Improving Targeting with Privacy-Protected Data: Honest Calibration of Treatment Effects[*]

Ta-Wei Huang Eva Ascarza

August 18, 2025

Abstract

Firms increasingly rely on conditional average treatment effect (CATE) models to target interventions and optimize marketing outcomes. However, the effectiveness of these models depends on access to high-quality individual-level data, a condition increasingly challenged by privacy regulations. This paper examines how *differential privacy* (DP), a state-of-the-art approach for protecting customer data, affects CATE model performance. We show that DP introduces model-dependent and heterogeneous distortions in both bias and variance, leading to suboptimal targeting. To address this challenge, we propose an *honest model calibration* method that improves CATE estimation without requiring access to unprotected data. Rather than denoising inputs as in traditional measurement error corrections, our approach calibrates model predictions using noisy but unbiased signals derived from doubly robust scores. To prevent overfitting to noise, we implement the *honesty* principle through structured sample splitting: first between training and calibration sets, and then within the calibration process itself. We provide theoretical guarantees that this method improves accuracy when the calibration sample is sufficiently large. Empirical results from both simulations and real-world data demonstrate that our approach improves CATE accuracy and targeting performance. It is fully compatible with DP constraints and can be easily integrated into existing targeting pipelines.

Keywords: Targeted Intervention, Conditional Average Treatment Effect, Differential Privacy, Model Calibration, Honesty, Causal Machine Learning

[*] Ta-Wei Huang is an Assistant Professor of Marketing at the National University of Singapore (email: thuang@nus.edu.sg). Eva Ascarza is a Professor of Business Administration at Harvard Business School (email: eascarza@hbs.edu). The data that support the findings of this article are publicly available in `https://github.com/jroessler/autoum/tree/intmar`. The authors received no financial support for the research, authorship, or publication of this article and declare no conflicts of interest. Generative AI (ChatGPT) was used solely to correct grammar and improve writing clarity; all ideas, analyses, and content are entirely the authors' own. We are grateful to Liangzong Ma and to participants at the HBS Marketing Seminar, ACIC 2023, Marketing Science 2023, CODE@MIT 2023, Choice Symposium 2023, and EMAC 2025 for their valuable comments and feedback.

1 Introduction

Targeted interventions, such as promotional offers (Hitsch et al. 2024, Huang and Ascarza 2024), proactive churn prevention (Ascarza 2018, Yang et al. 2023), and email campaigns (Ellickson et al. 2022), have become a cornerstone of modern customer value management. These strategies aim to optimize resource allocation by estimating each customer's sensitivity to a given intervention. However, their effectiveness depends on access to high-quality individual-level data, which is increasingly constrained by growing privacy concerns and tightening regulatory oversight. Traditional safeguards such as anonymization, hashing, and opt-in consent have shown important limitations, as data protected through these methods can often be re-identified when linked with external sources (Sweeney 1997, Narayanan and Shmatikov 2008, Cohen 2022). As a result, regulators such as the European Data Protection Board (European Data Protection Board 2025) and the U.S. Federal Trade Commission (Federal Trade Commission 2024) are advocating for stronger and more enforceable data protection standards.

In this context, differential privacy (DP) (Dwork 2006) has emerged as a leading solution. DP provides rigorous guarantees that the inclusion or exclusion of any individual has minimal influence on the output of an analysis. By *injecting the right amount of random noise* into data or query results, it reduces the risk of re-identification — even in worst-case scenarios where adversaries have access to both raw and auxiliary data. In addition, DP explicitly quantifies and bounds privacy loss, offering a level of transparency and rigor that traditional techniques lack. Because of these strong guarantees, DP has been widely adopted by national statistical agencies (e.g., Kenny et al. 2021, Near et al. 2023), major technology companies (e.g., Erlingsson et al. 2014, Apple 2017), and advertising platforms (e.g., Ghazi et al. 2024, Tullii et al. 2024).

To illustrate how DP protects consumer data, consider an online retailer conducting a randomized controlled experiment to test the effectiveness of a promotional offer. To estimate how different customers respond, the firm collects individual-level data, including demographics, purchase history, and spending outcomes. While names and email addresses may be stripped from the dataset, seemingly non-identifying attributes—such as age, gender, ZIP code, and shopping preferences—can still pose privacy risks. A motivated attacker could cross-reference these fields with publicly available information, such as social media profiles, loyalty program disclosures, or product review sites. In doing so, they could re-identify specific individuals and infer sensitive behavioral details, including how much they spend and what they buy.

Under DP protection, noise is added to both demographic attributes and purchase data before any analysis is conducted, significantly reducing the risk of re-identification. For example, a customer who reports as "female" might be randomly recorded as "male" with a small probability, and a true monthly purchase amount of $100 could be reported with added noise, such as $135.37. These distortions make it far more difficult for an attacker to link internal records with external data sources, even if they gain access to the firm's raw dataset. Importantly, even if an attacker manages to re-identify an individual using publicly available demographic information, DP ensures that they cannot accurately recover that customer's purchase behavior.

Although DP is appealing from a regulatory and ethical standpoint, its very strength unfortunately poses significant challenges for personalization. By injecting substantial noise into the data, DP protects consumer privacy and ensures compliance with increasingly strict regulations; yet these protections also distort the inputs used to train predictive models, potentially degrading their accuracy. This creates a pressing question for marketers: Can targeted interventions remain effective when the underlying data is deliberately perturbed? More specifically, how does DP-induced noise — whether applied to covariates, outcomes, or both — affect a firm's ability to estimate the *conditional average treatment effect* (CATE), the incremental impact of a marketing intervention given individual customer characteristics?

The objectives of this research are twofold. First, we investigate how DP protections alter the experimental data used for CATE estimation and, in turn, how these alterations affect the statistical properties of state-of-the-art CATE models. We show that the noise intentionally introduced by DP mechanisms can distort both the bias and variance of CATE predictions. Importantly, the magnitude and direction of these distortions vary across model classes and regions of the covariate space. These patterns have critical implications for targeted marketing interventions, as they can cause firms to incorrectly identify which customers are most likely to respond to a treatment, leading to inefficient targeting and suboptimal resource allocation.

Second, we propose a novel approach to improve the accuracy of existing CATE models when estimated on DP-protected data. Although there is a substantial literature on correcting bias from covariate measurement error, these methods are poorly suited to the unique challenges of CATE estimation under DP constraints. In particular, they (i) typically do not account for noise in the outcome variable, and (ii) often require additional information that is either difficult to obtain or violates privacy, such as repeated measurements of noisy variables, valid instrumental variables for all noisy covariates, or access to clean, non-private data.

To address these limitations, we propose a novel solution, *honest model calibration*, which calibrates model predictions on an independent DP-protected dataset using a proxy for CATE that remains unbiased under differential privacy. While loosely inspired by variable calibration methods (e.g., Sepanski and Carroll 1993), which use small amounts of clean data to denoise noisy variables, our method is fundamentally different in both purpose and design. Rather than denoising covariates or outcomes with external non-private data, it is tailored specifically for DP settings where no clean data are available and operates at a different stage of the modeling pipeline. By using a dataset independent from that used for model estimation, our approach prevents DP noise from compounding across stages, ensuring valid calibration. Importantly, it also preserves DP protection by the well-known post-processing property of DP (Dwork et al. 2014).

To construct the unbiased proxy for model calibration, we leverage the doubly robust (DR) score (Kennedy 2023), which combines predictions from outcome models with treatment assignment models to approximate individual treatment effects. A key advantage of the DR score is that, as long as the treatment assignment mechanism is correctly specified, either through complete randomization or a known assignment rule based on DP-protected covariates, it provides an unbiased proxy for the true CATE. This property makes the DR score a reliable target for calibrating model predictions.

However, the noise introduced by DP creates a fundamental challenge when using DR scores as a proxy: although remaining unbiased, the DR score becomes highly noisy under DP protection. Directly aligning (or calibrating) model predictions with such a noisy proxy on the same dataset risks overfitting to noise, ultimately degrading predictive accuracy. To address this, we propose an *honesty principle*, where the DP-protected data is partitioned into separate subsets for training, calibration, and validation. This separation ensures independence between model errors and proxy errors, reducing the likelihood of overfitting to DR score noise. In addition, we develop a new variant of the gradient boosting algorithm to implement honest model calibration, iteratively refining predictions by minimizing residuals with respect to the DR score. This variant incorporates a step-size determination strategy that explicitly separates calibration model construction from step-size optimization, ensuring that calibration leads to improvements in overall accuracy.

We provide formal guarantees for the proposed algorithm, showing that under the honesty condition, model calibration improves prediction accuracy when the calibration sample is sufficiently large. We then validate our method through extensive analyses on both simulated and real-world datasets, confirming its effectiveness under DP protection. Whether outcomes, covariates, or both are protected, our approach substantially improves CATE estimation accuracy, treatment prioritization, and targeting value — allowing firms to allocate resources more efficiently while maintaining DP protection.

Beyond accuracy gains and privacy preservation, our method offers several practical advantages for marketing practitioners. First, it is both model-agnostic and noise-agnostic: it improves accuracy without requiring customization for specific CATE models or noise injection mechanisms, making it broadly applicable across different settings. Second, it is cost-efficient, delivering reliable improvements even with modest experimental sample sizes. Third, it is easy to implement, computationally efficient[1], and integrates seamlessly into existing CATE prediction and personalization workflows, enabling scalable deployment in real-world applications.

Our contributions are twofold. Managerially, we examine how the growing adoption of DP protections affects firms' ability to perform targeted interventions. As regulators and digital platforms enforce stricter privacy standards, firms increasingly receive outcomes and covariates in noise-injected form. We show that this privacy-preserving mechanism could biases CATE estimates, increases their variance, and reduces the accuracy and stability of targeting decisions. While prior marketing research has highlighted the value of CATE estimation for personalization, most existing methods assume access to clean, high-fidelity data, a condition no longer guaranteed under DP protection.

Methodologically, we introduce a new model calibration approach to improve CATE estimation when the data are protected by DP. Our method departs from traditional variable calibration techniques used to correct measurement error bias, which rely on clean data to denoise variables and are incompatible with DP protection. Instead, we calibrate model predictions directly using an unbiased proxy for the true treatment effect. To guard against overfitting to privacy-induced noise, we employ a principled sample-splitting procedure that separates model training from calibration. The resulting method is fully privacy-preserving, requires no knowledge of the noise distribution, and is compatible with a wide range of CATE models, making it well-suited for real-world deployment.

This paper is structured as follows. Section 2 outlines the connections to existing literature. Section 3 characterizes how the implementation of differential privacy alters experimental data and, in turn, affects the performance of popular CATE models. Section 4 presents our proposed solution and discusses its key advantages. We assess the empirical performance of the method using simulation studies in Section 5 and two real-world datasets in Section 6. Finally, Section 7 concludes with implications for practice and directions for future research.

[1] In both our simulation and empirical applications, the proposed calibration procedure increases runtime by only 10% compared to the default approach without calibration.

2 Related Literature

Our research connects to multiple streams of literature across marketing, statistics, economics, and computer science.

2.1 Targeted Interventions and CATE Estimation

Our research contributes to the growing literature on targeted interventions via CATE estimation (Lemmens et al. 2025). Existing methods can be broadly classified into three categories (Hitsch et al. 2024): (i) Direct methods (e.g., Causal Forest; Wager and Athey 2018) that optimize for treatment effect heterogeneity directly; (ii) Transformed outcome regression methods (e.g., Nie and Wager 2021, Semenova and Chernozhukov 2021, Kennedy 2023), which estimate a proxy for the true CATE and regress it on covariates; and (iii) Indirect methods (e.g., T-learner; Künzel et al. 2019) that estimate potential outcomes separately and take their difference. Our work extends this literature by proposing a model calibration framework that improves accuracy for all three types, especially when the data is noisy.

CATE models are increasingly applied in marketing settings such as customer retention (Ascarza 2018), subscription (Simester et al. 2020), promotions (Smith et al. 2021, Huang et al. 2024), and catalog mailing (Hitsch et al. 2024). Recent work highlights challenges these models face in noisy environments. For instance, Huang and Ascarza (2024) show that outcome noise inflates the variance of CATE estimates and suggest using low-variance signals as a remedy. We extend this line of work by (1) extending the analysis to settings where both outcomes and covariates are noisy, and (2) proposing a new approach to improve CATE prediction that does not rely on low-variance proxies.

2.2 Differential Privacy

DP has received growing attention in both academic and applied settings due to its promise in protecting individual-level data. Most existing research has focused on developing new privacy mechanisms and establishing their theoretical guarantees (e.g., Kasiviswanathan et al.

2011, Erlingsson et al. 2014, Ding et al. 2017). Other work has studied the trade-off between privacy and accuracy, either from an information-theoretic perspective (Sarwate and Sankar 2014, Kalantari et al. 2018, Zhong and Bu 2022) or within a statistical framework (Showkatbakhsh et al. 2018, Amin et al. 2019). More recently, studies such as Niu et al. (2022) and Ponte et al. (2025) have examined differentially private CATE estimation and targeting, focusing on scenarios where either the CATE model or the targeting decisions must satisfy DP, while the underlying data remains unperturbed.

Despite these advances, little research has examined how DP-protected data — where noise is added at the source before CATE models are estimated — affects the accuracy of CATE prediction, or explored methods to address these challenges. This is an increasingly important issue for marketers, given the adoption of DP in many real-world platforms and data sources (e.g., Apple, Google, and the U.S. Census Bureau). Our study fills this gap by conducting one of the first investigations into the effects of DP-data on CATE estimation and targeting. In addition to documenting the privacy–accuracy trade-off, we introduce and evaluate a novel calibration method designed to improve CATE accuracy when working with DP-protected data. Our approach does not require access to additional non-private data and remains flexible across model classes and DP mechanisms, making it well-suited for practical deployment.

2.3 Classical Measurement Error

Our research contributes to the literature on classical measurement errors, which arise in our context when decision-makers introduce noise into individual data for privacy protection. Prior studies (e.g., Chesher 1991, Battistin and Chesher 2014, Abel 2018) demonstrate that such errors can bias regression coefficients and average treatment effect estimates in observational settings. We extend this line of research by conducting a comprehensive theoretical and empirical analysis of how measurement errors affect both the bias and variance of CATE predictions when using flexible machine learning models.

Existing bias correction methods for classical measurement errors typically follow four approaches: utilizing repeated measurements of variables (e.g., Hausman et al. 1991, Li and

Vuong 1998, Schennach 2004, Agarwal and Singh 2021), variable calibration through additional clean data (e.g., Sepanski and Carroll 1993, Lee and Sepanski 1995, Chen et al. 2005, Hu and Ridder 2012), applying instrumental variables (e.g., Hausman et al. 1991, Schennach 2007, Yang et al. 2022), and leveraging distribution information (e.g., Pal 1980, Wolter and Fuller 1982, Carroll et al. 1999, Schennach and Hu 2013). However, these methods face significant limitations when applied to correcting errors in CATE predictions under DP protection. Our research addresses these challenges by proposing a novel formulation: treating measurement error correction as a *model-level* correction problem rather than a variable-level adjustment. Importantly, our approach requires neither clean data, instrumental variables, nor repeated measurements, making it broadly applicable in real-world settings.

2.4 Model Calibration

Methodologically, our research builds on the emerging literature on model calibration — a post-processing approach that adjusts predictions to better match observed outcomes (Lichtenstein et al. 1977, Platt et al. 1999). Recent work extends this idea to CATE models, using calibration to assess alignment with empirical treatment effects (Leng and Dimmery 2024, Whitehouse et al. 2024) and improve generalizability across populations (Kern et al. 2024).

We contribute to this literature in three ways. First, we extend model calibration to address high-noise settings, whereas most prior work has focused on goals such as fairness (Hébert-Johnson et al. 2018), covariate shift (Kim et al. 2022), or multi-objective prediction (Gopalan et al. 2021). Second, while existing calibration methods aim to align predictions with difference-in-means estimates (e.g., Chernozhukov et al. 2018, Leng and Dimmery 2024), our approach goes further by also improving the ranking of treatment effects across individuals — an essential aspect for targeting decisions in marketing.

Third, unlike traditional subgroup-based calibration approaches that adjust predictions within each group individually (e.g., Whitehouse et al. 2024), our method leverages information across subgroups to address overfitting. Specifically, we determine adjustment magnitudes using data held out from the focal subgroup, drawing on the honesty principle from the

CATE estimation literature (Athey and Imbens 2016) to improve generalization and prevent overfitting . Fourth, we show that separating the data used for initial model training from the data used for calibration is essential. We provide formal theoretical guarantees that honest calibration improves accuracy asymptotically. While prior work on calibrating CATE models (e.g., Whitehouse et al. 2024) has emphasized the importance of sample splitting for estimating nuisance components such as propensity scores or outcome models, we extend this insight by showing that sample splitting is also critical during the calibration stage itself.

3 Problem: Impact of DP Protection on CATE Prediction

In this section, we illustrate how companies apply differential privacy in their data collection, examine how this integration alters experimental data, and discuss the resulting implications for CATE estimation and targeting.

3.1 Preliminary: Differential Privacy

Differential privacy is a rigorous framework for protecting individual-level data while still allowing meaningful statistical analysis. Its core principle is that the inclusion or exclusion of any single customer should have only a negligible impact on the results. To achieve this, DP introduces carefully designed random noise into the data or the outputs of queries, limiting what can be inferred about any individual.

There are two major frameworks for implementing DP: **central differential privacy (CDP)** and **local differential privacy (LDP)**. They differ primarily in two ways: (i) the level of trust required in the data collector and (ii) the stage of the data pipeline at which noise is introduced.

Central Differential Privacy (CDP). The CDP framework assumes that the firm is a trusted entity capable of securely collecting and storing raw, customer-level data in its internal systems. Under this model, differential privacy protections are applied not at the point of data collection but during data access and analysis. Specifically, when analysts (whether internal teams or authorized external collaborators) query the data (e.g., to compute descriptive statistics or estimate model parameters), random noise is intentionally added to the query output. This mechanism ensures that individual-level information is obscured in the released results, even though the underlying data remain unchanged. CDP is especially valuable in settings where (i) there is a high risk of internal misuse of outputs, or (ii) query results are stored or shared in environments that are less secure than the raw data repository.

Local Differential Privacy (LDP). The LDP framework assumes that the firm (or entity) storing the data cannot be fully trusted. As a result, noise is added to the data *before* it is transmitted to the firm. Specifically, calibrated random noise is added locally, on the customer's device or browser, ensuring that the true customer data are never observed by the platform. This approach provides strong privacy guarantees by design and is particularly well-suited for settings where the data collector is untrusted, or where the data being collected are highly sensitive and require strict privacy protection.

In practice, firms may adopt a hybrid approach to implementing DP protection, combining both local and central mechanisms depending on the type and sensitivity of the data. For example, consider a large e-commerce platform. Demographic variables such as gender, age, and income (which can be used to re-identify individuals when combined with external sources like social media) are privatized directly on the user's device before being transmitted. This privacy protection is achieved through LDP mechanisms, which ensure that the raw values are never seen by the platform. In contrast, transaction data, essential for operational processes and often serving as the ground truth for financial reporting, must be collected in their original, unperturbed form. However, when internal analysts derive individual-level features for targeting (e.g., RFM metrics), the company can implement CDP by adding calibrated noise to the query results before returning them to the analyst. This approach reduces internal privacy risks and helps prevent potential reidentification using the query outputs.

3.2 Implication for Targeted Interventions

Next, we examine how the implementation of DP protection affects a firm's ability to execute targeted marketing interventions. In such settings, firms aim to deliver interventions such as promotions or marketing messages only to customers who are expected to derive a positive incremental benefit (Lemmens et al. 2025). To identify these high-impact individuals, firms typically run randomized controlled experiments.n. The experimental data are then used to estimate CATEs, which capture the expected difference in outcomes between treated and untreated customers, conditional on their observed covariates. Customers with the highest predicted CATEs are prioritized for targeting, allowing firms to allocate resources more efficiently and maximize the overall impact of the intervention (e.g., Yoganarasimhan et al. 2022).

Within this paradigm, we first discuss how the implementation of DP protection perturbs the underlying experimental data. We then analyze how such noise injections affect the statistical properties of commonly used CATE models, focusing on the trade-offs between privacy protection, bias, and estimation variance.

3.2.1 Experimental Data

Typically, a firm collects three key pieces of data for each customer i participating in a randomized controlled experiment for designing targeted interventions:

- **Treatment condition** ($W_i \in \{0, 1\}$): whether the customer was exposed to the marketing intervention (e.g., receiving an email campaign or not).
- **Covariates** ($\mathbf{X}_i$): a set of pre-treatment characteristics used to model heterogeneity, such as demographic information or behavioral history (e.g., purchase frequency and recency). Note that firms may obtain these covariates from third-party providers that apply differential privacy protections—for example, geo-location and demographic information from the U.S. Census Bureau (Kenny et al. 2021).
- **Outcome variable** (Y_i): the observed response of interest, such as whether the customer clicked, converted, or how much they spent.

The firm's objective is to use this data to estimate a CATE model, which predicts the individualized impact of the intervention based on covariates. Formally, the CATE for customer i is defined as $\tau(\mathbf{X}_i) \equiv \mathbb{E}[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]$, where $Y_i(W_i)$ is the potential outcome (Rubin 1974) of individual i's response given the treatment condition W_i. Among these variables, the treatment allocation W_i is generally not considered sensitive from a privacy perspective. It is typically randomized and, on its own, does not uniquely identify any customer. In contrast, demographic information (such as age, gender, and income) is inherently sensitive and can often be cross-referenced with publicly available data sources, such as social media profiles, to reveal customers' true identity. As a result, differential privacy protections are typically applied to safeguard this information.

Beyond demographics, both the outcome variable Y_i and behavioral covariates (such as past purchase history) can also reveal highly sensitive information and thus require privacy protections. For example, consider a typical outcome in a promotional experiment: customer spending over a specific month. If a malicious actor gains access to raw spending data, they could cross-reference it with external sources, such as third-party credit card panels, to re-identify individuals — even when the data are anonymized. Because spending patterns (e.g., transaction timing, amounts, and merchant) are often uniquely identifying, an attacker could successfully match records to specific individuals. Once such linkage is established, additional private attributes (e.g., location, income bracket, or spending outside the platform) could also be inferred or directly revealed.

3.2.2 Data Distortion under Differential Privacy

When constructing a CATE model, DP protection implies that the analyst only has access to a noise-injected version of the experimental data, $\widetilde{\mathcal{D}} = \{(\widetilde{Y}_i, \widetilde{\mathbf{X}}_i, W_i)\}_{i=1}^N$, rather than the original dataset, $\mathcal{D} = \{(Y_i, \mathbf{X}_i, W_i)\}_{i=1}^N$. Specifically, instead of observing the true outcome Y_i and covariates $\mathbf{X}_i$, the analyst only has access to their privatized versions, $\widetilde{Y}_i$ and $\widetilde{X}_i$, when estimating CATEs using data from randomized experiments. We model DP protection as additive noise: $\widetilde{Y}_i = Y_i + \eta_i^Y$ and $\widetilde{\mathbf{X}}_i = \mathbf{X}_i + \boldsymbol{\eta}_i^X$. This additive noise formulation describes a broad class of

DP mechanisms. For numerical variables, this includes the Laplace mechanism (Dwork et al. 2006b, Kasiviswanathan et al. 2011), Gaussian mechanism (Dwork et al. 2006a), and geometric mechanism (Ghosh et al. 2009), where a numeric value is perturbed by adding random noise. For example, when an analyst queries a customer's 30-day spending, the DP-protected database may return a noisy value such as $135.37 instead of the true amount of $100, due to the addition of Laplace noise. For categorical variables, the framework also applies to the randomized response mechanism (Warner 1965, Erlingsson et al. 2014), in which the true value is randomly flipped according to a known probability distribution. For instance, a customer whose true gender is "female" might be reported as "male" with a certain probability.[2]

3.2.3 Impact on CATE Prediction

Conceptually, the additive noise introduced by DP mechanisms is related to the notion of *measurement error*, wherein observed values deviate from their true counterparts due to random perturbations. In linear regression, it is well established that measurement error in covariates leads to attenuation bias, shrinking estimated coefficients toward zero, whereas noise in the outcome variable does not introduce bias if the model is correctly specified and the error is additive and independent of the regressors. While DP-induced noise is related to this classical notion of measurement error, its implications for CATE estimation are considerably more complex, as we discuss next.

First, CATE estimation typically results in two implied outcome models, one for the treated group and one for the control group, whose difference defines the treatment effect for each individual (Huang and Ascarza 2024). If DP-induced measurement error introduces differential attenuation across these two models, the resulting CATE prediction may be biased in ways not captured by classical uniform attenuation effects. Second, modern CATE models increasingly rely on machine learning methods that incorporate nonlinearity and regularization (Wager and Athey 2018, Nie and Wager 2021, Farrell et al. 2021, Hitsch et al. 2024). In these

[2] The randomized response technique can be reformulated as the injection of additive noise. Consider a binary variable X. The randomized response with flipping probability p can be expressed as $\widetilde{X} = X + \mathbb{1}(X = 1)\eta_1 + \mathbb{1}(X = 0)\eta_0$, where $-\eta_1 \sim$ Bernoulli(f) and $\eta_0 \sim$ Bernoulli(p). A similar formulation can be applied to the dummy transformation of a discrete variable with multiple possible values.

settings, outcome noise, often dismissed as a source of bias in classic regression, can interact with regularization in ways that introduce bias into model predictions. Importantly, this bias is not uniform: its size and direction vary across model classes, making it difficult to address without understanding how noise propagates through complex learning algorithms.

Third, while classical measurement error literature emphasizes bias, variance plays an equally critical role in targeting decisions (Huang and Ascarza 2024, Fernández-Loría and Loría 2025). When covariates or outcomes are perturbed by DP-induced noise, the resulting increase in the variance of CATE estimates can substantially reduce targeting precision, even in the absence of bias. For firms aiming to deploy targeted interventions, this loss in precision directly undermines the value of targeting. As such, the impact of DP on model variance is just as consequential as its effect on bias.

To systematically examine these issues, we analyze how DP noise affects the bias and variance of standard CATE estimators. This analysis highlights the specific challenges introduced by DP and motivates the need for new methods to address them. Full results are provided in Web Appendix A, where we use both analytical approximations and simulations to study the behavior of popular CATE models under DP-induced noise. Below, we summarize the key patterns.

First, the results show that DP-induced distortions are highly model-dependent: different CATE estimators respond differently to the same level and type of DP noise. This has important implications: correction methods tailored to a specific estimator may work well in one context but fail in another, especially when a different CATE model is better suited to the data — a situation that is common in real-world applications (Rößler and Schoder 2022). Second, the magnitude and direction of these distortions vary across the covariate space, so a correction method that improves estimates in one region may inadvertently worsen them in another. Ideally, solutions should adaptively correct prediction errors at the individual customer level. However, such adaptive correction can itself lead to overfitting the DP noise. Finally, when DP is applied to outcomes, high privacy levels can induce either strong regularization bias or substantially higher variance, depending on the estimator's structure. Thus, an effective correction method must address distortions arising not only from covariate noise but also from outcome noise. Together, these results highlight the need for a new correction approach. In the next section, we present a method that improves CATE estimation and targeting by overcoming these three challenges.

