 2025

Uncertainty Estimation in Insurance Claims Modeling: A Conformal Prediction Approach (Chapter 4)
Jan Rabenseifner, Michael Merz

- Conducted literature review
- Defined research question
- Conceptualized statistical model

- Constructed proof
- Implemented algorithms and simulation study
- Acquired application data
- Preprocessed application data
- Visualized and interpreted results
- Prepared manuscript
- Presented project at the ELLIS Doctoral Symposium, Alicante 2022

A.2 Short Summaries of Papers Pursuant to §6(6) PromO

A.2.1 Short Summaries in English Language

Inference on Multiple Treatment Effects with an Application to Health Economics (Chapter 2)

In causal inference, the focus has traditionally been on the estimation of the causal effect of a single or very few treatment variables. However, in many relevant applications, numerous treatment variables are present, and their interactions can be highly complex. An example is the study of the causes of high blood pressure. In this paper, we develop methods for the estimation and inference of the marginal treatment effect of individual variables in such complex settings using modern machine learning methods. We extend the Double Machine Learning (DML) framework of Chernozhukov et al. (2018) to accommodate multiple, simultaneously administered binary treatments. We formalize causal attribution through a regularized tempered policy that restricts estimation to the empirically supported region of the treatment space, combining tempering with truncation to address structural overlap violations. To handle the computational intractability of the resulting 2^d-dimensional integration, we propose a simulation-based debiased estimator using the Method of Simulated Scores and establish its asymptotic properties. In an extensive simulation study, the small sample properties of the estimators are compared. Despite the challenging setting, our estimator overall delivers good results. In an empirical application, we estimate the marginal effects of different lifestyle risk factors for high blood pressure, drawing from U.S. data from The National Health and Nutrition Examination Survey (NHANES), finding that obesity shows the strongest association with systolic blood pressure, followed by smoking.

Calibration Strategies for Robust Causal Estimation: Theoretical and Empirical Insights on Propensity Score-Based Estimators (Chapter 3)

The partitioning of data for estimation and calibration critically impacts the performance of propensity score based estimators such as inverse probability weighting (IPW) and double/debiased machine learning (DML) frameworks. We extend recent advances in calibration techniques for propensity score estimation, improving the robustness of propensity scores in challenging settings such as limited overlap, small sample sizes, or unbalanced data. Our contributions are twofold: First, we provide a theoretical analysis of the properties of calibrated estimators in the context of DML. We establish that isotonic calibration preserves the convergence rate and complexity conditions required by DML theory, so that calibrated estimators inherit the asymptotic properties of the uncalibrated framework under mild additional assumptions. Second, through extensive simulations across four data-generating processes, we show that calibration reduces the variance of inverse-propensity-based estimators while also mitigating bias in IPW, even in small-sample regimes. We systematically evaluate the interaction between calibration methods (isotonic regression, Platt scaling, Venn-ABERS), sample-splitting schemes (nested cross-fitting, single-split, full-sample), and base learners (logistic regression, random forest, gradient boosting). A key insight is that calibration on the full sample using cross-fitted propensity scores provides the most stable improvements, while nested cross-fitting can be unstable for small sample sizes. Notably, calibration improves stability for flexible learners while preserving the doubly robust properties of DML, and incorporating a calibration step does not degrade performance even when it is unnecessary, provided that an appropriate sample-splitting approach is chosen.