4 Solution: Honest Model Calibration

4.1 Criteria for an Effective Solution

To address the challenge of CATE estimation with DP-protected experimental data, we outline three criteria for an effective solution. These criteria are not intended to be jointly sufficient for optimality; rather, they reflect important practical considerations that arise when developing targeted interventions with DP-protected data.

1. **Preserving Privacy Protection**: The method must not require access to clean, unperturbed data or attempt to reverse-engineer it. This ensures compliance with formal DP guarantees and avoids re-identification risks.

2. **Addressing Outcome Noise**: While much of the classical measurement error literature focuses on covariate noise, distortions from outcome noise are equally important. In high-privacy settings, noise injected into outcomes can create substantial regularization bias and inflate the variance of CATE models, both of which must be addressed to ensure effective targeting.

3. **Avoiding Challenging Data Collection**: To be practical and widely applicable, the solution should work with (DP-protected) data that are readily available in typical applications. Approaches that rely on auxiliary information such as noise-free data, external instruments, or repeated measurements may offer theoretical advantages but are often difficult to implement in practice.

Table 1 compares three widely used approaches from the covariate measurement error literature: repeated measurements, instrumental variables (IV), and variable calibration, against the key criteria for privacy-preserved data.[3] While each method has distinct strengths, they all face significant practical limitations when applied in the context of differential privacy protection. We describe these next and summarize them in Table 1.

First, *repeated measurement approaches* rely on obtaining multiple independent realizations of the same variable, which is often infeasible in applied settings. For instance, firms cannot repeatedly request sensitive information such as age or gender from users. Furthermore, collecting additional measurements may conflict with privacy protection goals, particularly when the purpose is to denoise the very variables subject to DP protection. Second, *IV methods* are theoretically compatible with DP and can address covariate measurement error. However, these methods do not address the problem of outcome noise. They also rely on valid instruments: auxiliary variables that are correlated with the true covariate but uncorrelated with the noise and outcome. Identifying such instruments is particularly challenging for demographic attributes, where no natural proxies may exist. Finally, variable calibration methods require access to a clean, non-privacy-protected dataset, which fundamentally conflicts with the protection of differential privacy.

These limitations motivate the development of a new solution that preserves DP protection, accounts for outcome noise, and avoids the need for additional data collection. While our approach is inspired by variable calibration methods, it departs from them in a fundamental way: rather than correcting the input data, we focus on post-estimation calibration — refining model predictions without altering the original DP-protected inputs. This shift not only ensures compliance with DP but also provides a framework for improving CATE accuracy in privacy-sensitive environments.

[3] There are two other potential solutions in the literature: deconvolution methods and moment calibration techniques. However, these approaches typically rely on strong assumptions about the noise distribution (Schennach 2022), are applicable only to a narrow range of model specifications, and do not address noise in the outcome variable. As such, they are not well-suited to our setting. For a detailed review, please refer to Web Appendix B.1.

Table 1: Evaluation of Classical Methods Against Critical Considerations

Solution	Preserve Privacy Protection	Address Outcome Noise	Does Not Require Difficult-to-Collect Data	Related Literature
Repeated Measurements[1]	✗	✓	✗	Hausman et al. (1991), Li and Vuong (1998), Schennach (2004), Agarwal and Singh (2021)
Instrumental Variables[2]	✓	✗	✗	Hausman et al. (1991), Schennach (2007), Hu and Schennach (2008), Yang et al. (2022)
Variable Calibration[3]	✗	✓	✗	Sepanski and Carroll (1993), Lee and Sepanski (1995), Chen et al. (2005), Hu and Ridder (2012)

Note: ✓= satisfies the criterion; ✗= does not satisfy the criterion.

[1] Repeated measurement approaches aim to recover the true value of a covariate by averaging across multiple independent noisy measurements. However, this strategy raises privacy concerns as collecting the same sensitive attribute multiple times increases the risk of re-identification and may breach DP protection. In addition, this approach is often impractical. For example, it might not possible to repeatedly ask customers for attributes such as gender or age.

[2] IV methods address measurement error bias by using auxiliary variables (instruments) that are correlated with the noisy covariate but uncorrelated with both the measurement noise and the outcome. While DP can be maintained by adding noise to the instruments, this often causes the weak instrument problem (Andrews et al. 2019). Besides, IV methods are not designed to correct outcome noise, and it is often difficult to find valid instruments for variables such as age or gender. Finally, while nonparametric IV models offer model flexibility, adapting them for covariate measurement error correction requires significant development.

[3] Variable calibration methods use a small, separate validation dataset with clean measurements to correct for noise in a larger, noisy dataset, and then use the denoised inputs to estimate the regression model. These methods are appealing because they can handle both covariate and outcome noise, do not require knowledge of the noise distribution, and are model-agnostic. However, they fundamentally violate privacy guarantees, as they depend on access to noise-free (non-DP-protected) data.

4.2 Honest Model Calibration

4.2.1 Solution Concept

Building on the idea of calibration, we introduce the honest model calibration approach, which operates entirely on DP-protected data and improves the accuracy of CATE models without reconstructing noise-free inputs. Traditional variable calibration methods denoise covariates or outcomes before model training (*pre-processing*). In contrast, our method focuses on *post-processing* by calibrating the predictions of a CATE model to improve accuracy. Specifically, we first train an initial CATE model using one privatized experimental dataset and then calibrate its predictions using an *independent* DP-protected dataset obtained via sample splitting. As discussed in Section 3.2.3, the initial estimates can suffer from substantial bias and variance. The goal of the calibration step is to align the model's predictions with an unbiased proxy of the true CATE using independent data. This alignment yields more accurate CATE estimates and, in turn, more reliable targeting decisions.

Importantly, this approach satisfies the three key criteria for an effective solution (Table 1). First, it is fully privacy-preserving. By the post-processing property of differential privacy, any transformation of DP-protected data remains differentially private (Dwork et al. 2014). Second, it addresses both outcome and covariate noise by directly correcting the estimation errors of the initial CATE model, without requiring assumptions about the source of error. Finally, it avoids the need for additional data collection, such as noise-free inputs, repeated measurements, or external instrumental variables.

Beyond the three core criteria discussed above, our proposed solution offers two additional practical advantages. First, it is *model-agnostic*. CATE estimation often relies on diverse machine learning models (e.g., tree-based algorithms or neural networks) to capture complex treatment effect heterogeneity. Model-specific corrections may work when the model class is fixed, but DP noise interacts with model structures in complex and often unpredictable ways (see Web Appendix A). As a result, corrections tailored to one model may not generalize well to others. A model-agnostic approach avoids this limitation, providing greater robustness and flexibility. This is especially valuable in real-world applications, where different teams or business units may use different models, model specifications often change over time, and practitioners usually choose algorithms based on what works best empirically rather than on strict assumptions about how the data are generated.

Second, it is *noise-agnostic* because it does not depend on the specific noise distribution. Although DP noise is often modeled as additive, real-world implementations frequently use more complex distributions, such as truncated geometric noise, randomized response, or discrete Gaussian noise. Moreover, the privacy parameters and noise distributions are often unknown to downstream analysts. For instance, advertisers may receive privatized data from a platform without being told the underlying DP mechanism, or internal analysts may lack visibility into how DP is implemented within their organization. In such cases, a flexible solution that avoids detailed knowledge or strict assumptions about the noise distribution is more broadly applicable.

4.2.2 Key Steps for Honest Model Calibration

To effectively calibrate initial CATE models that may suffer from high error, two key steps are required. The first is to construct a valid calibration target: an unbiased proxy for the true CATE that can be estimated using only DP-protected data. For this, we leverage the DR score, which yields an unbiased proxy of the CATE when either the conditional outcome models or the propensity score model is correctly specified (Funk et al. 2011). Although the DR score has been widely applied in CATE estimation (e.g., Semenova and Chernozhukov 2021, Kennedy 2023), we extend its role by repurposing it as a privacy-compatible target for model calibration.

The second step is to ensure that aligning the model's initial CATE predictions with this noisy but unbiased target meaningfully improves predictive accuracy. To accomplish this, we adopt the *honest* principle: splitting the privacy-protected dataset into two disjoint subsets. The initial CATE model is trained on one subset (training set) and calibrated on the other (calibration set). This design reduces estimation error because the DP noise injected into each customer's data is independent across splits. As a result, the calibration set can effectively correct prediction errors from the training set, rather than reinforcing the same noise patterns. We now formalize this framework and describe each key step in detail.

Calibration Target. When calibrating a model, the goal is to reduce the gap between its predictions and the true outcome of interest. In the case of CATE estimation, a key challenge is that individual treatment effects are never directly observed (Rubin 1974). This makes it essential to construct a reliable proxy for calibration. Following the literature on CATE estimation (Semenova and Chernozhukov 2021, Kennedy 2023), we adopt the DR score (Robins et al. 1994) as our proxy for the true CATE:

$$\check{\tau}_i = \mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i) + \frac{W_i}{e(\mathbf{X}_i)}\left[Y_i - \mu_1(\mathbf{X}_i)\right] - \frac{1 - W_i}{1 - e(\mathbf{X}_i)}\left[Y_i - \mu_0(\mathbf{X}_i)\right],$$

where $\mu_w(\mathbf{X}_i) = \mathbb{E}[Y_i \mid W_i = w, \mathbf{X}_i]$ denotes the conditional mean outcome under treatment condition $w \in \{0, 1\}$, and $e(\mathbf{X}_i) = \mathbb{P}[W_i = 1 \mid \mathbf{X}_i]$ is the propensity score given covariates $\mathbf{X}_i$. In practice, the conditional mean outcome models can be estimated using flexible machine learning algorithms combined with cross-fitting techniques (Semenova and Chernozhukov 2021).

The DR score is a particularly useful proxy because of its double robustness: it yields an unbiased estimate of the true CATE as long as either the outcome model or the propensity score model is correctly specified. If decision-makers have access to the true propensity score, this property ensures that $\check{\tau}_i$ remains an unbiased proxy, even when the outcome models are learned using data protected under DP, which may introduce bias. The following proposition formalizes that the DR score remains an unbiased estimator of the true CATE under DP protection, provided it is computed using the true propensity score:

Proposition 1 (Unbiased Signal)

Under standard assumptions of positivity, unconfoundedness, and no interference, the doubly robust score is an unbiased proxy for the true CATE when the additive noise introduced by the DP mechanism has mean zero. That is, $\mathbb{E}\left[\check{\tau}_i \mid \mathbf{X}_i\right] = \tau(\mathbf{X}_i)$, as long as the score is computed using the true propensity score, even if both the covariates and outcomes are DP-protected.

The proof is provided in Web Appendix C.1. Although assuming knowledge of the true propensity score may seem strong, it is practical in settings with DP protection. The assumption holds trivially in randomized controlled trials, where treatment is assigned independently of covariates. In cases of non-random treatment assignment, it remains valid if the assignment policy is known and based on privatized covariates. In such settings, even though the true covariates are not observed, the firm can still compute the true propensity score from the DP-protected data. This makes the use of AIPW both feasible and unbiased under DP constraints.

Effective Model Calibration with Honest Samples. Next, we turn to the task of calibrating a CATE model $\hat{\tau}$ using the DR score. The goal is to refine the model by aligning its predictions with the DR score $\check{\tau}_i$, typically by minimizing the mean squared residuals $(\hat{\tau}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \check{\tau}_i)^2$. At first glance, this objective resembles a boosting-based variant of the DR-learner (Kennedy 2023), where the DR score is treated as a pseudo-outcome for CATE and the model is iteratively updated to predict it. While the boosting-based DR-learner may perform well in settings without DP protection, its performance can deteriorate under DP because the boosting procedure is prone to overfitting the noise.

Specifically, when both the initial CATE model $\widehat{\tau}$ and the DR score $\check{\tau}_i$ are computed from the same dataset, the DP noise injected into the data affects both, creating a positive correlation between the model's prediction error and the proxy error in the DR score. In practice, this means that if noise causes the model to systematically over- or under-estimate treatment effects for certain observations, the DR score constructed from the same noisy data will tend to exhibit a similar error in the same direction, often with even larger magnitude. In such cases, aligning the model's predictions to the DR score reinforces the error instead of correcting it.

To illustrate this problem formally, suppose we evaluate the residual between a trained CATE model and the DR score on the same dataset $\widetilde{\mathcal{D}}$ used for model training. The expected squared difference can be decomposed as:

$$
\begin{aligned}
\mathbb{E}_{\widetilde{\mathcal{D}}}\left[(\widehat{\tau}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \check{\tau}_i)^2\right] &= \mathbb{E}_{\widetilde{\mathcal{D}}}[(\widehat{\tau}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_i) - (\underbrace{\check{\tau}_i - \tau(\mathbf{X}_i)}_{\equiv \check{\epsilon}_i})^2] \\
&= \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}}\left[(\widehat{\tau}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_i))^2\right]}_{\text{True expected squared prediction error}} \; - \; 2 \cdot \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}}\left[(\widehat{\tau}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_i)) \cdot \check{\epsilon}_i\right]}_{\text{Comovement w.r.t. the proxy error}} \; + \; \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}}\left[\check{\epsilon}_i^2\right]}_{\text{Variance}},
\end{aligned}
\tag{1}
$$

Here, $\check{\epsilon}_i$ denotes the approximation error of the DR score relative to the true (unobserved) treatment effect. The first term reflects the actual squared prediction error of the model—what we ideally want to minimize. The third term is the variance of the proxy target, which is fixed for any calibration procedure.

The key concern lies in the second term of the decomposition. When calibration is performed on the same dataset used to train the initial model, minimizing the mean squared residuals between the predicted CATE and the DR score can inflate the comovement term. In this case, the model may align with the idiosyncratic noise in the DR score rather than reduce true prediction error. Consequently, minimizing the squared residual between $\widehat{\tau}$ and $\check{\tau}_i$

does not necessarily improve accuracy and may even amplify error. This is a classic overfitting problem, where the model learns noisy patterns instead of improving accuracy.

We address this challenge by proposing an *honesty* principle for model calibration under differential privacy. Specifically, we partition the DP-protected dataset into two disjoint subsets: one for training the initial CATE model, and the other for calculating DR scores and calibrating the CATE model. We refer to this strategy as *honest* because it follows the same spirit as the honesty principle in Causal Trees (Athey and Imbens 2016), where separate samples are used for model fitting and prediction to prevent overfitting.

To understand how honesty helps mitigate the overfitting problem characterized in (1), let us consider two *disjoint* datasets: $\widetilde{\mathcal{D}}_{\text{train}}$, used for constructing the CATE model $\widehat{\tau}(\cdot)$, and $\widetilde{\mathcal{D}}_{\text{cal}}$, reserved for calibrating $\widehat{\tau}$ by aligning its predictions with the DR score of individuals in $\widetilde{\mathcal{D}}_{\text{cal}}$. By ensuring statistically independence between $\widetilde{\mathcal{D}}_{\text{train}}$ and $\widetilde{\mathcal{D}}_{\text{cal}}$, we can express the expected squared proxy residual for individual j in $\widetilde{\mathcal{D}}_{\text{cal}}$ as follows:

$$\mathbb{E}_{\widetilde{\mathcal{D}}_{\text{train}},\widetilde{\mathcal{D}}_{\text{cal}}}\left[(\widehat{\tau}(\widetilde{\mathbf{X}}_j)-\check{\tau}_j)^2\right] = \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}_{\text{train}},\widetilde{\mathcal{D}}_{\text{cal}}}\left[(\widehat{\tau}(\widetilde{\mathbf{X}}_j)-\tau(\mathbf{X}_j))^2\right]}_{\text{True prediction error}} + \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}_{\text{cal}}}\left[\check{\epsilon}_j^2\right]}_{\text{Constant}} - 2\underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}_{\text{train}}}\left[\widehat{\tau}(\widetilde{\mathbf{X}}_j)-\tau(\mathbf{X}_j)\right]}_{\text{Bias of }\widehat{\tau}(\widetilde{\mathbf{X}}_j)} \cdot \underbrace{\mathbb{E}_{\widetilde{\mathcal{D}}_{\text{cal}}}\left[\check{\epsilon}_j\right]}_{\text{Bias of }\check{\tau}_j}.$$

Since $\widetilde{\mathcal{D}}_{\text{train}}$ and $\widetilde{\mathcal{D}}_{\text{cal}}$ are independent, the prediction error of the CATE model and the approximation error in the DR score are statistically uncorrelated. By Proposition 1, the DR score is an unbiased proxy, implying that $\mathbb{E}[\check{\epsilon}_j] = 0$. This eliminates the second term in the decomposition. As a result, minimizing the observed squared residuals using $\widetilde{\mathcal{D}}_{\text{cal}}$ directly reduces the true prediction error without risk of overfitting, even though $\check{\tau}_j$ is also computed from DP-protected data in $\widetilde{\mathcal{D}}_{\text{cal}}$.

Importantly, our sample-splitting approach is conceptually different from sample splitting and cross-fitting methods used in the existing literature (e.g., Nie and Wager 2021, Semenova and Chernozhukov 2021, Kennedy 2023). Prior methods focus on ensuring that the nuisance components (e.g., propensity scores or outcome models) are estimated independently of the data used to calculate DR scores. In contrast, we perform sample splitting specifically to prevent overfitting to the proxy error in the DR score during model calibration. (See Web Appendix B.2 for a detailed discussion of this distinction.)

4.2.3 Theoretical Guarantee

In Web Appendix C, we provide theoretical guarantees showing that *honest* model calibration improves true CATE prediction accuracy whenever it reduces mean squared error against the unbiased signal. Specifically, Theorem App-1 establishes that, with high probability, the reduction in prediction error is lower-bounded by the observed residual improvement, minus complexity and overfitting terms that typically shrinks as the calibration sample size increases. This result suggests that when the calibration sample is sufficiently large, the accuracy improvement from calibration is guaranteed to be positive and converges to the observed reduction in mean squared residuals. In Web Appendix C.4, we also show that performing non-honest calibration may increase the risk of overfitting without improving accuracy, particularly when the initial CATE model is already sufficiently expressive. In this case, the potential benefit of using a larger sample for initial model training is limited, as it does not change the convergence rate of with respect to the training sample size.

4.3 An Algorithm for Honest Model Calibration

We propose a boosting-based algorithm for honest model calibration. Our approach builds on the standard gradient boosting framework (Friedman 2001), with two key modifications to ensure honesty. First, calibration is performed using data held out from the initial model training. Second, step-size determination is carried out using data held out from the calibration model estimation. This layered separation helps prevent overfitting in high-noise and small-sample settings by using independent data for each update.

Specifically, we partition the DP-protected experimental data $\widetilde{\mathcal{D}}$ into three disjoint subsets: a training set ($\widetilde{\mathcal{D}}_{\text{train}}$), a calibration set ($\widetilde{\mathcal{D}}_{\text{cal}}$), and a validation set ($\widetilde{\mathcal{D}}_{\text{val}}$). The initial CATE model,

$\widehat{\tau}^{[0]}(\cdot \mid \widetilde{\mathcal{D}}_{\text{train}})$, is trained exclusively on $\widetilde{\mathcal{D}}_{\text{train}}$. Calibration is then performed using only $\widetilde{\mathcal{D}}_{\text{cal}}$, ensuring that updates are statistically independent of the data used for initial training. The validation set $\widetilde{\mathcal{D}}_{\text{val}}$ is used to assess which updates lead to the greatest improvements in predictive accuracy and to determine when to stop the calibration process. This three-way data split preserves the honesty of the procedure and reduces the risk of overfitting to the approximation error in the DR score.

Gradient Boosting for Model Calibration. We implement the calibration procedure using a gradient boosting framework (Friedman 2001). Let $\widehat{\tau}^{[0]}(\cdot)$ denote the initial CATE model trained on the DP-protected training set $\widetilde{\mathcal{D}}_{\text{train}}$. The goal is to iteratively refine this model so that its predictions better align with the DR scores using the DP-protected calibration set $\widetilde{\mathcal{D}}_{\text{cal}}$.

At each iteration $r = 1, \ldots, R$, we estimate a residual calibration model $\widehat{c}^{[r]}(\cdot)$ that captures the discrepancy between the current model's predictions and the DR scores. The CATE model is then updated as follows:

$$\widehat{\tau}^{[r]}(\cdot) = \widehat{\tau}^{[r-1]}(\cdot) + \rho^{[r]}\widehat{c}^{[r]}(\cdot),$$

where $\rho^{[r]}$ is the step size that controls the update magnitude. Each residual model $\widehat{c}^{[r]}$ and its corresponding step size $\rho^{[r]}$ are estimated by minimizing the sum of squared residuals between the updated model prediction and the DR scores over $\widetilde{\mathcal{D}}_{\text{cal}}$:

$$(\rho^{[r]}, \widehat{c}^{[r]}) = \arg\min_{\rho, \widehat{c} \in \mathcal{C}} \sum_{j \in \widetilde{\mathcal{D}}_{\text{cal}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j) + \rho \widehat{c}(\widetilde{\mathbf{X}}_j) - \check{\tau}_j \right]^2,$$

where $\mathcal{C}$ denotes the model class used for the calibration model. This iterative process continues for R rounds, gradually improving the alignment between model predictions and DR scores on the calibration set.

In our setting, each calibration model $\widehat{c}^{[r]}$ is trained to predict the residual difference between the unbiased signal and the current CATE prediction, i.e., $\check{\tau}_j - \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j)$. This learning target follows the idea of gradient descent: at each step, the algorithm updates the model by fitting to the negative gradient of the mean squared loss, which indicates the direction that most effectively reduces the residual error. In our case, this corresponds to the residuals between the predicted CATEs of the current model and the DR scores since

$$-\frac{\partial(\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j) - \check{\tau}_j)^2}{\partial \widehat{\tau}^{[r-1]}} = 2\left[\check{\tau}_j - \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j)\right].$$

Once the calibration model $\widehat{c}^{[r]}$ is learned, the next step is to determine the step size $\rho^{[r]}$, which controls the size of the model update. This step size is typically chosen to minimize the squared difference between the updated prediction and the target signal, i.e., by solving:

$$\rho^{[r]} = \arg\min_{\rho} \sum_{j} \big(\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j) + \rho \widehat{c}^{[r]}(\widetilde{\mathbf{X}}_j) - \check{\tau}_j\big)^2.$$

We adopt gradient boosting as the calibration framework for both theoretical and practical reasons. Theoretically, boosting constructs flexible yet well-controlled function classes by aggregating simple base learners. As shown in Corollary App-1, the complexity of the additive boosting model $\sum_{r=1}^{R} \rho^{[r]} \widehat{c}^{[r]}(\cdot)$ grows linearly with the complexity of the base learner class $\mathcal{C}$. This allows us to control overall model complexity by choosing a simple class for $\widehat{c}^{[r]}$, such as shallow decision trees or linear regressions. Practically, gradient boosting is computationally efficient and straightforward to implement across a wide range of modeling environments. It can be deployed with minimal development effort, making it particularly suitable for targeting contexts that require scalable and easy-to-use solutions.

Honest Step Size Determination. In standard implementations of gradient boosting (e.g., Friedman 2001, Duan et al. 2020), both the calibration model $\widehat{c}$ and the step size ρ are estimated using the same dataset (in our case, $\widetilde{\mathcal{D}}_{\text{cal}}$). However, tuning the step size on the same data used to fit the calibration model can also lead to overfitting, as it may overreact to noise that drives the discrepancy between the current prediction and the DR score. To address this, we introduce an honest step-size determination technique that separates calibration model construction from step-size tuning. This is implemented through a data-efficient sample-splitting strategy that ensures statistical independence between model fitting and step-size selection. Specifically, before initiating the calibration process, we partition the calibration set $\widetilde{\mathcal{D}}_{\text{cal}}$ into Q disjoint subgroups:$\widetilde{\mathcal{D}}_{\text{cal}}$ into Q disjoint subgroups: $\widetilde{\mathcal{G}}^1_{\text{cal}}, \widetilde{\mathcal{G}}^2_{\text{cal}}, \ldots, \widetilde{\mathcal{G}}^Q_{\text{cal}}$. Then, in each boosting iteration, we perform the following steps:

1. For each $q \in \{1, \ldots, Q\}$, fit a calibration model $\widehat{c}_q^{[r]}$ using only the data in $\widetilde{\mathcal{G}}^q_{\text{cal}}$. The model predicts the residual between the current prediction and the DR score, i.e., $\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j) - \check{\tau}_j$.

2. For each fitted calibration model, determine the optimal step size $\rho_q^{[r]}$ by minimizing the squared residuals on the remainder of the calibration data (i.e., all data in $\widetilde{\mathcal{D}}_{\text{cal}} \backslash \widetilde{\mathcal{G}}^q_{\text{cal}}$):
$$\rho_q^{[r]} = \arg\min_{\rho} \sum_{j \in \widetilde{\mathcal{D}}_{\text{cal}} \backslash \widetilde{\mathcal{G}}^q_{\text{cal}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j) + \rho \widehat{c}_q^{[r]}(\widetilde{\mathbf{X}}_j) - \check{\tau}_j \right]^2 .$$

3. Evaluate each candidate update $(\widehat{c}_q^{[r]}, \rho_q^{[r]})$ on the validation set $\widetilde{\mathcal{D}}_{\text{val}}$, and select the one that minimizes the validation loss:
$$q^\star = \underset{q \in \{1, \cdots, Q\}}{\arg\min} \sum_{k \in \widetilde{\mathcal{D}}_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) + \rho_q^{[r]} \widehat{c}_q^{[r]}(\widetilde{\mathbf{X}}_k) - \check{\tau}_k \right]^2 .$$

4. Update the CATE model using the selected subgroup calibration model and step size:
$$\widehat{\tau}^{[r]} := \widehat{\tau}^{[r-1]} + \rho_{q^\star}^{[r]} \widehat{c}_{q^\star}^{[r]} .$$

By using (i) disjoint subsets for calibration model fitting and step-size determination, and (ii) a separate validation set to identify the most effective update, this approach reduces the risk of overfitting to noise in the DR score.[4]

Early Stopping. To determine when to stop the boosting procedure, we monitor improvements in predictive accuracy on the validation set, following standard early stopping practices in the boosting literature (Friedman 2001). Specifically, the procedure is terminated once the updated model no longer improves the squared residual loss[5] relative to the previous iteration:
$$\sum_{k \in \widetilde{\mathcal{D}}_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) + \rho_{q^\star}^{[r]} \widehat{c}_{q^\star}^{[r]}(\widetilde{\mathbf{X}}_k) - \check{\tau}_i \right]^2 - \sum_{k \in \mathcal{D}_{\text{val}}} \left[\widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) - \check{\tau}_k \right]^2 > 0.$$

[4] In Web Appendix D.4.2, we also compare our honest step-size determination approach to two alternatives: (i) determining the step size using the same subgroup used to train the calibration model, and (ii) a stochastic gradient descent approach, where a random subgroup is selected in each iteration to both train the calibration model and determine the step size. Our proposed method significantly outperforms both benchmarks.

[5] The improvement threshold can also be specified as a tuning parameter. A larger threshold reduces the number of updates, acting as a form of regularization.

Algorithm 1 presents the pseudo-code for the proposed method. To further enhance sample efficiency, we follow the approach in (Kennedy 2023): swap the roles of $\widetilde{\mathcal{D}}_{\text{train}}$, $\widetilde{\mathcal{D}}_{\text{cal}}$, and $\widetilde{\mathcal{D}}_{\text{val}}$, rerun Algorithm 1, and average the resulting predictions.

4.4 Implementation Considerations

Our proposed solution supports flexible model choices and hyperparameter tuning. We offer the following recommendations to guide implementation. First, the initial CATE model should be selected based on predictive accuracy or targeting effectiveness, as well-performing baselines tend to maintain their advantage after calibration in our analyses. For the nuisance models ($\widehat{\mu}_1$ and $\widehat{\mu}_0$), while the DR score remains unbiased, we recommend following the literature and selecting models based on their predictive performance (Nie and Wager 2021, Kennedy 2023). Finally, to reduce overfitting and control model complexity, simple calibration models such as linear regression are recommended.

Second, our algorithm partitions the calibration data into subgroups and updates the model iteratively to ensure honest step-size determination. When sample sizes are small, grouping individuals with similar initial CATE predictions or covariate values is more effective, as the resulting homogeneity within each subgroup helps capture more stable residual patterns. For larger samples, random grouping performs comparably well, and the choice of subgrouping strategy becomes less critical. For hyperparameters, we recommend setting the number of subgroups Q such that each contains at least 100 observations, and setting the number of iterations $R = Q$, as the benefit of reusing subgroups tends to diminish after the first update.

Algorithm 1: Proposed Honest Model Calibration Procedure

```
Input: $Q \in \mathbb{N}$, $R \in \mathbb{N}$
Output: Calibrated model $\widehat{\tau}$
Data: DP-protected Data $\widetilde{\mathcal{D}} = \{\widetilde{\mathcal{D}}_{\text{train}}, \widetilde{\mathcal{D}}_{\text{cal}}, \widetilde{\mathcal{D}}_{\text{val}}\}$; True Propensity Score $e_i$
Construct the initial CATE model $\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}_{\text{train}})$ using $\widetilde{\mathcal{D}}_{\text{train}}$ and generate predictions for $\widetilde{\mathcal{D}}_{\text{cal}}$.
Derive the DR score $\{\check{\tau}_j\}_{j \in \widetilde{\mathcal{D}}_{\text{cal}}}$ and $\{\check{\tau}_k\}_{k \in \widetilde{\mathcal{D}}_{\text{val}}}$ separately using the cross-fitting procedure.
Divide $\widetilde{\mathcal{D}}_{\text{cal}}$ into $Q$ subgroups ($\widetilde{\mathcal{G}}^1_{\text{cal}}, \cdots, \widetilde{\mathcal{G}}^Q_{\text{cal}}$) based on some criteria.
for $r = 1, \cdots, R$ do
    Calculate the proxy residual error $\check{\tau}_j - \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j)$ for $j \in \widetilde{\mathcal{D}}_{\text{cal}}$.
    for $q = 1, \cdots, Q$ do
        (Calibration Model Fittings) Construct a calibration model, denoted $\widehat{c}^{[r]}_q(\widetilde{\mathbf{X}}j)$, trained to predict the residual $\check{\tau}_j - \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j)$ using only the data from customers $j \in \widetilde{\mathcal{G}}^q_{\text{cal}}$.
        (Step Size Determination) Determine the step size by minimizing the squared proxy residuals across ciustomers in $\widetilde{\mathcal{D}}_{\text{cal}} \backslash \widetilde{\mathcal{G}}^q_{\text{cal}}$:
            $$\rho^{[r]}_q = \arg\min_{\rho} \sum_{j \in \widetilde{\mathcal{D}}_{\text{cal}} \backslash \widetilde{\mathcal{G}}^q_{\text{cal}}} \left[ \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_j) + \rho \widehat{c}^{[r]}_q(\widetilde{\mathbf{X}}_j) - \check{\tau}_j \right]^2$$
    end
    Pick the subgroup that leads to the smallest squared proxy residuals in the validation set $\widetilde{\mathcal{D}}_{\text{val}}$.
    Generate new CATE predictions for the validation set:
        $$\widehat{\tau}^{\text{new}}(\widetilde{\mathbf{X}}_k) = \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) + \rho^{[r]}_{q^\star} \widehat{c}^{[r]}_{q^\star}(\widetilde{\mathbf{X}}_k \mid \widetilde{\mathcal{G}}^{q^\star}_{\text{cal}}), \; k \in \widetilde{\mathcal{D}}_{\text{val}}.$$
    if $\sum_{k \in \widetilde{\mathcal{D}}_{val}} \left[ \widehat{\tau}^{new}(\widetilde{\mathbf{X}}_k) - \check{\tau}_k \right]^2 - \sum_{k \in \widetilde{\mathcal{D}}_{val}} \left[ \widehat{\tau}^{[r-1]}(\widetilde{\mathbf{X}}_k) - \check{\tau}_k \right]^2 < 0$ then
        Update the CATE model: $\widehat{\tau}^{[r]}(\cdot) = \widehat{\tau}^{[r-1]}(\cdot) + \rho^{[r]}_{q^\star} \widehat{c}^{[r]}_{q^\star}(\widetilde{\mathbf{X}}_k \mid \widetilde{\mathcal{G}}^{q^\star}_{\text{cal}})$.
    end
    else
        Stop calibration and return $\widehat{\tau}(\cdot) = \widehat{\tau}^{[r-1]}(\cdot)$.
    end
end
```

5 Empirical Performance: Simulation

We conduct simulation analyses for two primary purposes. First, to evaluate the performance of the proposed method relative to several alternative benchmarks. Second, to examine how our algorithm and competing methods behave across different sample sizes, thereby providing insight into their asymptotic properties.

5.1 Simulation Setup

We generate an experimental sample with a binary treatment variable ($W_i \in \{0, 1\}$) where the treatment assignment is completely random, with equal proportions for both treatment and control groups. The outcome variable is generated according to the following process:

$$\begin{aligned}
Y_i(W_i) &= b(\mathbf{X}_i) + (W_i - 0.5)\tau(\mathbf{X}_i) + \epsilon_i,\ \epsilon_i \sim_{i.i.d.} \mathcal{N}(0, 5), \\
b(\mathbf{X}_i) &= \frac{1}{4}\left[\sin(\pi X_{i,1} X_{i,2}) - 2(X_{i,1} - X_{i,3} - 0.5)^2 + X_{i,2}X_{i,4} + 2X_{i,5}^2 + X_{i,6}^2\right], \\
\tau(\mathbf{X}_i) &= \frac{1}{4}\left[2X_{i,1} - X_{i,3}\text{cos}(\pi X_{i,2}) + 0.5X_{i,3}^2 - X_{i,4} - \log(X_{i,1})(X_{i,4} - 2.5) - 8\right],
\end{aligned}$$

where each covariate is identically and independently distributed (i.i.d.) from Uniform$(0, 5)$.

We analyze two scenarios under the data-generating process described above. In the first scenario, only covariates are protected under DP. Specifically, the observed covariates are noisy versions of the true values: $\widetilde{X}_{i,p} = X_{i,p} + \eta_{i,p}$, where $\eta_{i,p} \sim$ Laplace$(0, \frac{\Delta_X}{\epsilon})$ are i.i.d Laplace noises. Here, Δ_X denotes the range of the covariates, and the ϵ values reflects the privacy budget for one covariate. In the second scenario, only the outcome variable is protected under DP. The observed outcome is a noisy version of the true potential outcome: $\widetilde{Y}_i(W_i) = Y_i(W_i) + \eta_i$, where $\eta_i \sim$ Laplace$(0, \frac{\Delta_Y}{\epsilon})$ and Δ_Y is the empirical range of the outcome.[6] For both scenarios, we vary the privacy level ϵ from 10 to 50 to examine the trade-off between privacy protection and predictive accuracy,[7] and also vary the experimental sample size to assess the sample efficiency of competing methods.

[6] Web Appendix D.3 presents additional results where both covariates and outcomes are protected by DP. The findings are consistent.

[7] These values are motivated by the 2020 U.S. Census, which used a total privacy budget of 19.61; internal evaluations showed that a budget of 12.2 offered strong privacy but introduced too much noise for downstream analyses. In our implementation, the corresponding standard deviations for Laplace noise are as follows: $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ for covariates; $\{1, 2, 3, 4, 5\}$ for the outcome.

5.2 Methods for Comparison

We compare the performance of our proposed solution against three alternatives: a standard CATE model and two simplified versions of our approach. This comparison not only establishes the superiority of our method but also quantifies the improvements gained from its individual components. We evaluate a total of four estimation methods:

1. **Default** (DEFAULT): This method constructs a CATE model using the full experimental dataset $\widetilde{\mathcal{D}} = \{\widetilde{\mathcal{D}}_{\text{train}}, \widetilde{\mathcal{D}}_{\text{cal}}, \widetilde{\mathcal{D}}_{\text{val}}\}$ without any model calibration. It represents the standard approach typically adopted by decision-makers for CATE estimation.

2. **Non-honest Calibration** (NON-HONEST): This method applies calibration by performing traditional gradient boosting on the initial CATE model using the full dataset $\widetilde{\mathcal{D}}$, without any sample splitting or honest step-size determination. Essentially, it is equivalent to the CATE estimation framework proposed by Kennedy (2023), with boosting methods as the model class for CATE modeling.

3. **Calibration with Sample Splitting** (SPLIT-ONLY): This method performs model calibration using $\widetilde{\mathcal{D}}_{\text{cal}}$ and stops when there is no improvement on a third validation set $\widetilde{\mathcal{D}}_{\text{val}}$. However, it determines the calibration model and step size using all individuals in $\widetilde{\mathcal{D}}_{\text{cal}}$ without honest step-size determination.

4. **Calibration with Sample Splitting and Honest Step-size Determination** (PROPOSED): Beyond using sample splitting for model calibration, this method also implements the proposed honest step-size determination procedure, in which subgroups are formed based on the predicted values from the initial CATE model.

For the DEFAULT and NON-HONEST methods, the entire dataset is used to train the initial CATE model, with NON-HONEST performing calibration on the same data. In contrast, the PROPOSED and SPLIT-ONLY methods partition the same dataset into three equal-sized folds, used respectively for model training, calibration, and validation.

5.3 Estimation Details

We summarize the key details of the model implementation below. Additional specifications and technical information are provided in Web Appendix D.1.

In our main analysis, we use the R-learner with regression forests (Nie and Wager 2021) as both the DEFAULT method and the initial CATE model across all approaches, as it delivers the best overall performance among the CATE models considered.

For DR socres, we construct the conditional mean outcome models used in the DR scores using ordinary least squares (OLS) regression with squared and interaction terms for both the initial CATE estimators and for the calibration targets. This specification was selected based on its predictive performance among the candidate models considered. The calibration models are similarly estimated using OLS with the same set of nonlinear terms.

5.4 Evaluation Procedure

We evaluate model performance using $B = 100$ bootstrap replications and report the average values of key metrics. As a first step, we generate a holdout set $\widetilde{\mathcal{D}}_{\text{holdout}}$ of $N_{\text{holdout}} = 10{,}000$ individuals. This set is excluded from all model estimation and calibration and is used solely for evaluation across bootstrap replications. In each replication b, we generate an experimental set $\widetilde{\mathcal{D}}^b$ and use it to estimate CATE using each of the four different methods. Predictions are then generated on the holdout set, and prediction errors are calculated for each method. Specifically, the prediction error for individual $l \in \widetilde{\mathcal{D}}_{\text{holdout}}$ is defined as $\text{err}_l^b = \widehat{\tau}_l^b - \tau(\mathbf{X}_l)$, where $\widehat{\tau}_l^b$ denotes the final prediction of individual l in the b-th replication. After completing this process across $B = 100$ bootstrap replications, we compute three key summary statistics to evaluate statistical accuracy and targeting efficiency:

Mean Squared Error (MSE). We first calculate the prediction error for individual $l \in \widetilde{\mathcal{D}}_{\text{holdout}}$ is defined as $\text{err}_l^b = \widehat{\tau}_l^b - \tau(\mathbf{X}_l)$, where $\widehat{\tau}_l^b$ denotes the final prediction of individual l in the b-th replication. After completing this process across $B = 100$ bootstrap replications, we compute the mean squared error (MSE)[8] for each method as follows:

$$\widehat{\mathbb{E}}\left[(\mathrm{err}_l^b)^2\right] = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_{\text{holdout}}}\sum_{l\in\widetilde{\mathcal{D}}_{\text{holdout}}}\left[\mathrm{err}_l^b\right]^2.$$

Area under the Targeting Operator Characteristic Curve (AUTOC). We use AUTOC (Yadlowsky et al. 2024) to assess the improvement in treatment prioritization performance achieved by our proposed method when decision-makers rely on CATE models estimated from DP-protected data. A higher AUTOC value indicates that the CATE model more effectively identifies individuals with the greatest incremental impact from the intervention and produces better treatment prioritization rules.

Targeting Value Improvement. We evaluate the value improvement for of the PROPOSED method for a simple targeting rule that assigns treatment to individuals with positive predicted CATEs. Specifically, we measure the incremental value gained from targeting individuals with positive predicted CATEs: $V(\widehat{\tau}) = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{1}{N_{\text{holdout}}}\sum_{l\in\widetilde{\mathcal{D}}_{\text{holdout}}}\tau(\mathbf{X}_l)\cdot\mathbb{1}(\widehat{\tau}_l^b > 0)\right]$, where $\widehat{\tau}_l^b$ is the predicted CATE of individual l in the b-th bootstrap replication. We then report the value improvement of a method $\widehat{\tau}$ compared to the DEFAULT method $\widehat{\tau}^{\text{DEFAULT}}$ as the percentage difference in their targeting value, i.e., $100\% \times \frac{V(\widehat{\tau})-V(\widehat{\tau}^{\text{DEFAULT}})}{V(\widehat{\tau}^{\text{DEFAULT}})}$.

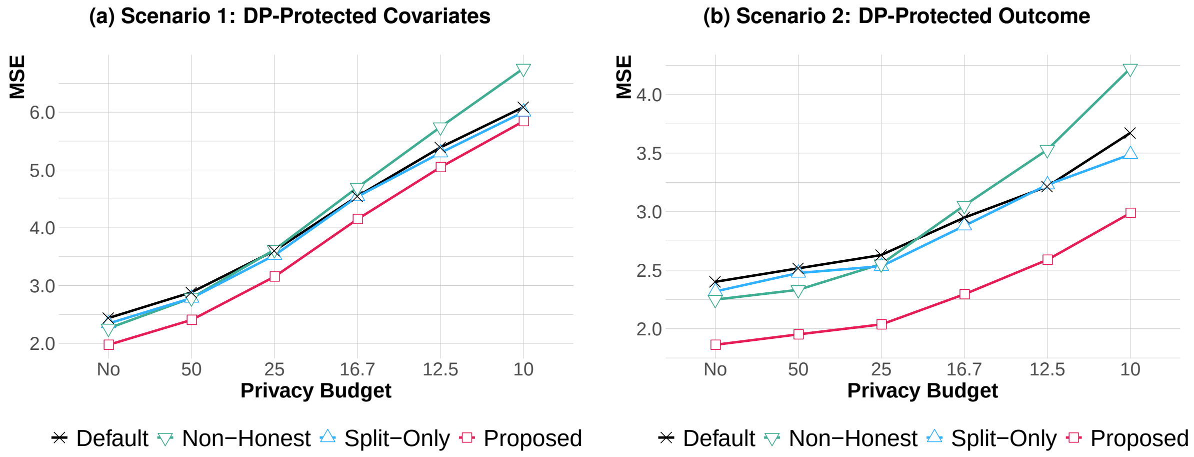
5.5 Results

5.5.1 Varying Privacy Level

Figure 1 reports the MSE of the different methods across a range of privacy levels, with the leftmost point on the x-axis representing the baseline without DP protection. Figure 1a presents results when covariates are protected under DP, and Figure 1b shows results when the outcome variable is protected. All models are trained on an experimental sample of 3,000 individuals, and performance is evaluated on a separate holdout set of 10,000 individuals.

[8] In Web Appendix D.2, we also present the mean squared bias and variance for model comparison.

Figure 1: Mean Squared Errors Across Varying Privacy Levels



Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap MSE over a holdout set of 10k individuals for each point. The results presented here are based on using R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Web Appendix D.5.

Several patterns emerge from the results. First, the PROPOSED method (red line) consistently achieves the lowest MSE, regardless of whether DP is applied to covariates or outcomes. Notably, our method also improves accuracy in non-DP settings (left-most results), suggesting that it can enhance targeting performance more broadly, even when the noise is moderate and not specific to DP. Second, neither the SPLIT-ONLY (blue line) nor the NON-HONEST (green line) approach improves accuracy relative to the DEFAULT (black line). While the SPLIT-ONLY method performs comparably to the NON-HONEST approach at lower privacy levels, it outperforms the NON-HONEST method under high noise level. This highlights the risk of overfitting when calibration is conducted without sample splitting, particularly in high-noise settings.

Next, we assess the effectiveness of treatment prioritization based on predicted CATEs from each method. Table 2 reports the AUTOC values for all approaches, along with the proportion of bootstrap replications (in parentheses) in which each method outperforms the DEFAULT. Consistent with earlier results, the PROPOSED method achieves the highest AUTOC across all privacy levels. The SPLIT-ONLY method also improves upon the DEFAULT, whereas the NON-HONEST method offers little to no improvement. These findings highlights the importance of maintaining honesty during model calibration.

Table 2: AUTOC Values Across Varying Privacy Levels

(a) Scenario 1: DP-Protected Covariates

Privacy	Proposed	Split-Only	Non-Honest	Default
No	2.39 (100%)	2.36 (90%)	2.34 (68%)	2.32
50.0	2.31 (100%)	2.27 (84%)	2.25 (74%)	2.23
25.0	2.17 (100%)	2.14 (90%)	2.11 (76%)	2.08
16.7	1.99 (100%)	1.96 (94%)	1.90 (56%)	1.90
12.0	1.8 (100%)	1.77 (94%)	1.71 (58%)	1.70
10.0	1.61 (96%)	1.59 (92%)	1.50 (32%)	1.52

(b) Scenario 2: DP-Protected Outcome

Privacy	Proposed	Split-Only	Non-Honest	Default
No	2.42 (96%)	2.39 (84%)	2.38 (71%)	2.36
50.0	2.42 (99%)	2.37 (84%)	2.36 (76%)	2.34
25.0	2.40 (97%)	2.37 (84%)	2.34 (62%)	2.32
16.7	2.37 (98%)	2.32 (88%)	2.28 (68%)	2.26
12.0	2.32 (96%)	2.29 (92%)	2.23 (61%)	2.22
10.0	2.25 (98%)	2.23 (96%)	2.14 (68%)	2.13

Note: We calculate the AUTOC values from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on using R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Web Appendix D.5.

Table 3 summarizes the improvement in targeting value for each method relative to the Default. We report both the percentage increase in value and the proportion of bootstrap replications in which each method outperforms the Default. The Proposed method consistently delivers substantial and significant gains across all levels of privacy protection, with its advantage becoming more pronounced as the privacy level increases. By contrast, neither the Split-only nor the Non-Honest method yields meaningful improvements, and the Non-Honest method performs even worse than the Default when covariates are heavily distorted by noise. These results highlight the risk of overfitting when applying a complex, data-adaptive calibration procedure for targeting without proper sample splitting.

Table 3: Targeting Value Improvement Across Varying Privacy Levels

(a) Scenario 1: DP-Protected Covariates

Privacy	Proposed	Split-Only	Non-Honest
No	7.0% (97%)	2.4% (65%)	2.8% (75%)
50.0	8.4% (98%)	2.5% (74%)	2.5% (63%)
25.0	10.6% (100%)	3.1% (71%)	2.8% (66%)
16.7	13.6% (100%)	0.6% (53%)	1.7% (50%)
12.5	18.6% (98%)	1.6% (52%)	1.5% (51%)
10.0	14.9% (85%)	2.6% (54%)	-10.3% (32%)

(b) Scenario 2: DP-Protected Outcome

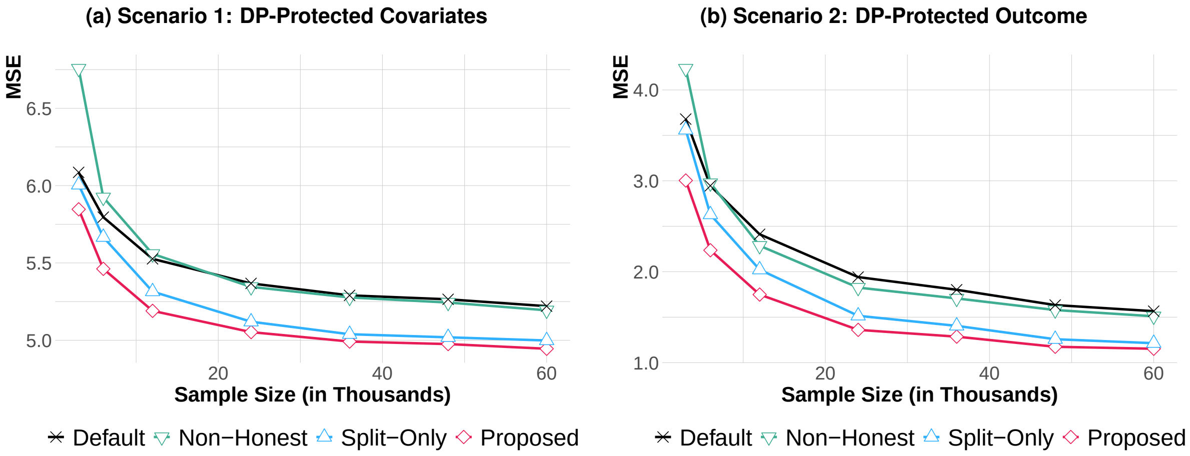
Privacy	Proposed	Split-Only	Non-Honest
No	6.9% (98%)	1.2% (69%)	3.5% (76%)
50.0	7.9% (98%)	1.2% (67%)	3.2% (69%)
25.0	8.8% (99%)	2.1% (66%)	3.5% (79%)
16.7	10.8% (96%)	0.9% (53%)	3.6% (66%)
12.5	11.5% (94%)	-0.1% (62%)	1.7% (58%)
10.0	13.8% (94%)	0.1% (52%)	1.0% (56%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner as the initial CATE model. Results derived from different CATE models are available in Web Appendix D.5.

5.5.2 Varying Sample Sizes

Next, to examine the asymptotic behavior of each method, we compare their performance when constructing models with experimental sample sizes ranging from 3,000 to 60,000 individuals, while keeping the privacy level fixed at a high level ($\epsilon = 10.0$). Figure 2 reports the MSE of each method at different sample sizes for model construction.

Figure 2: Mean Squared Errors Across Varying Sample Size


(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap MSE over a holdout set of 10k individuals for each point. The results presented here are based on using R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Web Appendix D.5.

First, as expected, MSE decreases as the sample size increases. Notably, the PROPOSED method (red line) consistently achieves the lowest MSE across all sample sizes, demonstrating both sample efficiency and superior performance, even with large experimental samples. Second, once the sample size is sufficiently large, the SPLIT-ONLY method (blue line) significantly outperforms the DEFAULT (black line) and approaches the performance of the PROPOSED method. This pattern is consistent with Theorem App-1, which predicts that calibration on independent data improves prediction accuracy when the calibration sample is large enough. Since both the PROPOSED and SPLIT-ONLY methods use the same additive boosting model, their improvements in mean squared residuals eventually converge, leading to similar accuracy as the calibration sample grows.

Third, the Non-Honest method (green line) yields only marginal improvements in MSE over the Default (black line). This result is consistent with the theoretical analysis in Web Appendix C.4, which shows that when the Default model is already highly expressive, non-honest calibration adds little value. In such cases, most residual error has already been addressed during the initial model construction, so further calibration without sample splitting provides limited gains in improving MSE.

Table 4 presents the AUTOC values for different methods across varying sample sizes. Consistent with previous findings, the Proposed method consistently outperforms all alternatives in treatment prioritization, achieving the highest AUTOC values across all sample sizes. Second, when sample sizes are large, both the Split-only and Non-Honest methods can outperform the Default approach, with Split-only achieving performance which is comparable to the Proposed specification with large samples.

Table 4: AUTOC Values Across Varying Sample Sizes

(a) Scenario 1: DP-Protected Covariates

Sample Size	Proposed	Split-Only	Non-Honest	Default
3k	1.61 (96%)	1.59 (92%)	1.50 (32%)	1.52
6k	1.67 (100%)	1.66 (100%)	1.59 (70%)	1.58
12k	1.72 (100%)	1.71 (100%)	1.66 (90%)	1.65
24k	1.75 (100%)	1.74 (100%)	1.70 (95%)	1.68
36k	1.76 (100%)	1.75 (100%)	1.71 (100%)	1.70
48k	1.76 (100%)	1.76 (100%)	1.72 (100%)	1.70
60k	1.77 (100%)	1.76 (100%)	1.72 (100%)	1.71

(b) Scenario 2: DP-Protected Outcome

Sample Size	Proposed	Split-Only	Non-Honest	Default
3k	2.25 (98%)	2.23 (96%)	2.14 (68%)	2.13
6k	2.37 (100%)	2.34 (98%)	2.29 (92%)	2.25
12k	2.44 (100%)	2.42 (98%)	2.37 (90%)	2.34
24k	2.49 (100%)	2.48 (100%)	2.43 (100%)	2.4
36k	2.50 (100%)	2.49 (100%)	2.45 (100%)	2.43
48k	2.52 (100%)	2.51 (100%)	2.47 (100%)	2.45
60k	2.52 (100%)	2.52 (100%)	2.48 (100%)	2.46

Note: We calculate the AUTOC values from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on using R-learner with regression forests as the initial CATE model. Results derived from different CATE models are available in Web Appendix D.5.

Table 5 summarizes the improvement in targeting value of each method relative to the Default across different sample sizes. Consistent with earlier results, the Proposed method outperforms all alternatives, achieving the highest performance at every sample size. As sample size increases, the Split-only method begins to approach the performance of the Proposed method, whereas the Non-Honest method delivers only marginal gains compared to the other two calibration approaches.

Table 5: Targeting Value Improvement Across Varying Sample Sizes

(a) Scenario 1: DP-Protected Covariates

Sample Size	Proposed	Split-Only	Non-Honest
3k	14.9% (85%)	2.6% (54%)	-10.3% (32%)
6k	17.5% (97%)	3.4% (68%)	0.7% (52%)
12k	13.7% (100%)	8.0% (93%)	0.0% (51%)
24k	11.5% (100%)	9.9% (100%)	1.1% (62%)
36k	10.8% (100%)	9.1% (100%)	0.9% (65%)
48k	10.6% (100%)	9.5% (100%)	1.3% (70%)
60k	11.4% (100%)	8.8% (100%)	1.5% (63%)

(b) Scenario 2: DP-Protected Outcome

Sample Size	Proposed	Split-Only	Non-Honest
3k	13.8% (94%)	0.1% (52%)	1.0% (56%)
6k	10.7% (100%)	5.6% (88%)	2.8% (74%)
12k	8.3% (100%)	5.3% (94%)	2.6% (82%)
24k	6.6% (100%)	5.1% (100%)	2.2% (97%)
36k	5.6% (100%)	4.9% (100%)	1.6% (98%)
48k	5.5% (100%)	4.5% (100%)	1.4% (92%)
60k	4.2% (100%)	3.6% (100%)	1.2% (93%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner as the initial CATE model. Results derived from different CATE models are available in Web Appendix D.5.