Uncertainty Estimation in Insurance Claims Modeling: A Conformal Prediction Approach (Chapter 4)

Claims frequency prediction is crucial for insurance operations; yet, standard models lack reliable uncertainty quantification. We address this by developing a conformal prediction framework for count data, providing valid prediction intervals under only the exchangeability assumption. Our contributions are threefold: First, we provide a comprehensive evaluation of conformal prediction methods for count data, tested on data-generating processes that mimic real insurance portfolios. Second, we propose DGCP Hybrid, a novel two-stage framework for zero-inflated data that combines Mondrian conformal prediction for the binary claim/no-claim decision with dynamically grouped conformal prediction for positive counts. This separation respects the distinct data-generating mechanisms for zeros and positive counts, maintaining balanced outcome-conditional coverage while producing substantially narrower intervals than parametric alternatives. Third, we introduce a Burt distance-based clustering diagnostic tool to segment portfolios and assess prediction interval performance across risk segments, identifying regions of the covariate space where prediction uncertainty is systematically elevated. In simulations, the method demonstrates robust coverage where parametric methods fail. In an application to German motor insurance data, our method corrects the severe coverage imbalances found in standard approaches, providing practical, model-agnostic uncertainty quantification for risk differentiation.

A.2.2 Kurzzusammenfassungen in Deutscher Sprache

Inferenz über multiple Behandlungseffekte mit einer Anwendung in der Gesundheitsökonomie (Kapitel 2)

In der kausalen Inferenz lag der Fokus traditionell auf der Schätzung des kausalen Effekts einer einzelnen oder weniger Behandlungsvariablen. In vielen relevanten Anwendungen sind jedoch zahlreiche Behandlungsvariablen vorhanden, deren Interaktionen hochkomplex sein können. Ein Beispiel ist die Untersuchung der Ursachen für Bluthochdruck. In diesem Beitrag entwickeln wir Methoden zur Schätzung und Inferenz marginaler Behandlungseffekte einzelner Variablen in solchen komplexen Settings unter Verwendung moderner Machine-Learning-Verfahren. Wir erweitern das Double-Machine-Learning-(DML)-Rahmenwerk von Chernozhukov et al. (2018) auf multiple, gleichzeitig verabreichte binäre Behandlungen. Dazu führen wir eine regularisierte, temperierte Policy ein, die die Schätzung auf den empirisch unterstützten Bereich des Behandlungsraums beschränkt und Glättung mit Trunkierung kombiniert, um strukturelle Overlap-Probleme zu bewältigen. Zur Bewältigung der rechnerischen Intraktabilität der resultierenden 2^d-dimensionalen Integration schlagen wir einen simulationsbasierten, bias-korrigierten Schätzer mittels der Methode der simulierten Scores vor und etablieren dessen asymptotische Eigenschaften. In einer umfangreichen Simulationsstudie vergleichen wir die Kleinstichprobeneigenschaften der Schätzer und zeigen, dass unser Ansatz trotz des anspruchsvollen Settings insgesamt gute Ergebnisse liefert. In einer empirischen Anwendung schätzen wir die marginalen Effekte verschiedener Lebensstilrisikofaktoren auf Bluthochdruck anhand von Daten des National Health and Nutrition Examination Survey (NHANES), wobei Adipositas die stärkste Assoziation mit dem systolischen Blutdruck zeigt, gefolgt vom Rauchen.

Kalibrierungsstrategien für robuste kausale Schätzung: Theoretische und empirische Einblicke in Propensity-Score-basierte Schätzer (Kapitel 3)