Together, these results highlight the value of our proposed honest calibration algorithm in improving model performance under DP protection. Specifically, the Proposed method consistently delivers the lowest MSE, the highest AUTOC, and the greatest value improvement across a wide range of privacy levels and sample sizes. Its performance gains are especially clear when the privacy constraint is strong and/or the sample size is small. The Split-only method also performs better than the Default method when the calibration sample is large, which is consistent with theoretical results on the benefits of honest calibration. In contrast, the Non-Honest method shows limited or no improvement and performs worse under high noise, highlighting the importance of using sample splitting during model calibration.

5.6 Robustness Checks

In Web Appendix D, we conduct a series of robustness checks. First, we replicate the simulation while applying DP protection to both outcomes and covariates. The results remain consistent with the main findings presented above. Second, we examine alternative implementation choices for the calibration step, including different subgroup partitioning strategies. We find that partitioning based on initial CATE predictions yields the strongest performance, while covariate-based subgrouping also delivers comparable improvements. Third, we evaluate the honest step-size determination against alternative boosting techniques and show that the proposed method consistently outperforms these alternatives. Finally, we assess the robustness of our results to different model classes used in the DEFAULT and for the initial CATE models. Across all specifications, the PROPOSED method consistently outperforms the alternatives.

6 Empirical Performance: Real-world Case Studies

We validate the proposed solution using two real-world applications that serve as established benchmarks for CATE models (e.g., Rößler and Schoder 2022). Both studies involve randomized marketing campaigns designed to encourage customer purchases.

6.1 Studies Overview

6.1.1 Study 1: Hillstrom E-mail Campaign.

The first study uses the Hillstrom dataset, originally from the MineThatData E-Mail Analytics and Data Mining Challenge (Hillstrom 2008). This dataset includes 64k customers, who were randomly assigned to one of three groups in an e-mail marketing experiment: a group that received an e-mail promoting men's merchandise, a group that received an e-mail promoting women's merchandise, and a control group that received no e-mail campaign. Following prior research using this dataset (Kane et al. 2014, Devriendt et al. 2018, Rößler and Schoder 2022), we focus on evaluating the effectiveness of the women's merchandise e-mail promotion compared to no e-mail at all. Our final sample consists of 42,693 customers, evenly split between the treatment group (21,387 customers) and the control group (21,306 customers). Further data descriptions can be found in Web Appendix E.

6.1.2 Study 2: Starbucks Promotional Campaign Data.

The second study uses data from a promotional campaign conducted through the Starbucks Rewards mobile app, made available by the Udacity Data Science Program. This dataset captures an experiment in which a subset of customers was randomly offered a promotion (the intervention) to encourage product purchases (the outcome variable). The dataset includes 126,184 customers with seven anonymous pre-treatment covariates. The sample is evenly split, with 63,112 customers in the treatment group and 63,072 in the control group. The response rates are notably low — 1.68% in the treatment group and 0.73% in the control group, resulting in an average treatment effect of 0.95%. Additional details can be found in Web Appendix F.

6.2 Implementation of DP

We consider two scenarios implementing the most common DP methods: one where pre-treatment covariates are protected by DP and another where the outcome is protected by DP. For discrete variables, we apply the randomized response mechanism (Dwork et al. 2014). Under this mechanism, the true value is observed with probability $1 - f$, while with probability f, a random draw from all possible values (each appearing with equal probability) is observed instead. For continuous variables, we introduce the Laplace(σ) noise. The privacy level is varied from a "Very Low" scenario (i.e., small f and σ) to a "Very High" (i.e., large f and σ) scenario. Additional implementation details can be found in Web Appendix E.3 and F.3.

6.3 Methods for Comparison and Estimation Details

As in Section 5, we compare the PROPOSED method with the DEFAULT method, which estimates the CATE model directly without honest model calibration, as well as the NON-HONEST and SPLIT-ONLY methods. The model specifications closely follow those described in Section 5, with the following key details.

In the main analysis, we use the R-learner with regression forests (Nie and Wager 2021) as the DEFAULT method and as the initial CATE model across different approaches. The DR score is computed using OLS regression for the conditional mean outcome models, as this specification provides the highest predictive accuracy among the candidate models tested. Similarly, the calibration models $\hat{c}_q^{[r]}$ are also constructed via OLS regression. For detailed model specifications and the complete set of results across different DEFAULT models, please refer to Web Appendix E and F.

6.4 Performance Evaluation

To evaluate the performance of each method, we adopt a bootstrap validation procedure similar to that used by Ascarza (2018). Specifically, we generate $B = 50$ random splits, each consisting of a model construction set (70%) and a holdout set (30%). For each split, we estimate CATE models using the four methods (Proposed, Default, Non-Honest, and Split-only) on the model construction set and generate CATE predictions for individuals in the holdout set.When covariates are protected by differential privacy, noise is added to both the model construction and holdout covariates to capture the full impact of privacy protection on targeting decisions. In contrast, when the outcome variable is protected, we do not inject noise into the holdout set, as it does not affect the evaluation of true model performance under DP constraints.

We evaluate each method using three metrics: First, we assess predictive accuracy using *group-level treatment effects*, as the true CATE for each individual is not directly observable in real-world data. Specifically, we divide the holdout set into ten equally sized groups based on predicted CATE rankings, compute the group average treatment effect (GATE) using both predicted CATEs and actual observed outcomes, and calculate the root mean squared error (RMSE) between them. This procedure is repeated across 50 bootstrap splits, and the RMSEs are averaged. Second, we assess *treatment prioritization* using the AUTOC metric, which measures how well each method ranks individuals according to their predicted treatment effects. AUTOC is computed on the holdout set for each split and averaged across all replications.

Third, we assess targeting value improvement by estimating the *policy value* of each method using the inverse probability-weighting estimator (Yoganarasimhan et al. 2022, Hitsch et al. 2024). For each method, we compute the value of a targeting policy that treats individuals with positive predicted CATEs and report its improvement relative to the Default method. Specifically, we calculate the percentage improvement in targeting value as $100\% \times \frac{V(\hat{\tau})-V(\hat{\tau}^{\text{Default}})}{V(\hat{\tau}^{\text{Default}})}$, where $V(\cdot)$ denotes the estimated policy value.

6.5 Results

Table 6 presents the RMSE of GATE (multiplied by 100) for different methods, along with the percentage of random splits where the focal method achieves a lower RMSE compared to the Default method.

Table 6: RMSE of GATE (Multiplied by 100) Across Varying Privacy Levels

(a) Study 1: Hillstrom Data

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	Proposed	Split-Only	Non-Honest	Default	Proposed	Split-Only	Non-Honest	Default
No	3.96 (100%)	3.98 (100%)	7.29 (52%)	7.29	3.96 (100%)	3.98 (100%)	7.29 (52%)	7.29
Very Low	3.14 (100%)	3.19 (100%)	5.20 (52%)	5.20	4.35 (100%)	4.43 (100%)	7.88 (44%)	7.86
Low	3.39 (100%)	3.43 (100%)	5.56 (46%)	5.55	4.39 (100%)	4.43 (100%)	8.18 (36%)	8.14
Medium	3.47 (100%)	3.55 (100%)	5.61 (60%)	5.62	4.92 (100%)	5.05 (100%)	8.76 (40%)	8.71
High	3.47 (100%)	3.50 (100%)	5.61 (34%)	5.56	4.99 (100%)	5.04 (100%)	8.91 (40%)	8.90
Very High	3.44 (100%)	3.48 (100%)	5.62 (42%)	5.60	5.10 (100%)	5.16 (100%)	9.33 (40%)	9.29

(b) Study 2 : Starbucks Data

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	Proposed	Split-Only	Non-Honest	Default	Proposed	Split-Only	Non-Honest	Default
No	0.71 (100%)	0.71 (100%)	1.40 (32%)	1.38	0.71 (100%)	0.71 (100%)	1.40 (32%)	1.38
Very Low	0.73 (100%)	0.74 (100%)	1.42 (34%)	1.41	1.59 (100%)	1.60 (100%)	2.85 (36%)	2.85
Low	0.78 (100%)	0.78 (100%)	1.44 (40%)	1.43	2.12 (100%)	2.14 (100%)	3.71 (48%)	3.71
Medium	0.80 (100%)	0.82 (100%)	1.47 (44%)	1.47	2.52 (100%)	2.52 (100%)	4.36 (48%)	4.36
High	0.84 (100%)	0.86 (100%)	1.47 (34%)	1.45	2.82 (100%)	2.84 (100%)	4.85 (36%)	4.84
Very High	0.84 (100%)	0.84 (100%)	1.42 (52%)	1.42	3.06 (100%)	3.09 (100%)	5.27 (32%)	5.27

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Consistent with the theoretical insights from Section 4 and the simulation results in Section 5, both the Proposed and Split-only methods improve accuracy relative to the Default, while the Non-Honest method fails to do so. Moreover, while RMSE rises with stronger protection on the outcome variable, this pattern does not always hold when noise is applied to covariates. While this may seem counterintuitive, it aligns with our bias–variance discussion in Web Appendix A. Adding noise to covariates reduces the model's ability to detect heterogeneity, which leads to more conservative (i.e., less variable) predictions. For example, in the Hillstrom dataset, we find that the variance of predicted CATEs from the Default method decreases by 20% under "Very Low" covariate protection compared to the no-privacy baseline.

Table 7 reports the AUTOC values (scaled by 100) for each method across different levels of privacy protection. First, both the Proposed and Split-only approaches consistently outperform the Default, whereas the Non-Honest method sometimes fails to improve AUTOC values. Second, consistent with earlier findings, the Proposed and Split-only methods both

Table 7: AUTOC Values (Multiplied by 100) Across Varying Privacy Levels

(a) Study 1: Hillstrom Data

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	1.38 (82%)	1.36 (82%)	1.18 (60%)	1.17	1.37 (86%)	1.38 (82%)	1.18 (60%)	1.17
Very Low	1.46 (92%)	1.42 (88%)	1.17 (80%)	1.13	1.10 (76%)	1.07 (72%)	0.92 (54%)	0.91
Low	1.24 (94%)	1.23 (98%)	0.98 (82%)	0.92	1.11 (82%)	1.08 (78%)	0.90 (70%)	0.87
Medium	1.12 (90%)	1.09 (86%)	0.83 (84%)	0.76	0.99 (74%)	0.95 (68%)	0.84 (62%)	0.80
High	0.98 (92%)	0.95 (90%)	0.74 (84%)	0.69	0.97 (78%)	0.96 (66%)	0.85 (72%)	0.81
Very High	0.94 (90%)	0.91 (88%)	0.69 (84%)	0.62	0.75 (86%)	0.73 (82%)	0.56 (62%)	0.52

(b) Study 2 : Starbucks Data

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	0.45 (96%)	0.44 (96%)	0.38 (64%)	0.38	0.45 (96%)	0.44 (96%)	0.38 (64%)	0.38
Very Low	0.43 (94%)	0.42 (86%)	0.37 (84%)	0.35	0.20 (84%)	0.20 (84%)	0.16 (92%)	0.14
Low	0.33 (92%)	0.32 (78%)	0.29 (84%)	0.27	0.13 (84%)	0.13 (80%)	0.11 (100%)	0.08
Medium	0.25 (92%)	0.24 (92%)	0.21 (82%)	0.19	0.11 (88%)	0.08 (76%)	0.06 (100%)	0.04
High	0.19 (78%)	0.19 (76%)	0.18 (78%)	0.16	0.09 (76%)	0.09 (68%)	0.07 (92%)	0.05
Very High	0.19 (74%)	0.18 (74%)	0.17 (76%)	0.15	0.09 (80%)	0.09 (68%)	0.08 (92%)	0.06

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

outperform the NON-HONEST, with the PROPOSED method showing slightly stronger performance than SPLIT-ONLY.

Table 8 reports the average targeting value improvement across different privacy levels. Consistent with the ranking in Table 7, both the PROPOSED and SPLIT-ONLY methods deliver substantial gains over the DEFAULT baseline at all levels of DP, whereas the NON-HONEST approach fails to improve value on the Hillstrom data. These results demonstrate the importance of honest calibration for enhancing targeting performance.

6.6 Small Sample Performance

While the PROPOSED and SPLIT-ONLY methods perform similarly in the main analyses (largely due to the ample sample sizes available in both datasets), we observe larger differences in smaller-sample settings. Specifically, we evaluate performance when only 10% of the data is used for model construction and 90% is reserved for holdout evaluation. In this setting, although SPLIT-ONLY continues to outperform the DEFAULT, the gap between PROPOSED and

Table 8: Targeting Value Improvement Across Varying Privacy Levels

(a) Study 1: Hillstrom Data

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	Proposed	Split-Only	Non-Honest	Proposed	Split-Only	Non-Honest
No No	2.25% (88%)	2.22% (92%)	-0.03% (56%)	2.25% (88%)	2.22% (92%)	-0.03% (56%)
Very Low	2.25% (92%)	2.13% (86%)	0.10% (50%)	2.32% (94%)	2.33% (98%)	0.06% (58%)
Low	2.28% (90%)	2.36% (92%)	0.03% (48%)	2.48% (92%)	2.51% (94%)	0.24% (66%)
Medium	2.73% (94%)	2.59% (92%)	0.10% (56%)	2.73% (98%)	2.64% (94%)	0.21% (70%)
High	2.64% (96%)	2.58% (96%)	0.12% (60%)	2.80% (94%)	2.85% (98%)	0.23% (72%)
Very High	2.88% (96%)	2.84% (98%)	0.16% (64%)	3.24% (100%)	3.18% (98%)	0.16% (66%)

(b) Study 2 : Starbucks Data

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	Proposed	Split-Only	Non-Honest	Proposed	Split-Only	Non-Honest
No	3.31% (100%)	3.16% (96%)	0.07% (52%)	3.31% (100%)	3.16% (96%)	0.07% (52%)
Very Low	4.84% (96%)	4.35% (94%)	-0.03% (48%)	5.80% (96%)	6.44% (100%)	1.09% (78%)
Low	4.80% (96%)	5.04% (96%)	-0.08% (44%)	4.53% (88%)	4.72% (80%)	1.05% (84%)
Medium	5.89% (96%)	5.48% (82%)	0.06% (56%)	5.32% (90%)	4.96% (80%)	1.16% (78%)
High	4.42% (94%)	3.94% (80%)	-0.21% (48%)	4.10% (82%)	3.41% (80%)	0.93% (76%)
Very High	5.12% (100%)	4.96% (94%)	0.37% (66%)	5.09% (84%)	4.82% (84%)	1.03% (78%)

Note: We calculate the value improvement from 50 splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Split-only widens in both predictive accuracy and targeting performance (see Web Appendix E.5 and F.5). In contrast, the Non-Honest method often lowers overall accuracy and fails to improve targeting value, though it may still yield modest gains in AUTOC.

6.7 Robustness Checks

We present robustness checks using alternative model classes for the Default approach and initial CATE models in Web Appendix E.6 and F.6. The Proposed method consistently outperforms the Default in both predictive accuracy and treatment prioritization, and generally outperforms the Split-Only and Non-Honest methods, especially in high-privacy settings. While the Split-Only method also provides improvement, the Non-Honest method often provides little to no improvement.[9] In sum, the robustness checks highlight the importance of

[9] The only minor exceptions are found in both datasets when using R-XGBoost and DR-XGBoost models, where there is no significant difference in value improvement across methods because nearly all approaches predict positive CATEs for 99.9% of customers. Nevertheless, we observe significant gains in both RMSE and AUTOC, with the Proposed method performing best. This indicates that targeting decisions could still be improved, especially when intervention costs are considered or when only a subset of customers can be targeted.

honesty in model calibration and demonstrate the consistent performance improvement of the PROPOSED method across a variety of model classes.

7 Discussion and Future Directions

Differential privacy has gained significant attention and adoption in recent years as a robust framework for protecting individual data while still allowing individual-level data collection. Yet its impact on the precision of existing CATE models and on firms' targeting capabilities remains largely unexplored. This paper takes a first step in addressing this gap and shows, both theoretically and empirically, that the noise introduced by DP can substantially reduce the predictive accuracy of CATE models and weaken targeting effectiveness. These challenges present significant hurdles for organizations that depend on privacy-protected data to deliver targeted interventions.

While existing methods for correcting measurement error bias have been developed, they are not well-suited for CATE estimation under DP protection. To address their limitation, we propose a novel model calibration approach that systematically refines CATE predictions using an unbiased proxy, while preserving formal privacy guarantees. A central feature of our method is the incorporation of the *honesty* principle, which helps prevent overfitting to the proxy's approximation errors, an issue that can be especially severe under DP noise. We further provide a formal error reduction guarantee: when the calibration sample is sufficiently large, the proposed algorithm leads to a provable improvement in prediction accuracy.

Our study has several limitations that point to promising directions for future research. First, while we quantify the impact of DP on CATE estimation and targeting performance, we are unable to directly measure the actual profitability loss due to data limitations. A more comprehensive analysis, such as evaluating firm-level performance before and after DP implementation across multiple companies, could yield deeper insights into the economic costs of privacy protection and inform policy design. Besides, our focus is limited to CATE prediction, but DP affects a broad range of marketing tasks. Future research could examine its implications for other areas, such as demand estimation, personalization, and product recommendation, where privacy constraints may similarly alter model performance and decision quality.

Second, our focus on treatment prioritization based on predicted CATEs reflects only one approach to targeting. The effects of differential privacy on alternative targeting strategies and possible remedies remain underexplored. For example, policy learning approaches (e.g., Kitagawa and Tetenov 2018, Athey and Wager 2021) aim to learn treatment rules that classify individuals based on whether their treatment effect exceeds a certain decision threshold. In these settings, minimizing mean squared error is not directly aligned with the learning objective of these methods, which is typically cost-sensitive classification. A promising direction for future research is to modify the learning objectives used in boosting-based calibration procedures by incorporating alternative loss functions, such as cross-entropy loss for classification (Mao et al. 2023) or expected profit loss for targeted sampling (Chen et al. 2024). Tailoring the calibration process to different objectives could better align model training with firms' real-world decision-making goals.

Third, our approach relies on access to an unbiased DR score for the true CATE, which is achievable when the decision-maker conducts a fully randomized experiment or assigns treatments based on a known propensity score and DP-protected covariates. However, in some real-world settings, such as advertising campaigns managed by third-party platforms, the true propensity score may be unknown, and only noisy, DP-protected covariates are shared with the focal company, even though targeting is based on the true covariates. In such cases, the estimated DR scores may be biased, especially when covariates are perturbed by differential privacy noise. This bias undermines the validity of the calibration procedure, as the DR score no longer serves as a reliable proxy for the true CATE. Addressing this limitation will require bias correction methods for propensity score estimation under privacy constraints. Developing such correction techniques remains an important direction for future research.

Finally, although our method is motivated by the challenge of improving CATE estimation under DP protection, it also provides a general framework for handling noisy or error-prone experimental data more broadly. Our simulations and empirical analyses demonstrate that the proposed approach not only enhances predictive accuracy under strict privacy constraints but also improves targeting performance even in the absence of DP noise. This positions our method as a general tool for firms seeking to target interventions when data quality is affected by noise or measurement error. While existing measurement-error correction techniques may also improve accuracy in non-DP settings, a systematic comparison with our approach lies beyond the scope of this study and represents an important direction for future research.

References

Abel AB (2018) Classical measurement error with several regressors. *Working Paper* .

Agarwal A, Singh R (2021) Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780* .

Amin K, Kulesza A, Munoz A, Vassilvtiskii S (2019) Bounding user contributions: A bias-variance tradeoff in differential privacy. *International Conference on Machine Learning*, 263–271 (PMLR).

Andrews I, Stock JH, Sun L (2019) Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics* 11(1):727–753.

Apple (2017) Learning with privacy at scale. Technical report, Apple.

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research* 55(1):80–98.

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Athey S, Wager S (2021) Policy learning with observational data. *Econometrica* 89(1):133–161.

Battistin E, Chesher A (2014) Treatment effect estimation with covariate measurement error. *Journal of Econometrics* 178(2):707–715.

Carroll RJ, Roeder K, Wasserman L (1999) Flexible parametric measurement error models. *Biometrics* 55(1):44–54.

Chen X, Hong H, Tamer E (2005) Measurement error models with auxiliary data. *The Review of Economic Studies* 72(2):343–366.

Chen YW, Ascarza E, Netzer O (2024) Policy-aware experimentation: Strategic sampling for optimized targeting policies. *Columbia Business School Research Paper* (5044549).

Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I (2018) Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india.

Chesher A (1991) The effect of measurement error. *Biometrika* 78(3):451–462.

Cohen A (2022) Attacks on deidentification's defenses. *31st USENIX Security Symposium (USENIX Security 22)*, 1469–1486.

Devriendt F, Moldovan D, Verbeke W (2018) A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data* 6(1):13–41.

Ding B, Kulkarni J, Yekhanin S (2017) Collecting telemetry data privately. *Advances in Neural Information Processing Systems* 30.

Duan T, Anand A, Ding DY, Thai KK, Basu S, Ng A, Schuler A (2020) Ngboost: Natural gradient boosting for probabilistic prediction. *International conference on machine learning*, 2690–2700 (PMLR).

Dwork C (2006) Differential privacy. *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, 1–12 (Springer).

Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M (2006a) Our data, ourselves: Privacy via distributed noise generation. *Annual international conference on the theory and applications of cryptographic techniques*, 486–503 (Springer).

Dwork C, McSherry F, Nissim K, Smith A (2006b) Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, 265–284 (Springer).

Dwork C, Roth A, et al. (2014) The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9(3–4):211–407.

Ellickson PB, Kar W, Reeder III JC (2022) Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science* .

Erlingsson Ú, Pihur V, Korolova A (2014) Rappor: Randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 1054–1067.

European Data Protection Board (2025) Guidelines 01/2025 on pseudonymisation. URL `https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-012025-pseudonymisation_en`, accessed: 2025-02-04.

Farrell MH, Liang T, Misra S (2021) Deep neural networks for estimation and inference. *Econometrica* 89(1):181–213.

Federal Trade Commission (2024) No, hashing still doesn't make your data anonymous. URL `https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/no-hashing-still-doesnt-make-your-data-anonymous`, accessed: 2025-02-04.

Fernández-Loría C, Loría J (2025) The amenability framework: Rethinking causal ordering without estimating causal effects. *arXiv preprint arXiv:2206.12532* .

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* 1189–1232.

Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M (2011) Doubly robust estimation of causal effects. *American journal of epidemiology* 173(7):761–767.

Ghazi B, Harrison C, Hosabettu A, Kamath P, Knop A, Kumar R, Leeman E, Manurangsi P, Sahu V (2024) On the differential privacy and interactivity of privacy sandbox reports. *arXiv preprint arXiv:2412.16916* .

Ghosh A, Roughgarden T, Sundararajan M (2009) Universally utility-maximizing privacy mechanisms. *Proceedings of the Forty-first Annual ACM symposium on Theory of Computing*, 351–360.

Gopalan P, Kalai AT, Reingold O, Sharan V, Wieder U (2021) Omnipredictors. *arXiv preprint arXiv:2109.05389* .

Hausman JA, Newey WK, Ichimura H, Powell JL (1991) Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics* 50(3):273–295.

Hébert-Johnson U, Kim M, Reingold O, Rothblum G (2018) Multicalibration: Calibration for the (computationally-identifiable) masses. *International Conference on Machine Learning*, 1939–1948 (PMLR).

Hillstrom K (2008) The minethatdata e-mail analytics and data mining challenge. Blog entry, Kevin Hillstrom: MineThatData, URL `https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html`.

Hitsch GJ, Misra S, Zhang WW (2024) Heterogeneous treatment effects and optimal targeting policy evaluation. *Quantitative Marketing and Economics* 22(2):115–168.

Hu Y, Ridder G (2012) Estimation of nonlinear models with mismeasured regressors using marginal information. *Journal of Applied Econometrics* 27(3):347–385.

Hu Y, Schennach SM (2008) Instrumental variable treatment of nonclassical measurement error models. *Econometrica* 76(1):195–216.

Huang TW, Ascarza E (2024) Doing more with less: Overcoming ineffective long-term targeting using short-term signals. *Marketing Science* .

Huang TW, Ascarza E, Israeli A (2024) Incrementality representation learning: Synergizing past experiments for intervention personalization. *Available at SSRN 4859809* .

Kalantari K, Sankar L, Sarwate AD (2018) Robust privacy-utility tradeoffs under differential privacy and hamming distortion. *IEEE Transactions on Information Forensics and Security* 13(11):2816–2830.

Kane K, Lo VS, Zheng J (2014) Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics* 2:218–238.

Kasiviswanathan SP, Lee HK, Nissim K, Raskhodnikova S, Smith A (2011) What can we learn privately? *SIAM Journal on Computing* 40(3):793–826.

Kennedy EH (2023) Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17(2):3008–3049.

Kenny CT, Kuriwaki S, McCartan C, Rosenman ET, Simko T, Imai K (2021) The use of differential privacy for census data and its impact on redistricting: The case of the 2020 us census. *Science Advances* 7(41):eabk3283.

Kern C, Kim MP, Zhou A (2024) Multi-accurate cate is robust to unknown covariate shifts. *Transactions on Machine Learning Research* .

Kim MP, Kern C, Goldwasser S, Kreuter F, Reingold O (2022) Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences* 119(4):e2108097119.

Kitagawa T, Tetenov A (2018) Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616.

Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Lee Lf, Sepanski JH (1995) Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association* 90(429):130–140.

Lemmens A, Roos JM, Gabel S, Ascarza E, Bruno H, Gordon BR, Israeli A, Feit EM, Mela CF, Netzer O (2025) Personalization and targeting: How to experiment, learn, & optimize. *International Journal of Research in Marketing* (Forthcoming).

Leng Y, Dimmery D (2024) Calibration of heterogeneous treatment effects in randomized experiments. *Information Systems Research* 35(4):1721–1742.

Li T, Vuong Q (1998) Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis* 65(2):139–165.

Lichtenstein S, Fischhoff B, Phillips LD (1977) Calibration of probabilities: The state of the art. *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*, 275–324 (Springer).

Mao A, Mohri M, Zhong Y (2023) Cross-entropy loss functions: Theoretical analysis and applications. *International conference on Machine learning*, 23803–23828 (PMLR).

Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111–125 (IEEE).

Near JP, Darais D, Lefkovitz N, Howarth G, et al. (2023) Guidelines for evaluating differential privacy guarantees. *National Institute of Standards and Technology, Tech. Rep* .

Nie X, Wager S (2021) Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108(2):299–319.

Niu F, Nori H, Quistorff B, Caruana R, Ngwe D, Kannan A (2022) Differentially private estimation of heterogeneous causal effects. *Conference on Causal Learning and Reasoning*, 618–633 (PMLR).

Pal M (1980) Consistent moment estimators of regression coefficients in the presence of errors in variables. *Journal of Econometrics* 14(3):349–364.

Platt J, et al. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10(3):61–74.

Ponte G, Wieringa J, Boot T (2025) Differentially private targeting strategies, working paper.

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89(427):846–866.

Rößler J, Schoder D (2022) Bridging the gap: A systematic benchmarking of uplift modeling and heterogeneous treatment effects methods. *Journal of Interactive Marketing* 57(4):629–650.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688.

Sarwate AD, Sankar L (2014) A rate-disortion perspective on local differential privacy. *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 903–908 (IEEE).

Schennach S (2022) Measurement systems. *Journal of Economic Literature* 60(4):1223–1263.

Schennach SM (2004) Estimation of nonlinear models with measurement error. *Econometrica* 72(1):33–75.

Schennach SM (2007) Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica* 75(1):201–239.

Schennach SM, Hu Y (2013) Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association* 108(501):177–186.

Semenova V, Chernozhukov V (2021) Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* 24(2):264–289.

Sepanski JH, Carroll RJ (1993) Semiparametric quasilikelihood and variance function estimation in measurement error models. *Journal of Econometrics* 58(1-2):223–256.

Showkatbakhsh M, Karakus C, Diggavi S (2018) Privacy-utility trade-off of linear regression under random projections and additive noise. *2018 IEEE International Symposium on Information Theory (ISIT)*, 186–190 (IEEE).

Simester D, Timoshenko A, Zoumpoulis SI (2020) Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science* 66(6):2495–2522.

Smith AN, Seiler S, Aggarwal I (2021) Optimal price targeting. *Available at SSRN 3975957* .

Sweeney L (1997) Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* 25(2-3):98–110.

Tullii M, Gaucher S, Richard H, Diemert E, Perchet V, Rakotomamonjy A, Calauzènes C, Vono M (2024) Position paper: Open research challenges for private advertising systems under local differential privacy .

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Warner SL (1965) Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 60(309):63–69.

Whitehouse J, Jung C, Syrgkanis V, Wilder B, Wu ZS (2024) Orthogonal causal calibration. *arXiv preprint arXiv:2406.01933* .

Wolter KM, Fuller WA (1982) Estimation of nonlinear errors-in-variables models. *The Annals of Statistics* 539–548.

Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S (2024) Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association* 1–14.

Yang J, Eckles D, Dhillon P, Aral S (2023) Targeting for long-term outcomes. *Management Science* .

Yang M, McFowland III E, Burtch G, Adomavicius G (2022) Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS Journal on Data Science* .

Yoganarasimhan H, Barzegary E, Pani A (2022) Design and evaluation of optimal free trials. *Management Science* .

Zhong H, Bu K (2022) Privacy-utility trade-off. *arXiv preprint arXiv:2204.12057* .

Web Appendix

Web Appendix A Bias and Variance of CATE Models with DP Data

In this appendix, we characterize the impact of DP protection on the prediction error of commonly-used CATE models. We begin by decomposing the MSE between the predicted and true CATE for a new customer (indexed by new) into squared bias and variance components:

$$\mathbb{E}\left[(\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_{\text{new}}))^2\right] = \underbrace{\mathbb{E}^2\left[\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \tau(\mathbf{X}_{\text{new}})\right]}_{\text{Squared Bias}} + \underbrace{\text{Var}\left[\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \widetilde{\mathcal{D}})\right]}_{\text{Variance}},$$

where $\widetilde{\mathcal{D}}$ is the DP-protected version of the experiment data $\mathcal{D}$.

We analyze each component separately and derive expressions for the bias and variance under common CATE modeling assumptions. Below, we first present the intuition using analytical derivations and then demonstrate the issues caused by DP using a simple simulation.

Web Appendix A.1 Analytical Characterization

To characterize the impact of DP, we apply the Taylor approximation, a common technique for analyzing covariate measurement error in the literature (Chesher 1991, Battistin and Chesher 2014). Let $\widehat{\tau}(\cdot \mid \mathcal{D})$ denote the CATE model trained on the non-private dataset. Define the injected noise into the covariates of the new individual as $\eta_{\mathbf{X}_{\text{new}}}$ and into the training data as $\boldsymbol{\eta}_{\mathcal{D}}$. The second-order approximation then yields:

$$\begin{aligned}\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}}|\widetilde{\mathcal{D}}) \approx\ & \widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D}) + \\ & \underbrace{\partial_{\mathbf{X},\mathcal{D}}\, \widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D}) \cdot (\eta_{\mathbf{X}_{\text{new}}}, \boldsymbol{\eta}_{\mathcal{D}})}_{\equiv A_1} + \underbrace{\frac{1}{2}\, \partial^2_{\mathbf{X},\mathcal{D}}\, \widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D}) \cdot (\eta^2_{\mathbf{X}_{\text{new}}}, \boldsymbol{\eta}^{\circ 2}_{\mathcal{D}})}_{\equiv A_2},\end{aligned} \tag{App-1}$$

where $\partial^k_{\mathbf{X},\mathcal{D}}\, \widehat{\tau}$ denotes the k-th order directional derivative of the CATE function with respect to covariates and training data, and $Z^{\circ k}$ indicates element-wise exponentiation of matrix or vector Z to the power k.

Bias. Finally, taking the expectation of the Taylor approximation in (App-1), the bias introduced by DP noise can be approximated as:

$$\mathbb{E}\left[\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}}) - \widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D})\right] \approx \frac{1}{2}\, \mathbb{E}\left[\partial^2_{\mathbf{X},\mathcal{D}}\, \widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D})\right]\ \mathbf{V}(\eta_{\mathbf{X}_{\text{new}}}, \boldsymbol{\eta}_{\mathcal{D}}),$$

where $\mathbf{V}(\eta_{\mathbf{X}_{\text{new}}}, \boldsymbol{\eta}_{\mathcal{D}})$ is a diagonal matrix whose entries correspond to the variances of noise added to each covariate and training example.

This result highlights two key drivers of bias under DP. First, the bias increases with the level of noise injected for privacy protection. Stricter privacy requirements lead to larger noise $(\eta_{\mathbf{X}_{\text{new}}}, \boldsymbol{\eta}_{\mathcal{D}})$ and thus greater bias. Second, and importantly, the magnitude of the bias *depends on the sensitivity of the CATE model to noise perturbations*, captured by the second-order derivative $\partial^2_{\mathbf{X},\mathcal{D}}\,\widehat{\tau}$. This stands in contrast to classical linear models where measurement error generally induces a uniform attenuation bias (Abel 2018).[1] Notably, bias under DP is not only amplified for more flexible models but can also be significant in highly regularized models, depending on how the model responds to perturbations in the input or training data. In addition, the bias is not uniform across the covariate space. This means that, even when average predictive metrics such as mean squared error (MSE) appear satisfactory, treatment effect rankings may be systematically distorted. This poses a particular challenge for targeting tasks, where accurate ranking is often more critical than overall error.

Variance. To analyze variance, we again use the Taylor expansion in (App-1). The difference in variance due to DP protection is:

$$\text{Var}\big[\widehat{\tau}(\widetilde{\mathbf{X}}_{\text{new}} \mid \widetilde{\mathcal{D}})\big] - \text{Var}\big[\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D})\big] \approx$$

$$\text{Var}\left[A_1\right] + \text{Var}\left[A_2\right] + \text{Cov}\left[\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D}), A_1\right] + \text{Cov}\left[\widehat{\tau}(\mathbf{X}_{\text{new}} \mid \mathcal{D}), A_2\right] + \text{Cov}\left[A_1, A_2\right].$$

While the variance of the first- and second-order terms is always non-negative, the covariance terms, which capture the relationship between the model's sensitivity to first- and second-order perturbations and the baseline predicted CATE, can be negative. As a result, DP protection does not necessarily lead to an increase in overall model variance.

This result can be understood intuitively. On one hand, adding noise through DP increases the variability of CATE predictions by perturbing the inputs, especially when the model is highly flexible and tries to capture fine-grained heterogeneity. This effect corresponds to the non-negative variance terms in the decomposition. On the other hand, when the noise level is high, it can obscure true variation in treatment effects, leading the model to shrink large

[1] The Taylor approximation results in Chesher (1991) focus on the difference between $\tau(\widetilde{\mathbf{X}}_{\text{new}})$ and $\tau(\mathbf{X}_{\text{new}})$, but do not account for the influence of model structure or estimation procedures. In our setting, these factors are essential to understanding the full impact of DP.

predicted CATE values (from the non-private setting) toward the average treatment effect. This shrinkage behavior is reflected in negative covariance terms and can offset the added variance. As a result, the overall predictive variance may actually decrease, leading to near-constant CATE estimates across customers. This explains why DP protection does not always increase model variance.

Web Appendix A.2 A Simulation Example

We now empirically illustrate these findings through a simple simulation exercise. Suppose a company conducts a randomized controlled experiment with two treatment conditions ($W_i \in \{0,1\}$) on a population of individuals (indexed by i). Let $Y_i \in \mathbb{R}$ represent the outcome of interest, and let X_i denote the pre-treatment covariate that captures heterogeneity in treatment effects. We generate simulated experimental data with $N = 2,000$ observations using the following data-generating process:

$$\begin{aligned} &Y_i = X_i^2 + W_i \cdot \tau(X_i) + e_i, \quad \tau(X_i) = (2 + X_i - 0.5X_i^2), \\ &W_i \sim \text{Ber}(0.5), \quad X_i \sim \text{Unif}(-2,2), \quad e_i \sim \mathcal{N}(0,1). \end{aligned} \tag{App-2}$$

For illustration, we consider two scenarios separately: covariate protection under DP and outcome protection under DP. In the first case, when covariates are protected using the Laplace mechanism, the decision-maker observes a noisy version of the true covariates, $\widetilde{X}_i = X_i + \eta_i^X$, where η_i^X is noise drawn from a Laplace(σ) distribution. In the second case, when the outcome of interest is protected under DP, the decision-maker observes only $\widetilde{Y}_i = Y_i + \eta_i^Y$ instead of the true outcome Y_i, with η_i^Y similarly drawn from a Laplace(σ) distribution.

To assess the effect of DP noise on CATE estimation and prediction, we generate 100 replications under each scenario and predict CATEs using three popular models:

1. *Causal Forest* (Wager and Athey 2018): We implement causal forest using the `grf` package with default parameters.

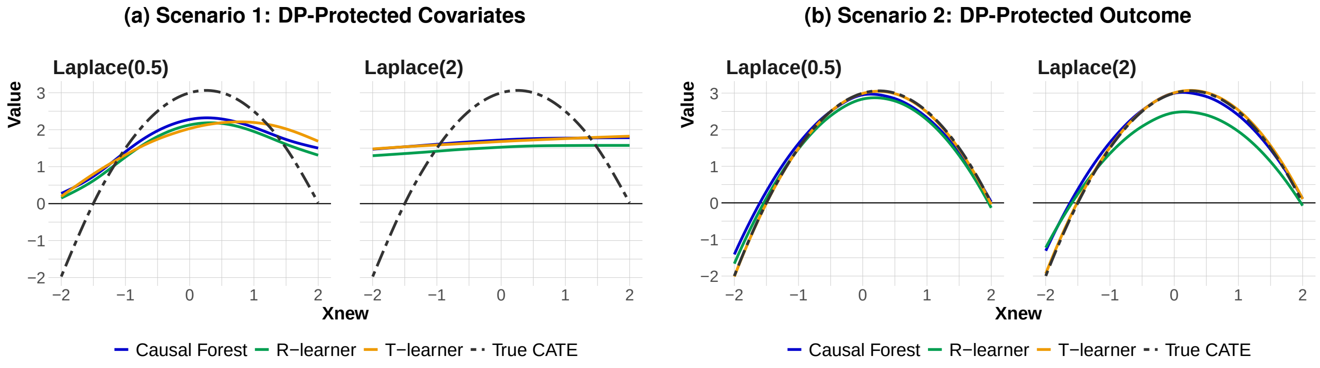
2. *R-XGBoost models* (Nie and Wager 2021): We implement R-XGBoost models for both the conditional mean outcome and CATE models. Hyperparameter tuning follows the methods implemented by the R-package `rlearner` provided by Nie and Wager (2021).

3. *T-learner with XGBoost models*: We implement the T-learner with correctly specified models. Specifically, for each treatment condition, we estimate a regression model of the outcome on covariates using second-order polynomial specifications. The CATE is then calculated by subtracting the predicted outcome from the control group model from that of the treatment group model.

For each value X_{new}, we calculate prediction bias by averaging the predicted CATE across 100 replications and comparing it to the true CATE. To approximate variance, we compute the variance of the predicted CATE for a given covariate value across the 100 replications.

Bias results. Figure App-1 presents the bias of predicted CATEs from different models under varying noise levels. Several key patterns emerge. First, as expected, bias in predicted CATE increases with stronger privacy protection. Second, as described in Section Web Appendix A.1, DP induced bias is model-dependent. When only the covariate is protected by DP (Figure App-1a), all models shrink the predicted CATE toward the average treatment effect, but the magnitude and pattern of bias vary across models. When the outcome is protected by DP (Figure App-1b), the differences are more noticeable, especially under high noise levels (right panel, Figure App-1b), R-learner exhibits significantly higher bias compared to the other two approaches. Third, the direction and magnitude of bias in CATE estimation depend on the covariate values. For $X_{\text{new}} > 1.5$ and $X_{\text{new}} < -1$, all models overestimate the CATE (exacerbation bias), whereas for $-1 < X_{\text{new}} < 1.5$, they underestimate it (attenuation bias). This variation in bias direction occurs because DP noise in CATE estimation pulls estimates toward the average treatment effect rather than toward zero.

Figure App-1: Bias of CATE Predictions in the Simulation Setting

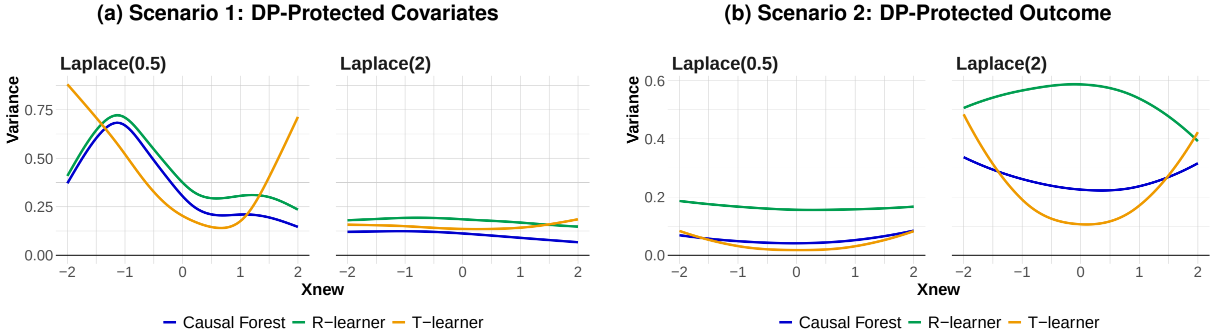


(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Variance results. Figure App-2 presents the variance of CATE predictions across different covariate values and models. Similar to the results when analyzing the bias, the variance of CATE predictions is model-dependent and heterogeneous across covariate values. Moreover, when the covariate is protected by DP (Figure App-2a), we observe a counterintuitive result: increasing the noise level decreases the variance. This occurs because all CATE models underestimate treatment effect heterogeneity, producing predictions that cluster around the average treatment effect. Conversely, when the outcome is protected by DP (Figure App-3b), the variance of CATE predictions increases substantially as the noise level rises.

Figure App-2: Variance of CATE Predictions in the Simulation Setting



Web Appendix B Additional Literature Review

This appendix provides supplementary literature context to support the main text. It reviews classical approaches to correcting measurement error and discusses their limitations in the context of differential privacy. It also summarizes related work on sample-splitting strategies for honest estimation and model calibration, in order to highlight the conceptual differences between existing approaches and our proposed method.

Web Appendix B.1 Classical Measurement Error Bias Correction

Table App-1 summarizes canonical approaches for addressing measurement error bias in covariates, including repeated measurements, instrumental variables, variable calibration, distribution-based corrections, and higher-order moments calibration. For each method, the table outlines the core identification strategy, the typical assumptions required for validity, and the key limitations that arise when these methods are applied in DP settings.

Table App-1: Comprehensive Overview of Solutions for Measurement Error Bias

Study	Flexible with Many Covariates	Model Specification	Estimation Method
Repeated Measurements of Variables			
Hausman et al. (1991a)	Yes	Polynomial	Least Square
Li (2002)	Yes	Nonlinear	Least Square
Schennach (2004)	Yes	Polynomial	GMM
Agarwal and Singh (2021)	Yes	Linear	PCA + Least Square
Instrumental Variables			
Hausman et al. (1991)	No	Polynomial	GMM
Newey (2001)	No	Nonlinear	GMM
Schennach (2007)	No	Nonlinear	GMM
Hu and Schennach (2008)	No	Nonlinear	MLE
Additional Clean Data			
Bound et al. (1989)	Yes	Linear	OLS
Sepanski and Carroll (1993)	No	Nonlinear	OLS
Chen et al. (2005)	Yes	Nonlinear	GMM
Hu and Ridder (2012)	No	Nonlinear	Deconvolution + MLE
Information of Noise Distribution			
Wolter and Fuller (1982)	No	Polynomial	Parametric MLE
Fan and Truong (1993)	No	Nonlinear	Deconvolution + Kernel Estimator
Bonhomme and Robin (2010)	No	Linear	Deconvolution
Higher-order Moments in Data			
Pal (1980)	No	Linear	GMM
Erickson and Whited (2002)	Yes	Linear	GMM
Schennach and Hu (2013)	No	Known functional form	MLE

We also discuss these limitations in greater detail below and explain why existing methods, though effective in traditional settings, are not well suited to the privacy-preserving context studied in this paper.

Repeated Measurements of Variables. These techniques aim to mitigate random noise by averaging multiple measurements of the same variable, reducing the impact of individual errors. While they are theoretically capable of handling complex functional forms, they become impractical in high-dimensional settings due to the exponential increase in the number of repeated measurements required as the dimensionality of the data grows. This requirement can be resource-intensive or even infeasible in practice. Additionally, these techniques pose significant privacy risks, as repeated data collection increases the likelihood of re-identifying individuals. This undermines privacy protections, making such methods unsuitable for use in contexts requiring strict privacy constraints.

Instrumental Variables. These techniques use additional instrumental variables (IVs) that are correlated with the noisy covariate but uncorrelated with the measurement error or outcome, serving as proxies to correct bias. However, identifying valid instruments in high-dimensional settings is challenging, as suitable instruments are needed for many covariates. Additionally, IV approaches often rely on linear or parametric assumptions, limiting their ability to capture heterogeneity. Their integration with specific models also reduces model-agnostic flexibility. Moreover, protecting instrumental variables under LDP can exacerbate the weak instruments problem, further complicating the estimation process.

Additional Clean Data. These methods calibrate the model using a separate dataset with noise-free measurements to correct estimates from the larger dataset. It is theoretically well-suited in high-dimensional settings with flexible functional forms and model-agnostic applicability. However, it raises significant privacy concerns because the clean dataset itself may contain sensitive information.

Noise Distribution Information. These methods rely on maximum likelihood or deconvolution kernel estimators to correct bias by incorporating knowledge of the noise distribution. However, they are not well-suited to our setting for several reasons. First, they require strong assumptions about the form of the noise distribution, which may not hold in practice — especially under complex or unknown DP mechanisms. Second, they are often developed for low-dimensional or parametric models and may not scale well to flexible, high-dimensional CATE models commonly used in marketing applications. Third, these approaches focus primarily on correcting noise in covariates and do not account for noise in the outcome variable, which is equally important for CATE prediction under DP protection.

Higher-order Moments. These methods use overidentifying information from higher-order moments to correct measurement error bias, typically under the assumption that the measurement errors satisfy certain independence conditions. However, they often rely on strong distributional and functional form assumptions, do not address noise in the outcome variable, and fail to tackle the increased variance introduced by privacy-preserving noise.

Web Appendix B.2 Existing Sample Splitting Approaches in Causal Inference

Table App-2 outlines key distinctions between our proposed method and the sample-splitting strategies employed in causal forests (Athey and Imbens 2016, Wager and Athey 2018) and transformed outcome regressions for CATE estimation (Chernozhukov et al. 2018, Nie and Wager 2021, Kennedy 2023, Whitehouse et al. 2024). In causal forests, sample splitting is primarily used to separate the *partitioning* stage from the *treatment effect estimation* within each leaf, thereby reducing overfitting by ensuring that model selection and estimation are conducted on independent data. In transformed outcome regressions, sample splitting serves a different purpose: it separates the estimation of nuisance components (e.g., conditional outcome and propensity score models) from the construction of proxy CATE scores. This ensures that errors in nuisance models do not bias the final CATE estimates.

Table App-2: Key Distinctions Between Our Method and Existing Sample-splitting Approaches

Method	Where Does Splitting Occur?	Primary Goal
Causal Tree / Forest	Partition tree structure vs. Estimate leaf-level treatment effect.	Decouple treatment effect estimation from tree partitioning
Transformed Outcome Regression	Estimate nuisance models vs. Construct the proxy score and CATE models.	Eliminate bias from nuisance estimates
Proposed Honest Calibration	1. Train the initial CATE model vs. Calibrate the initial model vs. Validate the calibration process. 2. Estimate calibration model vs. Determine the step size.	Avoid overfitting to the noisy proxy during the calibration procedure

Our setting addresses a distinct challenge: calibrating an already-fitted CATE model using a noisy proxy for the true treatment effect—a problem that naturally arises under differential privacy. Theoretically, the goal of our sample splitting strategy is to ensure that the calibration target (i.e., the residual between the DR score and the predicted CATE) is not correlated with the prediction error of the initial CATE model. This differs from other sample splitting strategies in causal inference. For example, Causal Forests use sample splitting to decouple tree construction from treatment effect estimation, reducing the risk of overfitting to observed heterogeneity. Similarly, in transformed outcome regressions, sample splitting ensures that bias in the nuisance models does not have a first-order effect on the second-stage CATE model.

Furthermore, we propose a new structured sample-splitting strategy. Specifically, we divide the data into three disjoint subsets: one for training the initial CATE model, one for calibration, and one for validation. This additional layer of separation ensures that the calibration step is statistically independent from both model training and evaluation. Furthermore, within our boosting framework, we introduce an additional honesty constraint by ensuring that the construction of each calibration model is independent of the step-size determination. This level of modularity introduces a novel safeguard against overfitting—one not present in standard boosting-based CATE methods—and enhances model robustness under privacy constraints.

Web Appendix C Theoretical Guarantees

This appendix presents the formal proof of Proposition 1 and establishes the theoretical foundation for the proposed honest model calibration method. Web Appendix C.1 begins by proving Proposition 1, which shows that the DR score serves as an unbiased proxy for the true conditional average treatment effect (CATE) in randomized experiments or in settings where treatment assignment probabilities are known (e.g, the firms using a known treatment rule based on DP-protected covariates).

Building on this result, Web Appendix C.2 provides a high-probability lower bound on the gain from performing honest model calibration using an unbiased signal such as the DR score. This lower bound consists of three components: (i) the observed reduction in mean squared residuals between the initial prediction $\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_j)$ and the DR score $\check{\tau}_j$ on the calibration set, (ii) a complexity penalty that reflects the potential of the calibration model $\widehat{c}$ to overfit noise, and (iii) an overfitting error term that converges to zero as the calibration sample size increases. The proof relies critically on sample splitting between the training and calibration sets. This independence ensures that standard concentration bounds from (Wainwright 2019) can be applied to control the generalization error on the calibration set.

Next, Web Appendix C.2 derives an upper bound on the complexity of the proposed additive boosting model class introduced in Section 4.3. This bound depends on the calibration model used. When the calibration model is relatively simple—such as a linear model or a shallow decision tree—the complexity penalty decreases more quickly with sample size, meaning that the observed reduction in mean squared residuals more closely reflects the true improvement in predictive accuracy.

Finally, Web Appendix C.4 explain why non-honest calibration—where the same data are used for both model training and calibration—may fail to improve performance. This is particularly likely when the initial model $\widehat{\tau}^{[0]}$ is already highly expressive. In such cases, recalibrating on the same data adds little value and may instead increase the risk of overfitting.

Web Appendix C.1 DR Score as an Unbiased Signal

Proposition 1 (Unbiased Signal)

Under standard assumptions of positivity, unconfoundedness, and no interference, the DR score is an unbiased proxy for the true CATE when the additive noise introduced by the DP mechanism has mean zero. That is, $\mathbb{E}\left[\check{\tau}_i \mid \mathbf{X}_i\right] = \tau(\mathbf{X}_i)$*, as long as the score is computed using the true propensity score, even if both the covariates and outcomes are DP-protected.*

Proof. First, we rearrange the DR score as follows:

$$\check{\tau}_j = \underbrace{\left[\frac{W_i\widetilde{Y}_i}{e_i} - \frac{(1-W_i)\widetilde{Y}_i}{1-e_i}\right]}_{\text{IPW term}} \quad + \quad \underbrace{\left[\widehat{\mu}_1(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \cdot \left(1 - \frac{W_i}{e_i}\right) - \widehat{\mu}_0(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \left(1 - \frac{1-W_i}{1-e_i}\right)\right]}_{\text{Noise term}} ,$$

where $\widehat{\mu}_1(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}})$ and $\widehat{\mu}_0(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}})$ are the estimated conditional mean outcome models for each treatment condition. The conditional expectation for the IPW term can be written as

$$\begin{aligned}
&\mathbb{E}\left[\frac{W_i\widetilde{Y}_i}{e_i} - \frac{(1-W_i)\widetilde{Y}_i}{1-e_i} \mid \mathbf{X}_i\right] \\
&\quad = \underbrace{\mathbb{P}(W_i = 1 \mid \mathbf{X}_i)}_{=e_i} \cdot \mathbb{E}\left[\frac{\widetilde{Y}_i}{e_i} \mid W_i = 1, \mathbf{X}_i\right] \quad + \quad \underbrace{\mathbb{P}(W_i = 0 \mid \mathbf{X}_i)}_{=1-e_i} \cdot \mathbb{E}\left[\frac{\widetilde{Y}_i}{1-e_i} \mid W_i = 0, \mathbf{X}_i\right] \\
&\quad = \mathbb{E}[\widetilde{Y}_i \mid W_i = 1, \mathbf{X}_i] - \mathbb{E}[\widetilde{Y}_i \mid W_i = 0, \mathbf{X}_i] \\
&\quad = \mathbb{E}[Y_i \mid W_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i \mid W_i = 0, \mathbf{X}_i] \quad \text{(since the additive noise has mean zero)} \\
&\quad = \mathbb{E}[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}\left[Y_i(0) \mid \mathbf{X}_i\right] = \tau(\mathbf{X}_i) \quad \text{(by the unconfoundedness assumption).}
\end{aligned}$$

The conditional expectation for the noise term can be written as

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mu}_1(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \cdot \left(1 - \frac{W_i}{e_i}\right) - \widehat{\mu}_0(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \left(1 - \frac{1-W_i}{1-e_i}\right) \mid \mathbf{X}_i\right] = \\
&\quad \mathbb{E}\left[\widehat{\mu}_1(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \mid \mathbf{X}_i\right] \cdot \underbrace{\left(1 - \frac{\mathbb{E}[W_i|\mathbf{X}_i]}{e_i}\right)}_{=0 \text{ since } \mathbb{E}[W_i|\mathbf{X}_i]=e_i} + \mathbb{E}\left[\widehat{\mu}_0(\widetilde{\mathbf{X}}_i \mid \widetilde{\mathcal{D}}) \mid \mathbf{X}_i\right] \cdot \underbrace{\left(1 - \frac{1-\mathbb{E}[W_i|\mathbf{X}_i]}{1-e_i}\right)}_{=0 \text{ since } \mathbb{E}[W_i|\mathbf{X}_i]=e_i} = 0.
\end{aligned}$$

Combining these two results, we have $\mathbb{E}\left[\check{\tau}_i \mid \mathbf{X}_i\right] = \tau(\mathbf{X}_i)$. ■

Web Appendix C.2 Error Improvement Bound for Honest Calibration

Next, we prove a lemma showing that, for an unseen customer, the expected squared residual between the predicted CATE and an unbiased signal (in our setting, the DR score) is equivalent — up to a constant (i.e., the variance of the unbiased signal) — to the true mean squared error between the predicted and true CATE. This result implies that, the risk bound on the population mean squared residuals also serves as a valid bound on the true CATE prediction error. Therefore, bounding the generalization error of the empirical residuals with respect to the DR score is equivalent to bounding the generalization error of the true CATE.