Die Aufteilung von Daten für Schätzung und Kalibrierung beeinflusst die Leistung von Propensity-Score-basierten Verfahren wie Inverse Probability Weighting (IPW) und Double/Debiased Machine Learning (DML) maßgeblich. Wir erweitern jüngste Fortschritte in der Kalibrierung von Propensity-Score-Schätzungen und verbessern deren Robustheit in anspruchsvollen Settings wie begrenztem Overlap, kleinen Stichproben oder unausgewogenen Daten. Unsere Beiträge sind zweifach: Erstens liefern wir eine theoretische Analyse der Eigenschaften kalibrierter Schätzer im Kontext von DML. Wir zeigen, dass die isotonische Kalibrierung die Konvergenzrate und Komplexitätsbedingungen des DML-Rahmenwerks erhält, sodass kalibrierte Schätzer dessen asymptotische Eigenschaften unter milden zusätzlichen Annahmen übernehmen. Zweitens zeigen wir in umfangreichen Simulationen über vier datengenerierende Prozesse hinweg, dass Kalibrierung die Varianz inverse-Propensity-basierter Schätzer reduziert und gleichzeitig den Bias von IPW mindert. Wir evaluieren systematisch die Interaktion zwischen Kalibrierungsmethoden (isotonische Regression, Platt Scaling, Venn-ABERS), Sample-Splitting-Schemata (verschachteltes Cross-Fitting, Single-Split, Full-Sample) und Basislernern (logistische Regression, Random Forest, Gradient Boosting). Ein zentrales Ergebnis ist, dass die Kalibrierung auf der Gesamtstichprobe unter Verwendung kreuzgefitteter Propensity-Scores die stabilsten Verbesserungen liefert, während verschachteltes Cross-Fitting bei kleinen Stichproben instabil sein kann. Insgesamt verbessert die Kalibrierung die Stabilität flexibler Lernverfahren, ohne die doppelt robuste Eigenschaft von DML zu beeinträchtigen.

Unsicherheitsschätzung in der Modellierung von Versicherungsschäden: Ein Conformal Prediction-Ansatz (Kapitel 4)

Die Prognose der Schadenshäufigkeit ist für den Versicherungsbetrieb von zentraler Bedeutung, doch Standardmodelle bieten keine verlässliche Unsicherheitsquantifizierung. Wir adressieren dieses Problem durch die Entwicklung eines Conformal-Prediction-Rahmenwerks für Zähldaten, das valide Prognoseintervalle allein unter der Annahme der Austauschbarkeit liefert. Unsere Beiträge sind dreifach: Erstens präsentieren wir eine umfassende Evaluation von Conformal Prediction-Methoden für Zähldaten, getestet an datengenerierenden Prozessen, die reale Versicherungsportfolios nachbilden. Zweitens schlagen wir DGCP Hybrid vor, ein neuartiges zweistufiges Verfahren für nullinflationierte Daten, das Mondrian Conformal Prediction für die binäre Schaden/ Kein-Schaden Entscheidung mit dynamisch gruppierter Conformal Prediction für positive Schadenszahlen kombiniert. Diese Trennung respektiert die unterschiedlichen datengenerierenden Mechanismen für Nullen und positive Werte und erreicht eine ausgewogene, ergebnisabhängige Abdeckung bei gleichzeitig deutlich schmaleren Intervallen als parametrische Alternativen. Drittens führen wir ein auf der Burt-Distanz basierendes Clustering-Diagnosetool ein, um Portfolios zu segmentieren und die Performanz der Prognoseintervalle über Risikosegmente hinweg zu bewerten, wodurch Bereiche des Kovariatenraums identifiziert werden können, in denen die Prognoseunsicherheit systematisch erhöht ist. In Simulationen zeigt die Methode robuste Abdeckung, wo parametrische Verfahren versagen. In einer Anwendung auf deutsche Kfz-Versicherungsdaten korrigiert unser Ansatz die starken Abdeckungsungleichgewichte standardmäßiger Methoden und liefert eine praktische, modellunabhängige Unsicherheitsquantifizierung für die Risikodifferenzierung.

A.3 List of Publications Pursuant to §6 (6) PromO

Journal Article	**Publication Status**
Chen, C. Y.-H., Chernozhukov, V., Rabenseifner, J. and Spindler, M. (2025). Inference on Multiple Treatment Effects with an Application to Health Economics.	Working Paper
Klaassen, S., Rabenseifner, J., Kueck, J. and Bach, P. (2025). Calibration Strategies for Robust Causal Estimation: Theoretical and Empirical Insights on Propensity Score-Based Estimators.	Under Review at Journal of Machine Learning Research (since August 2025)
Rabenseifner, J. and Merz, M. (2025). Uncertainty Estimation in Insurance Claims Modeling: A Conformal Prediction Approach.	Working Paper

A.4 Statement on the Usage of Generative Artificial Intelligence

Recent advances in large language models (LLMs) based on generative pre-trained transformers have influenced scholarly workflows worldwide, including my own. In the early stages of this dissertation, prior to the release of *ChatGPT*, I relied primarily on conventional tools for literature review (e.g., search engines such as *Google Scholar*), programming support (e.g., *Stack Overflow*), and basic translation or proofreading (e.g., *DeepL* and built-in language tools of LaTeX editors). With the increasing availability of LLMs, I gradually adopted generative AI systems (e.g., *ChatGPT*, *Gemini*, *GitHub Copilot*) for the same types of tasks in support of the research process, including literature questioning and summarization, drafting or refining code snippets, debugging, and receiving feedback on language clarity and presentation.

All generative AI outputs were treated exclusively as suggestions. They were critically assessed, verified, and, where necessary, corrected by me. This dissertation constitutes my own original scientific work: I defined the research questions, developed the methodological approaches, designed and interpreted the simulation studies and real world-applications, and wrote the manuscripts, as detailed in my contribution statement in Appendix A.1. No part of the scientific content—neither the formulation of research questions nor methodological choices, experimental results, interpretations, or manuscript writing—was delegated to generative AI.

I did not enter confidential, proprietary, or personal data into generative AI systems. AI-generated text was never used as a factual source; all scientific claims in this dissertation are supported by peer-reviewed literature or my own analyzes. Responsibility for the content, originality, and any remaining errors lies entirely with me.

Eidesstattliche Versicherung

Hiermit erkläre ich, Jan Thilo Rabenseifner, an Eides statt, dass ich die Dissertation mit dem Titel

Causal Estimation and Predictive Uncertainty: Essays on Robust Statistical Learning

selbstständig – und bei einer Zusammenarbeit mit anderen Wissenschaftler:innen – gemäß der beigefügten Darlegung nach § 6 Abs. 4, 5, 7 der geltenden Promotionsordnung der Fakultät für Betriebswirtschaft (University of Hamburg Business School) verfasst und keine anderen als die von mir angegebenen Hilfsmittel benutzt habe. Die den herangezogenen Werken wörtlich oder sinngemäß entnommenen Stellen sind als solche gekennzeichnet.

Ich versichere, dass auch im Anwendungsfall von generativer Künstlicher Intelligenz (genKI) meine eigene schöpferische Leistung der erhebliche Anteil in dieser Dissertation ist und ich die genutzte genKI in einem Anhang in meiner Dissertation aufgeführt und die Zitate in der Dissertation deutlich gekennzeichnet habe. Dieser Anhang ist Teil meiner Dissertation. Ich bin für ggfs. durch genKI generierte Inhalte, die Einhaltung urheberrechtlicher Bestimmungen, meine eigenständige Erstellung sowie für die wissenschaftliche Integrität meiner Dissertation selbst verantwortlich. Mir ist bekannt, dass fehlende oder fehlerhafte Angaben als Täuschungsversuch gewertet werden können. Ich erkläre, dass ich die Bestimmungen zum Urheberrecht und Datenschutz (DSGVO) sowie die jeweils geltenden Richtlinie der Fakultät für Betriebswirtschaft (University of Hamburg Business School) zur Anwendung von genKI-Tools erfüllt habe und erfüllen werde.

Ich versichere, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe und die Arbeit nicht schon in einem früheren Promotionsverfahren im In- oder Ausland angenommen oder als ungenügend beurteilt worden ist.

Ich versichere, dass diese Druckausführung mit meiner Originalunterschrift zu 100% der pdf-Datei entspricht, die ich zur Begutachtung eingereicht habe, und dass alle möglichen später erforderlichen Drucke dieser Dissertation ebenfalls zu 100% diesem Druckexemplar und der eingereichten pdf-Datei entsprechen.

Hamburg, den 19. Februar 2026