Lemma 1 (Equivalence of Population Excess Risk and True Risk)
Let $\widehat{\tau}(\widetilde{\mathbf{X}}_i)$ denote a CATE estimate trained on $\widetilde{\mathcal{D}}_{train}$, and let $\tau(\mathbf{X}_i)$ represent the true CATE. We define the population excess risk for a holdout customer from the same distribution as the difference between (i) the mean squared residual between the model's predictions and the unbiased signal $\check{\tau}_i$, and (ii) the mean squared proxy error of the doubly robust score relative to the true CATE:

$$\mathcal{E}(\widehat{\tau},\tau) = \mathbb{E}\big[\big(\widehat{\tau}(\widetilde{\mathbf{X}}_i) - \check{\tau}_i\big)^2\big] - \mathbb{E}\left[\big(\tau(\mathbf{X}_i) - \check{\tau}_i\big)^2\right].$$

Then, this excess risk is exactly equal to the mean squared error between the model's CATE predictions and the true CATE, i.e., $\mathcal{E}(\widehat{\tau},\tau) = \mathbb{E}\big[\big(\widehat{\tau}(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)\big)^2\big]$.

Proof. We first note that

$$\begin{aligned}
&\mathbb{E}\big[\big(\widehat{\tau}(\widetilde{\mathbf{X}}_i) - \check{\tau}_i\big)^2 \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big] \\
&\quad= \mathbb{E}\big[\widehat{\tau}^2(\widetilde{\mathbf{X}}_i) \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big] - 2\mathbb{E}\big[\widehat{\tau}(\widetilde{\mathbf{X}}_i) \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big]\mathbb{E}\big[\check{\tau}_i \mid \mathbf{X}_i\big] + \mathbb{E}\big[\check{\tau}_i^2 \mid \mathbf{X}_i\big] \\
&\quad= \mathbb{E}\big[\widehat{\tau}^2(\widetilde{\mathbf{X}}_i) \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big] - 2\mathbb{E}\big[\widehat{\tau}(\widetilde{\mathbf{X}}_i) \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big]\mathbb{E}\left[\check{\tau}_i \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\right] + \mathbb{E}^2\big[\check{\tau}_i \mid \mathbf{X}_i\big] + \mathrm{Var}\big[\check{\tau}_i \mid \mathbf{X}_i\big] \\
&\quad= \mathbb{E}\big[\widehat{\tau}^2(\widetilde{\mathbf{X}}_i) \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big] - 2\mathbb{E}\big[\widehat{\tau}(\widetilde{\mathbf{X}}_i) \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big] \cdot \tau(\mathbf{X}_i) + \tau^2(\mathbf{X}_i) + \mathrm{Var}\big[\check{\tau}_i \mid \mathbf{X}_i\big].
\end{aligned}$$

since $\check{\tau}_i$ is unbiased and $\check{\tau}_i$ is independent of $\widehat{\tau}$. Similarly, we have

$$\begin{aligned}
\mathbb{E}\left[\big(\tau(\mathbf{X}_i) - \check{\tau}_i\big)^2 \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\right] &= \tau^2(\mathbf{X}_i) - 2\tau(\mathbf{X}_i)\mathbb{E}\left[\check{\tau}_i \mid \mathbf{X}_i\right] + \mathbb{E}^2\big[\check{\tau}_i \mid \mathbf{X}_i\big] + \mathrm{Var}\big[\check{\tau}_i \mid \mathbf{X}_i\big] \\
&= \tau^2(\mathbf{X}_i) - 2\tau^2(\mathbf{X}_i) + \tau^2(\mathbf{X}_i) + \mathrm{Var}\big[\check{\tau}_i \mid \mathbf{X}_i\big] = \mathrm{Var}\big[\check{\tau}_i \mid \mathbf{X}_i\big].
\end{aligned}$$

Combining these two observations together, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\widehat{\tau}(\widetilde{\mathbf{X}}_i) - \check{\tau}_i\right)^2 \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\right] - \mathbb{E}\big[\left(\tau(\mathbf{X}_i) - \check{\tau}_i\right)^2 \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big] \\
&\quad = \mathbb{E}\big[\widehat{\tau}^2(\widetilde{\mathbf{X}}_i) \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big] - 2\mathbb{E}\big[\widehat{\tau}^2(\widetilde{\mathbf{X}}_i) \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\big] \cdot \tau(\mathbf{X}_i) + \tau^2(\mathbf{X}_i) \\
&\quad = \mathbb{E}\left[\left(\widehat{\tau}(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)\right)^2 \mid \widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i\right].
\end{aligned}$$

Taking expectation of the above result over the distribution of $\widetilde{\mathcal{D}}_{\text{train}}, \mathbf{X}_i$ gives the lemma. ■

We now establish an upper bound on the true error improvement for a generic honest calibration algorithm, which combines a baseline predictor $\widehat{\tau}^{[0]}$ trained on a training set with a calibration model h fitted on an independent calibration set.

Theorem App-1 (Error Reduction Bound for Honest Model Calibration)

Assume that the DR score $\check{\tau}_j$ is bounded for every customer. Let $\widetilde{\mathcal{D}}_{train}$ and $\widetilde{\mathcal{D}}_{cal}$ be two independent *i.i.d. samples from the same distribution. Let $\widehat{\tau}^{[0]}$ be a bounded baseline CATE estimator trained* only *on $\widetilde{\mathcal{D}}_{train}$. Assume that the calibration model $\widehat{h} \in \mathcal{H}$ is estimated by minimizing the mean-squared residuals between the prediction and the unbiased signal on $\widetilde{\mathcal{D}}_{\text{cal}}$, i.e.,*

$$\widehat{\tau}^{\text{cal}}(\widetilde{\mathbf{X}}_j) := \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{j \in \widetilde{\mathcal{D}}_{cal}} \left(\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_j) + h(\widetilde{\mathbf{X}}_j) - \check{\tau}_j\right)^2,$$

where $\mathcal{H}$ is a bounded model class. Then, for every $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\mathbb{E}\big[\left(\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)\right)^2\big] - \mathbb{E}\big[\left(\widehat{\tau}^{\text{cal}}(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)\right)^2\big] \;\geqslant\; \Delta_{\text{cal}} - 4\widehat{\mathfrak{R}}_n(\mathcal{H}) - 4B\sqrt{\frac{\log(2/\delta)}{n}},$$

where n denotes the sample size of the calibration data, B is a constant, and Δ_{cal} denotes the empirical improvement in mean squared residuals, i.e.,

$$\Delta_{\text{cal}} := \frac{1}{n} \sum_{j \in \widetilde{\mathcal{D}}_{cal}} \left(\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_j) - \check{\tau}_j\right)^2 - \frac{1}{n} \sum_{j \in \widetilde{\mathcal{D}}_{cal}} \left(\widehat{\tau}^{cal}(\widetilde{\mathbf{X}}_j) - \check{\tau}_j\right)^2 \;\; (\geqslant 0 \textit{ by construction of } \widehat{h}).$$

Proof. Define the empirical mean squared residual over the calibration set as

$$\widehat{L}_n(h) := \frac{1}{n} \sum_{j \in \widetilde{\mathcal{D}}_{\text{cal}}} \left(\widehat{\tau}^0(\widetilde{\mathbf{X}}_j) + h(\widetilde{\mathbf{X}}_j) - \check{\tau}_j\right)^2.$$

Also, define the corresponding population mean squared residual as

$$L(h) := \mathbb{E}\left[\left(\widehat{\tau}(\widetilde{\mathbf{X}}_i) + h(\widetilde{\mathbf{X}}_i) - \check{\tau}_i\right)^2\right].$$

Step 1 (General Rademacher Bound). We apply a standard generalization bound using empirical Rademacher complexity. Specifically, by Theorem 4.10 in Wainwright (2019), for all $h \in \mathcal{H}$ and with probability at least $1-\delta$, the difference between the empirical loss and the population loss can be bounded as:

$$|L(h) - \widehat{L}_n(h)| \;\leqslant\; 2\,\widehat{\mathfrak{R}}_n(\mathcal{H}) + 2B\sqrt{\frac{\log(2/\delta)}{n}}, \qquad \text{(App-3)}$$

where $\widehat{\mathfrak{R}}_n(\mathcal{H})$ denotes the empirical Rademacher complexity of the function class $\mathcal{H}$ evaluated on the calibration sample, and B is the upper bound on the loss (such a bound exists since $\check{\tau}_j$ is bounded and both $\mathcal{T}$ and $\mathcal{H}$ are classes of bounded functions).

Step 2 (Rademacher Bound for Baseline and Calibrated Predictors). We now apply the general bound (App-3) to two specific functions in $\mathcal{H}$: the baseline (uncalibrated) predictor and the calibrated predictor.

- For the baseline predictor $\widehat{\tau}^{[0]}$, we consider the case where the calibrated predictor corresponds to no correction. Plugging into (App-3) yields:

$$L(0) \;\geqslant\; \widehat{L}_n(0) - 2\widehat{\mathfrak{R}}_n(\mathcal{H}) - 2B\sqrt{\tfrac{\log(2/\delta)}{n}}.$$

- For the calibrated predictor, we denote by $\hat{g}$ the function selected via calibration to minimize empirical loss. Applying (App-3) gives:

$$L(\hat{h}) \;\leqslant\; \widehat{L}_n(\hat{h}) + 2\widehat{\mathfrak{R}}_n(\mathcal{H}) + 2B\sqrt{\tfrac{\log(2/\delta)}{n}}.$$

Step 3 (Rearrange). Subtract the lower bound for $L(0)$ from the upper bound for $L(\hat{h})$:

$$L(\hat{h}) - L(0) \;\leqslant\; \widehat{L}_n(\hat{h}) - \widehat{L}_n(0) + 4\,\widehat{\mathfrak{R}}_n(\mathcal{H}) + 2B\sqrt{\tfrac{\log(2/\delta)}{n}}.$$

By definition, the improvement in empirical calibration loss is given by $\Delta_{\text{cal}} = \widehat{L}_n(0) - \widehat{L}_n(\hat{h})$. Also, by Lemma 1, we have

$$\begin{aligned} L(\hat{h}) - L(0) &= \left[L(\hat{h}) - \mathbb{E}\left[\left(\tau(\mathbf{X}_i) - \check{\tau}_i\right)^2\right]\right] - \left[L(0) - \mathbb{E}\left[\left(\tau(\mathbf{X}_i) - \check{\tau}_i\right)^2\right]\right] \\ &= \mathbb{E}\left[\left(\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_i) + \hat{h}(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)\right)^2\right] - \mathbb{E}\left[\left(\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)\right)^2\right]. \end{aligned}$$

Combining these results together, the inequality becomes:

$$\mathbb{E}\left[\left(\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_i) + \hat{h}(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)\right)^2\right] - \mathbb{E}\left[\left(\widehat{\tau}^{[0]}(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)\right)^2\right] \;\leqslant\; -\Delta_{\text{cal}} + 4\,\widehat{\mathfrak{R}}_n(\mathcal{H}) + 2B\sqrt{\tfrac{\log(2/\delta)}{n}}.$$

Finally, multiplying -1 on both sides of the inequality gives the result. ■

The bound indicates that when the calibration model h is trained to reduce empirical mean squared residuals on a held-out calibration set, this improvement generalizes to unseen customers at the population level. By Lemma 1, such residual reduction also translates into improved prediction accuracy, albeit limited by a model-dependent overfitting term and a random noise term that diminishes with calibration sample size. This result highlights an important trade-off in the design of the calibration procedure. On one hand, richer function classes provide greater flexibility to fit complex residual patterns and reduce in-sample error. On the other hand, they also introduce a higher risk of overfitting, which is captured by larger Rademacher complexity terms in the bound. From a practical standpoint, this suggests that careful selection of the calibration model class — balancing expressive power with regularization — is important for ensuring reliable gains in prediction performance, particularly when working with limited calibration data.

Web Appendix C.3 Error Improvement Bound for Adaptive Boosting

Motivated by this trade-off, we adopt the function class used in the additive boosting algorithm as our calibration model. This class provides sufficient expressive power to improve upon a broad range of baseline models, while maintaining control over model complexity through constraints on the base learners and step sizes at each iteration. To formalize this idea, we first define the model class induced by the boosting procedure under consideration.

Definition App-1 (Additive Boosting Class)

Let $\mathcal{C}$ be a class of functions such that for every $c \in \mathcal{C}$, we have $|c(\cdot)| \leqslant C$ almost surely, for some constant $C > 0$. For an integer $R \geqslant 1$ and a constant $\bar{\rho} > 0$, we define the calibration function class $\mathcal{H}$ as

$$\mathcal{H} := \Big\{ h(\cdot) \mid \ h(\cdot) = \sum_{r=1}^{R} \rho^{[r]} c^{[r]}(\cdot), c^{[r]} \in \mathcal{C}, \ 0 \leqslant |\rho^{[r]}| \leqslant \bar{\rho} \Big\}.$$

Next, we derive the upper bound on the empirical Rademacher complexity of $\mathcal{H}$ and incorporate it into Theorem App-1 to obtain the final improvement bound.

Corollary App-1 (Error Reduction Bound for Honest Additive Boosting)

Consider the additive boosting algorithm for the honest model calibration procedure in Theorem App-1.

Then, the expected error improvement follows the following bound:

$$\mathbb{E}\left[\left(\hat{\tau}(\widetilde{\mathbf{X}}_i)+\hat{h}(\widetilde{\mathbf{X}}_i)-\tau(\mathbf{X}_i)\right)^2\right]-\mathbb{E}\left[\left(\hat{\tau}(\widetilde{\mathbf{X}}_i)-\tau(\mathbf{X}_i)\right)^2\right] \leqslant -\Delta_{\text{cal}}+8\,\bar{\rho}R\widehat{\mathfrak{R}}_n(\mathcal{C})+2B\sqrt{\tfrac{\log(2/\delta)}{n}}.$$

Proof. Write $S(f) := n^{-1}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i)$. First, observe that

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{|\rho^{[r]}|\leqslant\bar{\rho},\ c^{[1]},\ldots,c^{[R]}\in\mathcal{C}}\sum_{r=1}^{R}\rho^{[r]}S(c^{[r]})\right].$$

Since $\sup_{|\rho|\leqslant\bar{\rho}}\rho\, s=\bar{\rho}\,|s|$ for any $s\in\mathbb{R}$, we obtain

$$\sup_{|\rho^{[r]}|\leqslant\bar{\rho},\ c^{[1]},\ldots,c^{[R]}\in\mathcal{C}}\sum_{r=1}^{R}\rho^{[r]}S(c^{[r]})=\bar{\rho}\sum_{r=1}^{R}\sup_{c\in\mathcal{C}}\left|S(c)\right|.$$

Furthermore, since $|S(c)|\leqslant\sup_{c\in\mathcal{C}}S(c)+\sup_{c\in\mathcal{C}}[-S(c)]$, we have $\sup_{c\in\mathcal{C}}\left|S(c)\right|\leqslant 2\widehat{\mathfrak{R}}_n(\mathcal{C})$. Combining the steps and pulling out the deterministic factors, we obtain:

$$\widehat{\mathfrak{R}}_n(\mathcal{H})\leqslant 2\bar{\rho}R\widehat{\mathfrak{R}}_n(\mathcal{C}).$$

Finally, inserting the above bound to Theorem App-1 gives the result. ■

This result suggests that overfitting can be controlled by adopting simpler model classes — such as linear models or shallow decision trees—commonly used as base learners in gradient boosting algorithms. These models strike a favorable balance between expressive power and statistical stability and, under standard conditions,— typically achieve an $O(1/\sqrt{n})$ convergence rate for generalization error.

Web Appendix C.4 Generalization Error for Non-honest Model Calibration

Note that Theorem App-1 fails to hold without honesty in the calibration step, i.e., if the training set $\widetilde{\mathcal{D}}_{\text{train}}$ and the calibration set $\widetilde{\mathcal{D}}_{\text{cal}}$ overlap. In this case, the initial estimator $\hat{\tau}^{[0]}$ is no longer independent of the calibration data, and thus the functional class $\mathcal{H}$ used in the calibration procedure becomes a random class that depends on the data.

Then, what happens if the calibration procedure is performed on the same dataset used to construct $\hat{\tau}^{[0]}$? In this case, we lose the independence between model training and calibration, and as a result, we can only bound the gap between the true prediction error $(\hat{\tau}^{[0]}-\tau)^2$ and the empirical mean squared residuals $(\hat{\tau}^{[0]} - \check{\tau})^2$ using the Rademacher complexity of the composite model class $\mathcal{T} + \mathcal{H} = \{\hat{\tau}^{[0]} + h : \hat{\tau}^{[0]} \in \mathcal{T}, \ h \in \mathcal{H}\}$. Specifically, using the same technique, we can show that with probability $1 - \delta$,

$$\mathbb{E}\big[\big(\hat{\tau}^{[0]}(\widetilde{\mathbf{X}}_i) + h(\widetilde{\mathbf{X}}_i) - \tau(\mathbf{X}_i)\big)^2\big] \leqslant \frac{1}{m} \sum_{j \in \widetilde{\mathcal{D}}_{\text{train}}} \big(\hat{\tau}^0(\widetilde{\mathbf{X}}_j + h(\widetilde{\mathbf{X}}_i) - \check{\tau}_j\big)^2 + 2\,\widehat{\mathfrak{R}}_m(\mathcal{T} + \mathcal{H}) + 2B\sqrt{\frac{\log(2/\delta)}{m}},$$

where m denotes the training sample size. Note that if the baseline model class $\mathcal{T}$ is already sufficiently expressive such that the empirical mean squared residuals are small using $\hat{\tau}^{[0]}$ alone, then performing additional calibration on the same data (i.e., without honesty) offers limited potential for further error reduction, even when the training sample is large. At the same time, it increases the risk of overfitting, since $\widehat{\mathfrak{R}}_m(\mathcal{T} + \mathcal{H})$ is greater or equal to $\widehat{\mathfrak{R}}_m(\mathcal{T})$. Thus, without sample splitting, calibration may degrade rather than improve generalization performance. While using the entire sample (i.e., avoiding splitting) would improve estimation due to a larger sample size, this benefit is limited in practice. Specifically, neither the Rademacher complexity term $\widehat{\mathfrak{R}}_m(\mathcal{T} + \mathcal{H})$ nor the stochastic error term $\sqrt{\frac{\log(2/\delta)}{m}}$ changes in order of m.

Web Appendix D Additional Details and Results for the Simulation Analyses in Section 5

In this appendix, we provide implementation details about the simulation analyses described in Section 5 of the main document and present robustness checks.

Web Appendix D.1 Details on Model Specifications

Constructing the Initial CATE Model. In our main analysis, we implement the R-learner (Nie and Wager 2021) using regression forests from the `grf` package with default parameters for both the conditional mean outcome models and the CATE model. The true propensity score is used to construct Robinson's transformation. In Web Appendix D.5, we provide additional robustness checks for different initial CATE models.

Constructing the DR Score for Model Calibration. To construct the DR score, we use a fixed propensity score ($e = 0.5$) since the treatment assignment is completely randomized. We evaluate three candidate models for the conditional mean: linear regression, linear regression with interactions, and regression forests. The training set consists of 1,000 individuals (500 treated and 500 non-treated), while the holdout evaluation set includes 2,000 individuals (1,000 treated and 1,000 non-treated). Table App-3 reports the mean squared error (MSE) of each model in predicting the outcome. Since linear regression with interactions yields the lowest prediction error, we select it as our preferred model for the DR score. Additionally, we use the same model to construct calibration models. Notably, the conditional mean model used in the DR transformation and calibration process does not align exactly with the data-generating process. However, the algorithm still effectively reduces the MSE of CATE predictions.

Table App-3: MSE of Conditional Outcome Models Across Varying Privacy Levels

(a) Scenario 1: DP-Protected Covariates

Privacy	Regression	Regression with Interactions	Regression Forest
No	33.4	29.6	33.1
50.0	34.0	30.6	34.2
25.0	36.1	33.7	36.3
16.7	39.2	37.8	39.7
12.5	41.4	40.9	41.3
10.0	43.4	43.0	43.1

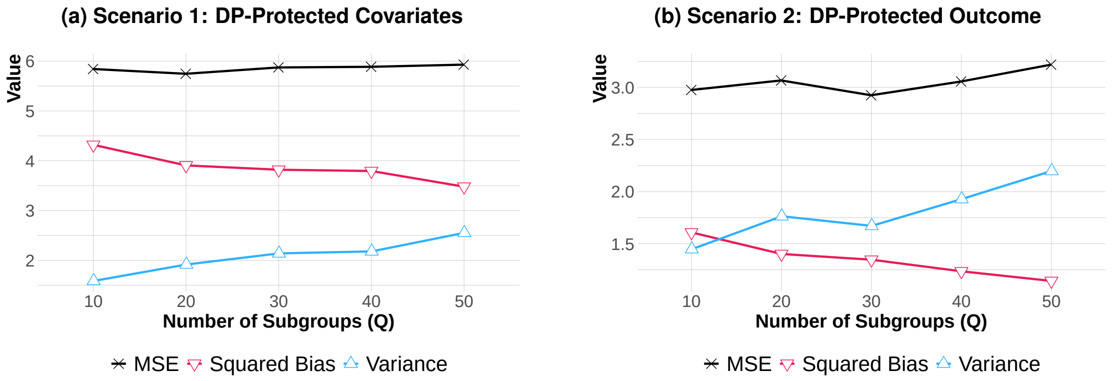
(b) Scenario 2: DP-Protected Outcome

Privacy	Regression	Regression with Interactions	Regression Forest
No	33.4	29.5	33.6
50.0	35.1	31.1	35.0
25.0	41.3	37.5	41.4
16.7	51.6	48.6	51.7
12.5	65.8	63.3	65.6
10.0	83.9	81.7	84.4

Determine Number of Subgroups and Number of Iterations. Next, we investigate the sensitivity of our solution to the pre-specified number of subgroups (Q) and the number of iterations (R). We start by illustrating the bias-variance trade-offs for the number of subgroups. Figure App-3 displays the predictive error metrics when applying the proposed solution Across Varying values of Q under two scenarios with high privacy protection. As we increase the number of subgroups, the bias decreases for both scenarios, while the variance correspondingly increases. This outcome showcases a classic bias-variance trade-off commonly observed in machine learning models: increasing model complexity may enhance precision but concurrently introduce additional variability. Notably, when the outcome is protected by DP, changes in Q do not significantly affect overall accuracy. On the other hand, when the covariates are protected by DP, using the smallest value of Q achieves the lowest MSE.

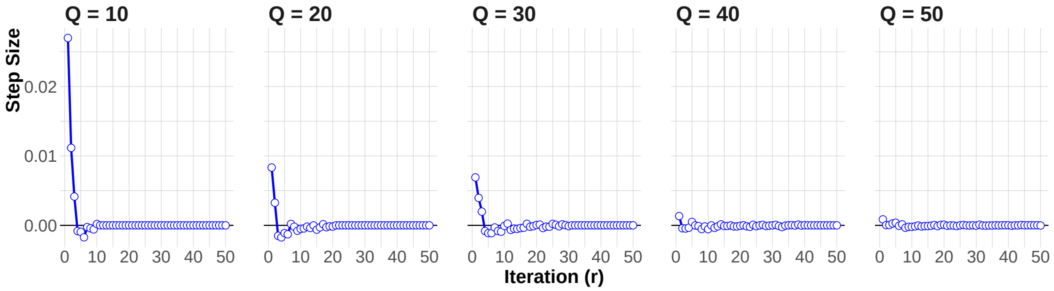
Figure App-4 illustrates the step size $\rho_{q^\star}^{[r]}$ in each iteration (capped at 50 iterations) when applying the proposed solution Across Varying numbers of subgroups (i.e., varying values of Q). Generally, the step sizes approach zero before the completion of the first Q iterations, regardless of the chosen value for Q. This observation suggests that performing $R = Q$ iterations is

Figure App-3: Predictive Errors of Proposed Method Across Varying Numbers of Subgroups



typically sufficient for the proposed solution, thereby providing a practical guideline for setting this hyperparameter.

Figure App-4: Step Size in Each Iteration



Note: Each point is calculated by averaging the mean step size over 100 bootstrap replications. We present results from the case when the covariates are protected by DP with $\epsilon = 10.0$, but the pattern remains the same across different privacy levels. We use R-learner with regression forests as the initial CATE model.

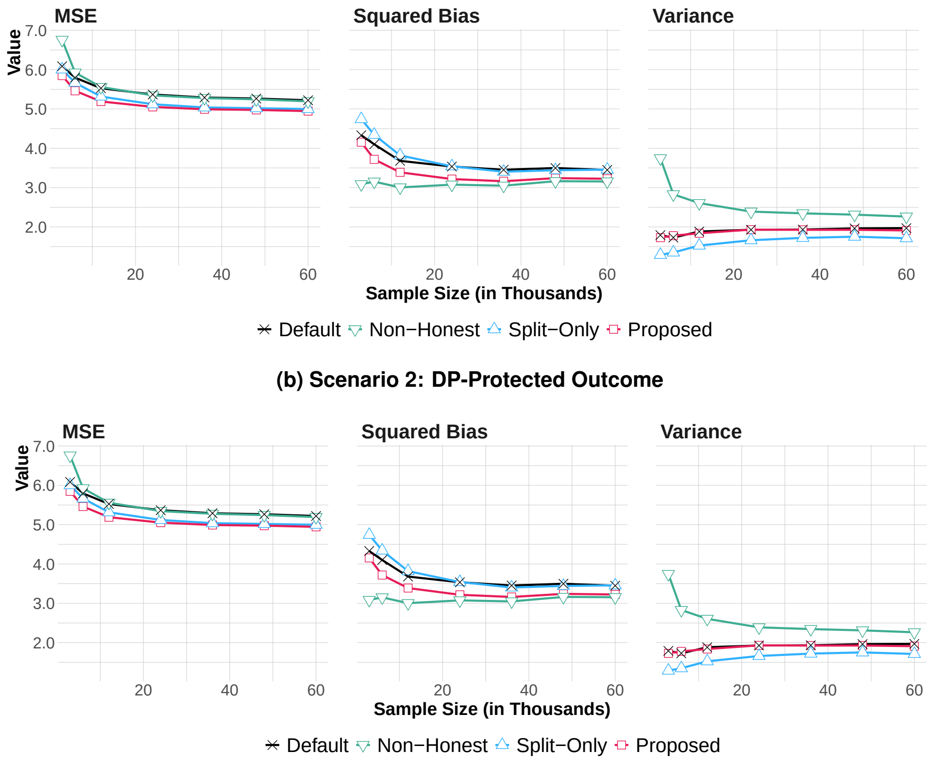
Web Appendix D.2 Bias and Variance

To further examine how the PROPOSED method enhances predictive accuracy, we extend the analysis in Section 5 by decomposing the mean squared error (MSE) into its squared bias and variance components. These metrics are computed as follows:

1. *(Mean Squared Bias)* We calculate the average deviation of model predictions from the actual CATEs, i.e., $\widehat{\mathbb{E}}^2\left[(\widehat{err}_i^b)\right] = \frac{1}{N_{\text{holdout}}}\sum_{i\in D_{\text{holdout}}}\left[\frac{1}{B}\sum_{b=1}^{B}\widehat{err}_i^b\right]^2$.
2. *(Mean Variance)* We calculate the average variability in the model's predictions across individuals in the holdout set when the model is trained with different model construction sets, i.e., $\widehat{\text{Var}}\left[\widehat{\tau}\right] = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_{\text{holdout}}}\sum_{i\in D_{\text{holdout}}}\left\{\left(\widehat{\tau}^b(\mathbf{X}_i) - \widehat{\mathbb{E}}[\widehat{\tau}^b(\mathbf{X}_l)]\right)^2\right\}$.

Figure App-5: Statistical Accuracy Across Varying Sample Sizes

(a) Scenario 1: DP-Protected Covariates



(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

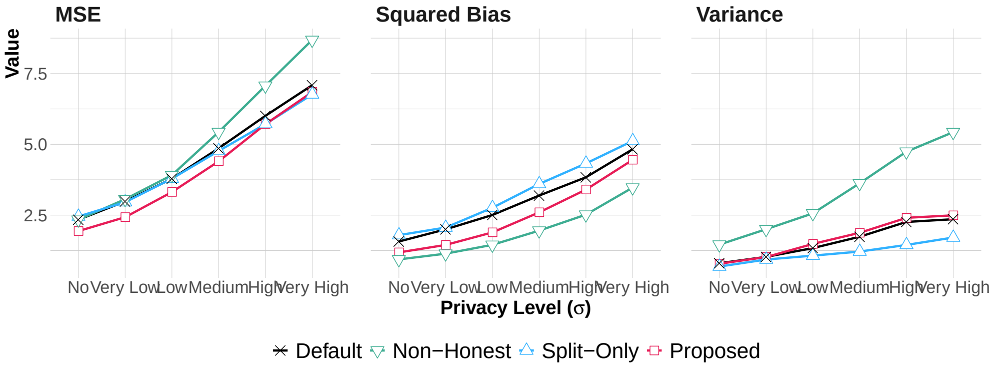
Figure App-5 presents the mean squared error, squared bias, and variance of different methods across varying sample sizes. First, we find that the NON-HONEST method achieves the smallest squared bias because its initial CATE model is trained on a sample three times larger than that used by the PROPOSED and SPLIT-ONLY methods. However, as the sample size increases, the squared bias of the PROPOSED method decreases to a level comparable to that of the NON-HONEST method, whereas the SPLIT-ONLY method does not exhibit the same improvement. Second, despite its lower bias, the NON-HONEST method consistently shows significantly higher variance than the other methods. This is due to overfitting to noise in the DR score. Together, these findings confirm the importance of the *honesty principle* in mitigating overfitting when the DR score is noisy.

Web Appendix D.3 Results for DP-protection on Both Outcome and Covariates

In this appendix, we present results where both the outcome and covariates are protected under differential privacy. Specifically, we consider five privacy levels (very low, low, medium, high, and very high) corresponding to total privacy budgets ranging from 4 to 20 for both the outcome and covariates. We then replicate the previous analyses under these settings to assess model performance across different levels of privacy protection.

Varying Privacy Levels. Figure App-6 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply R-learners with regression forests as the Default method and the initial CATE models. The results are consistent with the findings from the main analysis.

Figure App-6: Statistical Accuracy Across Varying Privacy Levels



Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Table App-4 reports the AUTOC values across different privacy levels. The Proposed method consistently outperforms other approaches, achieving the best performance across all privacy levels. Table App-5 presents the value improvement of different methods across various privacy levels. The results align with the findings from the main analysis.

Table App-4: AUTOC Values Across Varying Privacy Levels

Privacy	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	2.37 (96%)	2.34 (75%)	2.33 (48%)	2.33
Very Low	2.3 (100%)	2.26 (76%)	2.22 (52%)	2.22
Low	2.14 (100%)	2.11 (83%)	2.08 (75%)	2.06
Medium	1.92 (100%)	1.90 (96%)	1.81 (46%)	1.81
High	1.64 (92%)	1.66 (97%)	1.54 (57%)	1.55
Very High	1.36 (83%)	1.40 (92%)	1.26 (48%)	1.28

Note: We calculate the AUTOC value from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

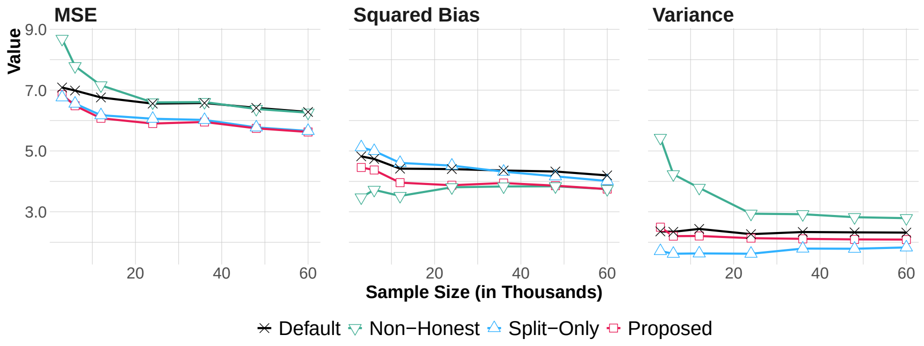
Table App-5: Targeting Value Improvement Across Varying Privacy Levels

Privacy	PROPOSED	SPLIT-ONLY	NON-HONEST
No	5.3% (84%)	−1.4% (36%)	1.0% (64%)
Very Low	8.8% (100%)	0.7% (56%)	0.2% (56%)
Low	12.5% (100%)	1.1% (56%)	4.5% (64%)
Medium	15.3% (100%)	1.1% (56%)	−4.3% (40%)
High	14.2% (84%)	6.4% (72%)	−18.1% (24%)
Very High	24.7% (72%)	21.8% (68%)	−68.2% (12%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Varying Sample Sizes. Figure App-7 presents the MSE, squared bias, and variance of different methods across varying sample sizes under the high privacy level. Table App-6 reports the AUTOC values across different sample sizes. Table App-7 presents the value improvement of different methods across various sample sizes. The results align with the findings in Section 5.5.

Figure App-7: Statistical Accuracy Across Varying Sample Sizes



Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Table App-6: AUTOC Values Across Varying Sample Sizes

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
3k	1.36 (83%)	1.4 (92%)	1.26 (48%)	1.28
6k	1.45 (96%)	1.42 (92%)	1.33 (76%)	1.30
12k	1.57 (100%)	1.54 (100%)	1.45 (88%)	1.40
24k	1.6 (100%)	1.55 (100%)	1.47 (100%)	1.42
36k	1.61 (100%)	1.58 (100%)	1.49 (100%)	1.45
48k	1.62 (100%)	1.60 (100%)	1.50 (100%)	1.46
60k	1.65 (100%)	1.64 (100%)	1.55 (100%)	1.50

Note: We calculate the AUTOC value from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Table App-7: Value Improvement Across Varying Sample Sizes

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST
3k	24.7% (72%)	21.8% (68%)	−68.2% (12%)
6k	75.8% (92%)	74.6% (92%)	−50.6% (20%)
12k	81.6% (100%)	77.9% (100%)	-13.5% (24%)
24k	67.5% (100%)	55.5% (100%)	3.8% (65%)
36k	86.1% (100%)	83.9% (100%)	1.7% (72%)
48k	70.2% (100%)	74% (100%)	11.2% (80%)
60k	57% (100%)	60.4% (100%)	4.4% (80%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Web Appendix D.4 Additional Benchmarks for Honest Model Calibration

In this appendix, we provide additional benchmarks for model calibration to further demonstrate the value of our PROPOSED solution. Specifically, we evaluate (i) different subgroup partitioning strategies within the PROPOSED method to assess the impact of within-group homogeneity in the subgroup cross-learning strategy and (ii) stochastic gradient boosting (Friedman 2002) for honest calibration to compare the effectiveness of our proposed subgroup cross-learning approach against a standard machine learning technique.

Web Appendix D.4.1 Alternative Subgroup Partitioning Strategies.

We compare three alternative subgroup partitioning strategies to splitting individuals based on their predicted CATEs from the initial CATE predictions.

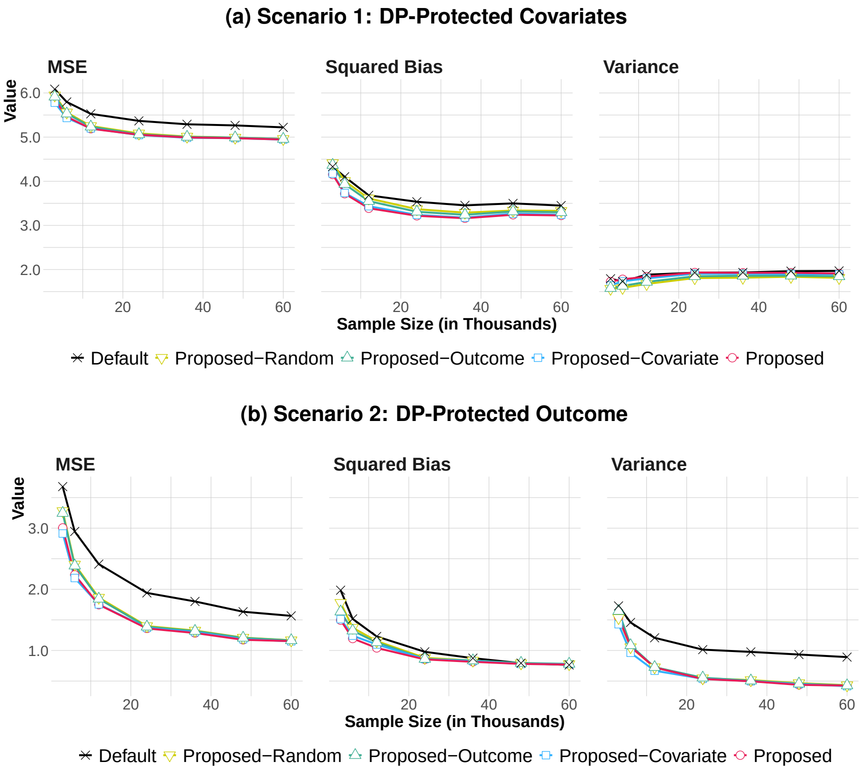
1. *Partitioning based on covariate values*: We use K-means clustering on covariate values to define the subgroups.

2. *Partitioning based on the outcome*: We categorize individuals based on the outcome variable.

3. *Partitioning randomly*: We generate random subgroups of individuals.

Figure App-8 presents the MSE, squared bias, and variance across different sample sizes under high privacy levels. Overall, all subgroup partitioning strategies improve accuracy compared to the DEFAULT method. When the sample size is large, performance remains similar across different subgroup partitioning strategies.

Figure App-8: Statistical Accuracy Across Varying Sample Sizes: Subgroup Partitioning



(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Web Appendix D.4.2 Alternative Boosting Methods

In this appendix, we compare several alternative boosting procedures for model calibration. We first consider the alternative honest model calibration methods.

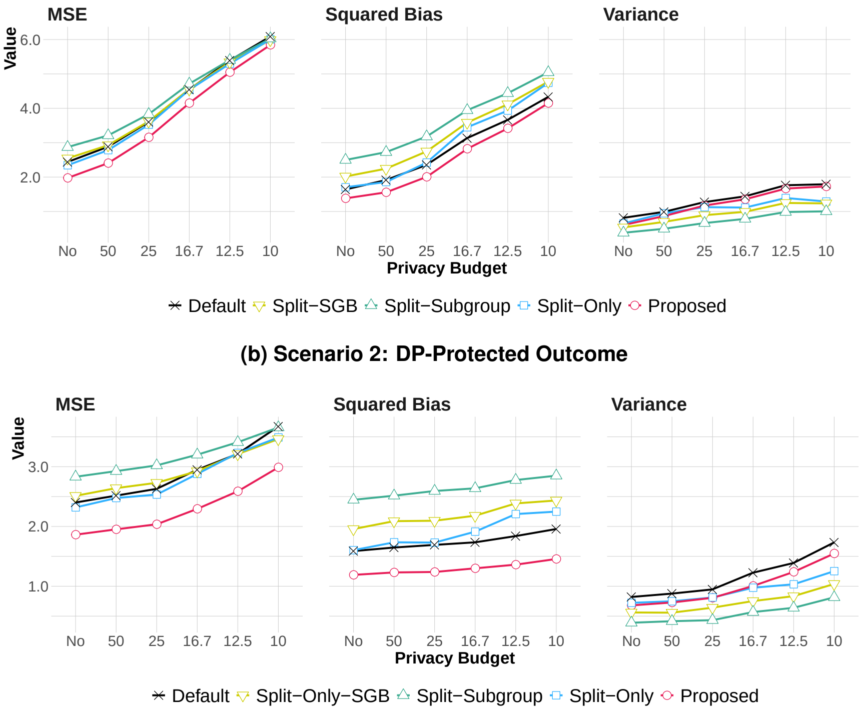
Calibration with Sample Splitting and Subgroup Updates (SPLIT-SUBGROUP). This method performs model calibration using a separate calibration dataset $\widetilde{\mathcal{D}}_{\text{cal}}$ and stops when there is no improvement on a third validation set $\widetilde{\mathcal{D}}_{\text{val}}$. We apply subgroup-based calibration (e.g., Whitehouse et al. 2024), where subgroups are formed based on the predicted CATEs from the initial model. Calibration models are trained separately within each subgroup, and the one that yields the greatest reduction in mean squared error (MSE) is selected. Unlike our proposed method, the step size is determined using data from the *same* subgroup rather than other subgroups.

Calibration with Stochastic Gradient Boosting (Split-SGB). This method performs model calibration using a separate calibration dataset $\widetilde{\mathcal{D}}_{\text{cal}}$ and stops when there is no improvement on a third validation set $\widetilde{\mathcal{D}}_{\text{val}}$. In each iteration, we apply the stochastic gradient descent method to update the model. We randomly sample 20% of individuals from the calibration set, then use them to construct the calibration model and determine the step size.

Figure App-9 presents the MSE, squared bias, and variance of different methods across different privacy levels. Overall, the alternative honest model calibration methods do not improve accuracy and instead exhibit significantly higher bias.

Figure App-9: Statistical Accuracy Across Varying Privacy Levels: Alternative Boosting Methods

(a) Scenario 1: DP-Protected Covariates



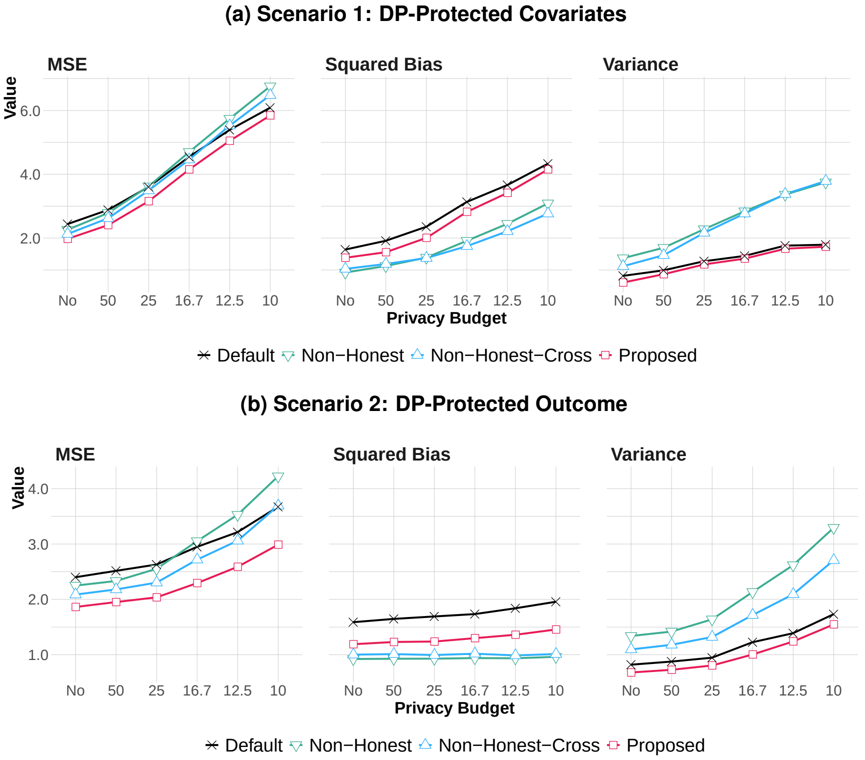
(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

In addition, we examine the value of the proposed honest step-size determination in settings without a separate calibration set (Non-honest-Cross). Figure App-10 presents the

MSE, squared bias, and variance of different methods across different privacy levels. Consistent with our theoretical insights, we find that honest step-size determination enhances predictive accuracy, even without using a separate dataset for model calibration.

Figure App-10: Statistical Accuracy Across Varying Privacy Levels: Honest Step-size Learning


(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point. The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Web Appendix D.5 Robustness Checks for Different Initial CATE models

In this section, we further evaluate the performance of two different approaches for the initial CATE model:

1. *Causal Forest* (Wager and Athey 2018): We implement causal forest using the `grf` package with default parameters.

2. *DR-learner with regression forests* (Kennedy 2023): We construct the DR-learner by regressing the cross-fitted doubly robust scores on the covariates. All regression models (including the conditional mean outcome models and the final CATE model) are estimated using regression forests with default parameters from the `grf` package.

3. *R-XGBoost models*: Similar to the R-learner with regression forests used in the main analysis, we implement R-XGBoost models for both the conditional mean outcome and CATE models. Hyperparameter tuning follows the methods implemented by Nie and Wager (2021).

4. *DR-XGBoost* models: Similar to the DR-learner with regression forests, we implement DR-XGBoost for both the conditional mean outcome and CATE models. Hyperparameter tuning follows the methods implemented by Nie and Wager (2021).

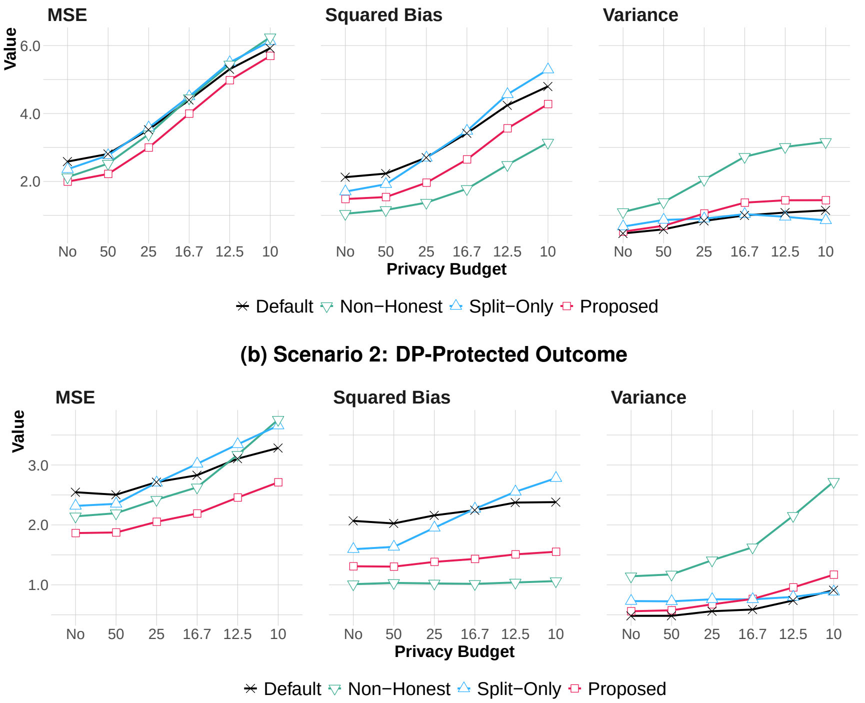
Results for the T-learner (Künzel et al. 2019) using both regression forests and XGBoost are available upon request. Overall, regardless of the initial CATE models, our proposed solution yields to the smallest predictive error and best targeting performance Across Varying privacy levels and experiment sample sizes.

Web Appendix D.5.1 Results for Causal Forest

Varying Privacy Levels. Figure App-11 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply Casual Forest as the DEFAULT method and the initial CATE models. Table App-8 reports the AUTOC values across different privacy levels. Table App-9 presents the value improvement of different methods across various privacy levels. Overall, the results align with the findings from the main analysis in Section 5.5.

Varying Sample Sizes. Figure App-12 presents the MSE, squared bias, and variance of different methods across different sample sizes, where we apply Casual Forest as the DEFAULT method and the initial CATE models. Table App-10 reports the AUTOC values across different sample sizes. Table App-11 presents the value improvement of different methods across various sample sizes. Overall, the results align with the findings from the main analysis in Section 5.5.

Figure App-11: Statistical Accuracy Across Varying Privacy Levels: Causal Forest


(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table App-8: AUTOC Values Across Varying Privacy Levels: Causal Forest

(a) Scenario 1: DP-Protected Covariates

Privacy	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	2.38 (82%)	2.34 (56%)	2.35 (54%)	2.35
50.0	2.33 (94%)	2.28 (70%)	2.28 (56%)	2.28
25.0	2.19 (90%)	2.15 (66%)	2.13 (50%)	2.13
16.7	2.00 (82%)	1.98 (68%)	1.94 (32%)	1.96
12.5	1.81 (80%)	1.79 (72%)	1.73 (28%)	1.77
10.0	1.63 (82%)	1.63 (82%)	1.55 (24%)	1.60

(b) Scenario 2: DP-Protected Outcome

Privacy	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	2.38 (78%)	2.34 (56%)	2.34 (45%)	2.34
50.0	2.38 (78%)	2.35 (50%)	2.34 (37%)	2.35
25.0	2.36 (84%)	2.32 (54%)	2.31 (38%)	2.32
16.7	2.34 (69%)	2.31 (49%)	2.29 (26%)	2.32
12.5	2.31 (77%)	2.27 (59%)	2.22 (31%)	2.26
10.0	2.26 (74%)	2.23 (63%)	2.17 (23%)	2.22

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-9: Targeting Value Improvement Across Varying Privacy Levels: Causal Forest

(a) Scenario 1: DP-Protected Covariates

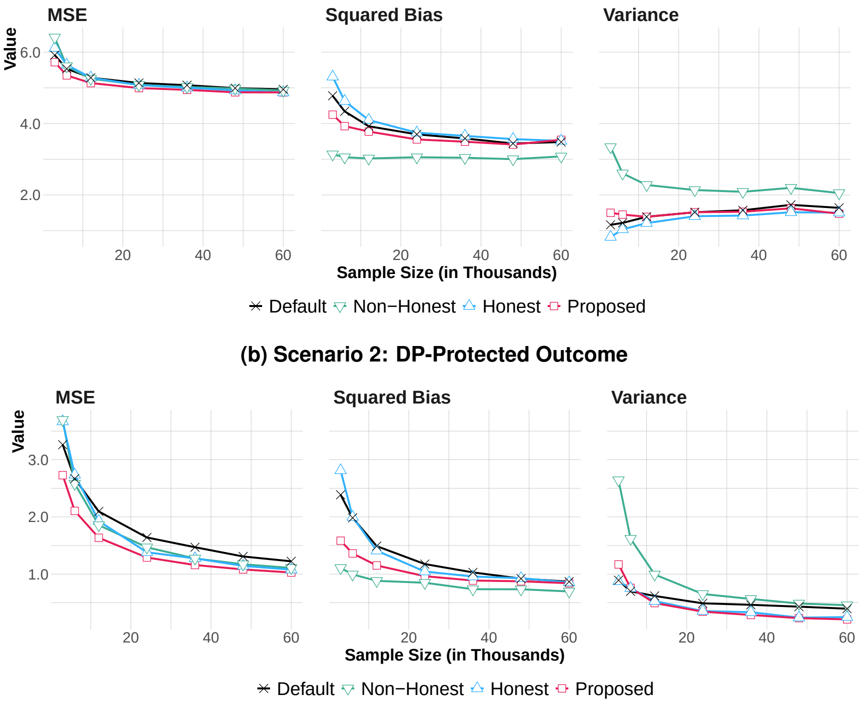
Privacy	PROPOSED	SPLIT-ONLY	NON-HONEST
No	10.1% (96%)	3.6% (68%)	8.9% (84%)
50.0	10.4% (98%)	1.9% (64%)	6.4% (78%)
25.0	12.7% (98%)	−3.4% (48%)	7.9% (86%)
16.7	13.4% (84%)	−3.4% (42%)	5.3% (56%)
12.5	14.9% (86%)	−15.9% (24%)	11.0% (60%)
10.0	16.8% (92%)	−26.4% (22%)	4.4% (58%)

(b) Scenario 2: DP-Protected Outcome

Privacy	PROPOSED	SPLIT-ONLY	NON-HONEST
No	9.7% (97%)	2.8% (72%)	5.6% (76%)
50.0	9.1% (92%)	1.8% (61%)	5.3% (72%)
25.0	10.6% (93%)	−1.7% (52%)	6.6% (73%)
16.7	12% (90%)	−5.1% (41%)	8.4% (76%)
12.5	13.8% (92%)	−6.9% (36%)	7.1% (71%)
10.0	11.5% (86%)	−11.1% (29%)	2.5% (53%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Figure App-12: Statistical Accuracy Across Varying Sample Sizes: Causal Forest


(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table App-10: AUTOC Values Across Varying Sample Sizes: Causal Forest

(a) Scenario 1: DP-Protected Covariates

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
3k	1.63 (82%)	1.63 (82%)	1.55 (24%)	1.60
6k	1.69 (92%)	1.69 (92%)	1.64 (30%)	1.58
12k	1.74 (98%)	1.73 (98%)	1.7 (50%)	1.65
24k	1.75 (100%)	1.75 (100%)	1.73 (70%)	1.68
36k	1.76 (100%)	1.76 (100%)	1.74 (85%)	1.70
48k	1.77 (100%)	1.77 (100%)	1.75 (90%)	1.70
60k	1.78 (100%)	1.78 (100%)	1.76 (90%)	1.71

(b) Scenario 2: DP-Protected Outcome

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
3k	2.26 (74%)	2.23 (63%)	2.17 (23%)	2.22
6k	2.35 (92%)	2.31 (58%)	2.29 (40%)	2.30
12k	2.42 (94%)	2.4 (80%)	2.38 (74%)	2.37
24k	2.45 (100%)	2.45 (95%)	2.43 (85%)	2.42
36k	2.48 (92%)	2.47 (90%)	2.46 (70%)	2.45
48k	2.49 (98%)	2.48 (100%)	2.47 (100%)	2.47
60k	2.50 (96%)	2.49 (80%)	2.49 (90%)	2.48

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-11: Targeting Value Improvement Across Varying Sample Sizes: Causal Forest

(a) Scenario 1: DP-Protected Covariates

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST
3k	16.8% (92%)	−26.4% (22%)	4.4% (58%)
6k	12.9% (90%)	−12% (26%)	5.5% (58%)
12k	5.1% (94%)	−1.6% (56%)	1.4% (58%)
24k	5.8% (95%)	2.5% (74%)	1.5% (60%)
36k	5.7% (100%)	3.1% (83%)	3% (85%)
48k	3.5% (96%)	2% (87%)	0.3% (52%)
60k	2.5% (90%)	2.9% (100%)	0.5% (61%)

(b) Scenario 2: DP-Protected Outcome

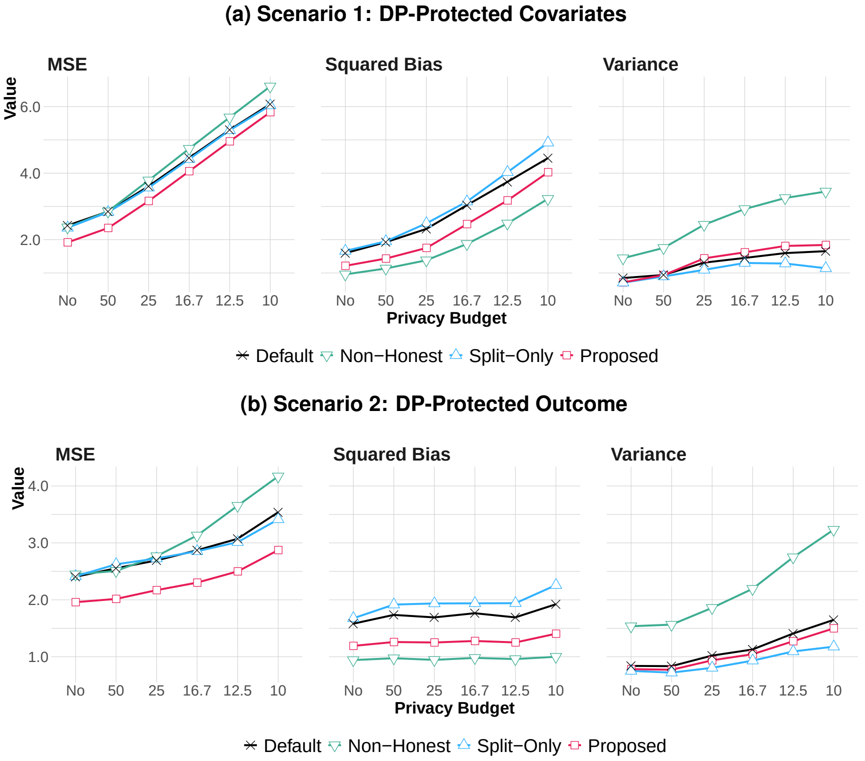
Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST
3k	10.1% (82%)	−11.6% (28%)	3.0% (52%)
6k	8.2% (98%)	−0.1% (52%)	2.6% (66%)
12k	6.6% (100%)	2.9% (84%)	3.8% (92%)
24k	3.6% (100%)	2.8% (95%)	1.8% (95%)
36k	3.3% (95%)	2.0% (90%)	2.1% (90%)
48k	1.8% (93%)	1.7% (90%)	0.9% (90%)
60k	1.1% (88%)	0.7% (85%)	0.6% (82%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Web Appendix D.5.2 Results for DR-learner with Regression Forests

Varying Privacy Levels. Figure App-13 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply DR-learner with regression forests as the DEFAULT method and the initial CATE models.

Figure App-13: Statistical Accuracy Across Varying Privacy Levels: DR-learner with RFs



Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table App-12 reports the AUTOC values across different privacy levels. Table App-13 presents the value improvement of different methods across various privacy levels. Overall, the results align with the findings from the main analysis in Section 5.5.

Table App-12: AUTOC Values Across Varying Privacy Levels: DR-learner with RFs

(a) Scenario 1: DP-Protected Covariates

Privacy(ϵ)	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	2.35 (92%)	2.32 (80%)	2.3 (62%)	2.29
50.0	2.28 (96%)	2.24 (76%)	2.22 (66%)	2.20
25.0	2.13 (98%)	2.10 (87%)	2.06 (63%)	2.05
16.7	1.96 (96%)	1.93 (90%)	1.87 (48%)	1.88
12.5	1.76 (96%)	1.73 (92%)	1.68 (43%)	1.68
10.0	1.56 (100%)	1.55 (95%)	1.47 (29%)	1.49

(b) Scenario 2: DP-Protected Outcome

Privacy(ϵ)	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	2.37 (96%)	2.34 (72%)	2.31 (54%)	2.31
50.0	2.36 (96%)	2.33 (84%)	2.29 (55%)	2.3
25.0	2.34 (98%)	2.31 (83%)	2.27 (56%)	2.27
16.7	2.32 (99%)	2.28 (85%)	2.23 (56%)	2.23
12.5	2.30 (98%)	2.26 (88%)	2.20 (57%)	2.19
10.0	2.22 (100%)	2.19 (88%)	2.13 (60%)	2.10

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-13: Targeting Value Improvement Across Varying Privacy Levels: DR-learner with RFs

(a) Scenario 1: DP-Protected Covariates

Privacy(ϵ)	Proposed	Split-only	Non-Honest
No	7.2% (94%)	1% (64%)	1.8% (58%)
50.0	9.9% (100%)	2.2% (66%)	4.3% (78%)
25.0	10.3% (98%)	1.1% (56%)	2.2% (60%)
16.7	12.6% (96%)	-0.8% (58%)	1.4% (54%)
12.5	16.2% (94%)	-3.2% (46%)	-0.1% (50%)
10.0	17.5% (94%)	-4.8% (44%)	-11.6% (30%)

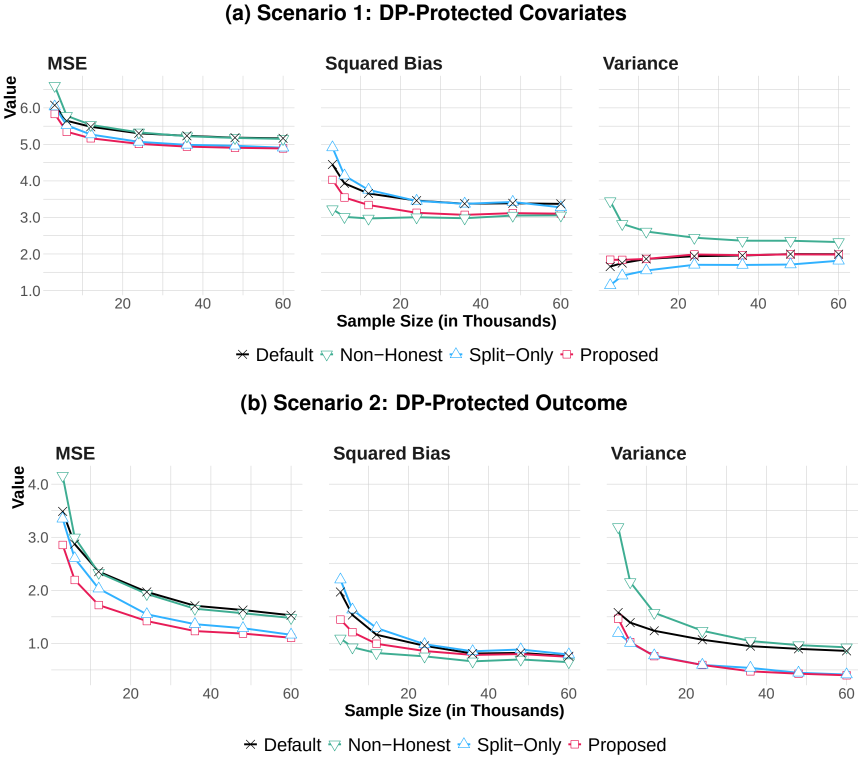
(b) Scenario 2: DP-Protected Outcome

Privacy(ϵ)	Proposed	Split-only	Non-Honest
No	7% (96%)	1.6% (58%)	1.1% (52%)
50.0	8.1% (97%)	-0.1% (62%)	2.0% (60%)
25.0	9.9% (98%)	0.5% (56%)	3.4% (66%)
16.7	11.1% (96%)	2.1% (66%)	3.1% (58%)
12.5	11.3% (99%)	1.7% (68%)	0.1% (46%)
10.0	13.2% (100%)	3.4% (76%)	0.8% (56%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses). The results presented here are based on the use of DR-learner with regression forests as the initial CATE model.

Varying Sample Sizes. Figure App-14 presents the MSE, squared bias, and variance of different methods across different sample sizes, where we apply DR-learner with regression forests as the Default method and the initial CATE models. Table App-14 reports the AUTOC values across different sample sizes. Table App-15 presents the value improvement of different methods across various sample sizes. Overall, the results align with the findings from the main analysis in Section 5.5.

Figure App-14: Statistical Accuracy Across Varying Sample Sizes: DR-learner with RFs


(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table App-14: AUTOC Values Across Varying Sample Sizes: DR-learner with RFs

(a) Scenario 1: DP-Protected Covariates

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	1.56 (100%)	1.55 (94%)	1.47 (29%)	1.49
6k	1.66 (100%)	1.64 (94%)	1.60 (76%)	1.58
12k	1.70 (100%)	1.68 (100%)	1.64 (88%)	1.62
24k	1.73 (100%)	1.72 (100%)	1.67 (93%)	1.66
36k	1.74 (100%)	1.73 (100%)	1.69 (95%)	1.67
48k	1.75 (100%)	1.74 (100%)	1.70 (97%)	1.69
60k	1.75 (100%)	1.74 (100%)	1.70 (96%)	1.69

(b) Scenario 2: DP-Protected Outcome

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	2.22 (100%)	2.19 (88%)	2.13 (60%)	2.10
6k	2.33 (100%)	2.29 (88%)	2.25 (88%)	2.21
12k	2.39 (100%)	2.37 (100%)	2.33 (100%)	2.30
24k	2.44 (100%)	2.43 (100%)	2.38 (100%)	2.35
36k	2.47 (100%)	2.45 (100%)	2.42 (100%)	2.40
48k	2.47 (100%)	2.47 (100%)	2.42 (100%)	2.41
60k	2.48 (100%)	2.48 (100%)	2.44 (100%)	2.42

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table App-15: Targeting Value Improvement Across Varying Sample Sizes: DR-learner with RFs

(a) Scenario 1: DP-Protected Covariates

Sample Size	Proposed	Split-only	Non-Honest
3k	37.3% (94%)	3.4% (54%)	9.8% (62%)
6k	35.1% (100%)	19.9% (88%)	16.9% (84%)
12k	25.2% (100%)	18.3% (98%)	9.6% (94%)
24k	17.5% (100%)	16.0% (100%)	4.0% (92%)
36k	16.5% (100%)	15.0% (100%)	3.0% (94%)
48k	13.5% (100%)	12.5% (100%)	3.3% (95%)
60k	15.7% (100%)	14.7% (100%)	3.6% (95%)

(b) Scenario 2: DP-Protected Outcome

Sample Size	Proposed	Split-only	Non-Honest
3k	29.4% (98%)	10.9% (80%)	18.2% (82%)
6k	23.2% (100%)	14.9% (100%)	11.8% (94%)
12k	18.2% (100%)	15.2% (98%)	9.1% (100%)
24k	14.2% (100%)	12.2% (100%)	5.8% (100%)
36k	11.9% (100%)	11.0% (100%)	4.1% (100%)
48k	10.4% (100%)	9.7% (100%)	3.8% (100%)
60k	9.9% (100%)	9.7% (100%)	3.4% (100%)

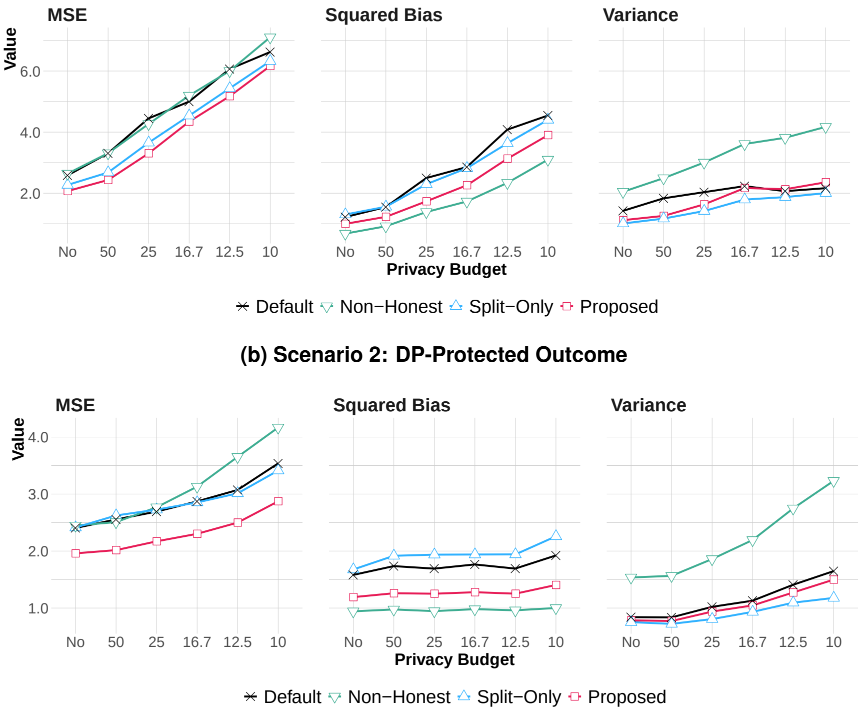
Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Web Appendix D.5.3 Results for R-learners with XGBoost Models

Varying Privacy Levels. Figure App-15 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply R-XGBoost models as the Default method and the initial CATE models. Table App-16 reports the AUTOC values across different privacy levels. Table App-17 presents the value improvement of different methods across various privacy levels. Overall, the results align with the findings from the main analysis in Section 5.5.

Varying Sample Sizes. Figure App-16 presents the MSE, squared bias, and variance of different methods across different sample sizes, where we apply R-XGBoost models as the Default method and the initial CATE models. Table App-18 reports the AUTOC values across different sample sizes. Table App-19 presents the value improvement of different methods across various sample sizes. Overall, the results align with the findings from the main analysis in Section 5.5.

Figure App-15: Statistical Accuracy Across Varying Privacy Levels: R-XGBoost



(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table App-16: AUTOC Values Across Varying Privacy Levels: R-XGBoost

(a) Scenario 1: DP-Protected Covariates

Privacy(ϵ)	Proposed	Split-only	Non-Honest	Default
No	2.32 (72%)	2.31 (72%)	2.27 (48%)	2.27
50.0	2.26 (100%)	2.23 (84%)	2.17 (72%)	2.13
25.0	2.10 (96%)	2.07 (92%)	2.00 (88%)	1.93
16.7	1.91 (88%)	1.89 (86%)	1.84 (68%)	1.81
12.5	1.72 (88%)	1.69 (88%)	1.65 (88%)	1.56
10.0	1.49 (80%)	1.48 (80%)	1.42 (72%)	1.38

(b) Scenario 2: DP-Protected Outcome

Privacy(ϵ)	Proposed	Split-only	Non-Honest	Default
No	2.36 (93%)	2.33 (93%)	2.27 (67%)	2.25
50.0	2.35 (87%)	2.33 (83%)	2.28 (67%)	2.24
25.0	2.33 (97%)	2.29 (83%)	2.24 (83%)	2.18
16.7	2.27 (97%)	2.25 (97%)	2.18 (87%)	2.10
12.5	2.22 (90%)	2.17 (70%)	2.13 (73%)	2.06
10.0	2.16 (90%)	2.10 (83%)	2.07 (87%)	1.96

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table App-17: Targeting Value Improvement Across Varying Privacy Levels: R-XGBoost

(a) Scenario 1: DP-Protected Covariates

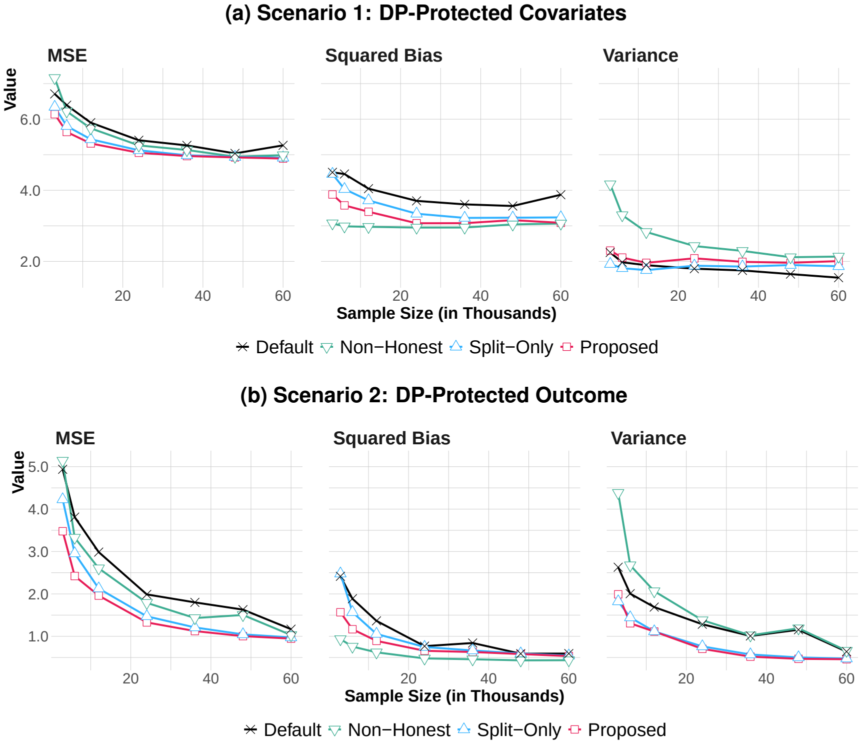
Privacy(ϵ)	Proposed	Split-only	Non-Honest
No	4.8% (72%)	4.1% (64%)	0.3% (40%)
50.0	14.1% (96%)	11.7% (84%)	4.1% (73%)
25.0	25.0% (98%)	22% (88%)	11.0% (92%)
16.7	18.2% (80%)	16% (76%)	4.0% (60%)
12.5	36.8% (88%)	28.7% (85%)	3.9% (63%)
10.0	40.3% (76%)	37.7% (79%)	−0.2% (36%)

(b) Scenario 2: DP-Protected Outcome

Privacy(ϵ)	Proposed	Split-only	Non-Honest
No	10.0% (93%)	7.3% (90%)	1.7% (63%)
50.0	9.3% (77%)	7.7% (70%)	2.5% (67%)
25.0	14.3% (90%)	10.5% (80%)	5.3% (70%)
16.7	21.4% (97%)	18.6% (97%)	9.1% (73%)
12.5	20.1% (95%)	13.1% (73%)	7.7% (73%)
10.0	35.7% (93%)	26.9% (82%)	18.6% (83%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Figure App-16: Statistical Accuracy Across Varying Sample Sizes: R-XGBoost


(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table App-18: AUTOC Values Across Varying Sample Sizes: R-XGBoost

(a) Scenario 1: DP-Protected Covariates

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	1.49 (80%)	1.48 (80%)	1.42 (72%)	1.38
6k	1.59 (92%)	1.57 (82%)	1.53 (88%)	1.46
12k	1.66 (92%)	1.64 (86%)	1.6 (84%)	1.54
24k	1.71 (83%)	1.7 (70%)	1.68 (90%)	1.66
36k	1.73 (81%)	1.72 (80%)	1.70 (80%)	1.68
48k	1.73 (85%)	1.73 (77%)	1.73 (83%)	1.72
60k	1.74 (94%)	1.74 (87%)	1.72 (88%)	1.70

(b) Scenario 2: DP-Protected Outcome

Sample Size	Proposed	Split-only	Non-Honest	Default
3k	2.16 (90%)	2.10 (83%)	2.07 (87%)	1.96
6k	2.30 (84%)	2.25 (72%)	2.25 (72%)	2.11
12k	2.37 (96%)	2.36 (96%)	2.36 (96%)	2.23
24k	2.46 (75%)	2.45 (74%)	2.45 (69%)	2.37
36k	2.49 (81%)	2.48 (75%)	2.48 (75%)	2.42
48k	2.51 (93%)	2.50 (85%)	2.50 (83%)	2.43
60k	2.52 (80%)	2.52 (77%)	2.52 (60%)	2.50

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table App-19: Targeting Value Improvement Across Varying Sample Sizes: R-XGBoost

(a) Scenario 1: DP-Protected Covariates

Sample Size	Proposed	Split-only	Non-Honest
3k	40.3% (76%)	37.7% (79%)	−0.2% (36%)
6k	51.6% (82%)	47.5% (84%)	23.8% (84%)
12k	27.6% (88%)	26.7% (84%)	11.7% (84%)
24k	11% (85%)	9.5% (80%)	4.2% (75%)
36k	9.0% (86%)	8.1% (78%)	2.5% (78%)
48k	8.5% (83%)	7.6% (76%)	1.4% (71%)
60k	7.7% (90%)	7.4% (90%)	4.8% (100%)

(b) Scenario 2: DP-Protected Outcome

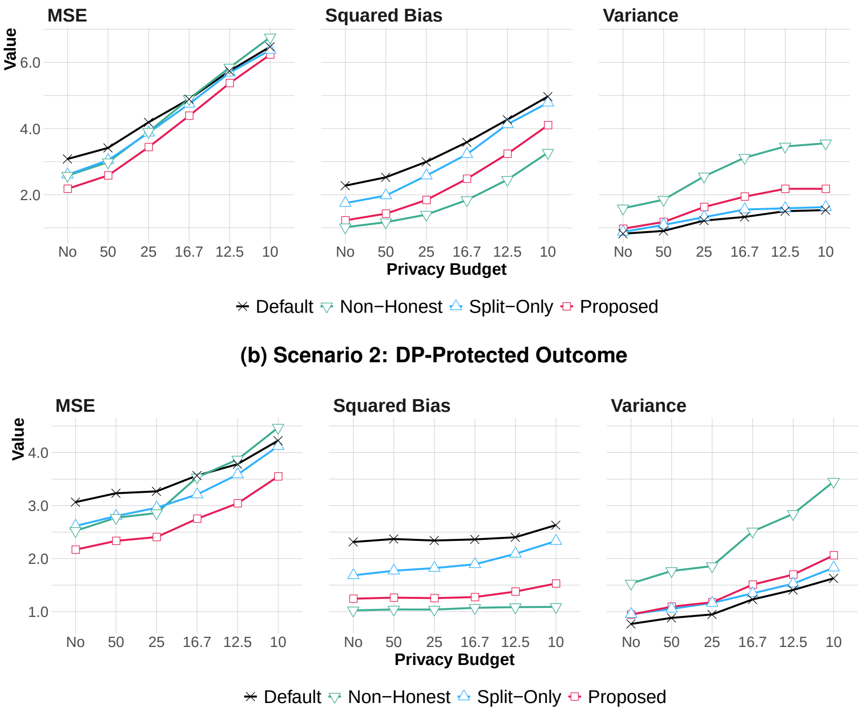
Sample Size	Proposed	Split-only	Non-Honest
3k	35.7% (93%)	26.9% (82%)	18.6% (83%)
6k	24.2% (96%)	17.3% (68%)	14.6% (80%)
12k	13.8% (88%)	12.6% (92%)	7.1% (84%)
24k	7.3% (86%)	6.2% (65%)	3.1% (85%)
36k	5% (83%)	4.3% (80%)	2.4% (75%)
48k	7.6% (80%)	7.1% (77%)	2.6% (90%)
60k	1.5% (75%)	1.30% (69%)	0.8% (60%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Web Appendix D.5.4 Results for DR-learner with XGBoost Models

Varying Privacy Levels. Figure App-17 presents the MSE, squared bias, and variance of different methods across different privacy levels, where we apply DR-lerarner with XGBoost models as the DEFAULT method and the initial CATE models.

Figure App-17: Statistical Accuracy Across Varying Privacy Levels: DR-XGBoost



(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table App-20 reports the AUTOC values across different privacy levels. Table App-21 presents the value improvement of different methods across various privacy levels. Overall, the results align with the findings from the main analysis in Section 5.5.

Table App-20: AUTOC Values Across Varying Privacy Levels: DR-XGBoost

(a) Scenario 1: DP-Protected Covariates

Privacy(ϵ)	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	2.34 (98%)	2.30 (88%)	2.31 (85%)	2.24
50.0	2.27 (98%)	2.22 (81%)	2.24 (92%)	2.17
25.0	2.11 (98%)	2.06 (79%)	2.08 (86%)	2.01
16.7	1.92 (97%)	1.88 (66%)	1.89 (83%)	1.86
12.5	1.72 (87%)	1.67 (63%)	1.7 0(87%)	1.65
10.0	1.51 (82%)	1.48 (65%)	1.48 (59%)	1.47

(b) Scenario 2: DP-Protected Outcome

Privacy(ϵ)	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	2.34 (98%)	2.30 (88%)	2.31 (85%)	2.24
50.0	2.32 (99%)	2.26 (85%)	2.28 (89%)	2.22
25.0	2.31 (96%)	2.25 (77%)	2.27 (86%)	2.21
16.7	2.26 (97%)	2.21 (84%)	2.21 (83%)	2.15
12.5	2.22 (96%)	2.14 (70%)	2.18 (82%)	2.11
10.0	2.14 (92%)	2.05 (62%)	2.10 (84%)	2.03

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-21: Targeting Value Improvement Across Varying Privacy Levels: DR-XGBoost

(a) Scenario 1: DP-Protected Covariates

Privacy(ϵ)	PROPOSED	SPLIT-ONLY	NON-HONEST
No	11.6% (98%)	6.6% (85%)	7.8% (82%)
50.0	13.0% (97%)	7.6% (82%)	9.7% (87%)
25.0	13.4% (96%)	6.4% (79%)	8.7% (83%)
16.7	11.9% (88%)	3.8% (64%)	4.8% (74%)
12.5	13.3% (85%)	5.3% (67%)	3.1% (72%)
10.0	14.3% (74%)	7.6% (66%)	2.5% (51%)

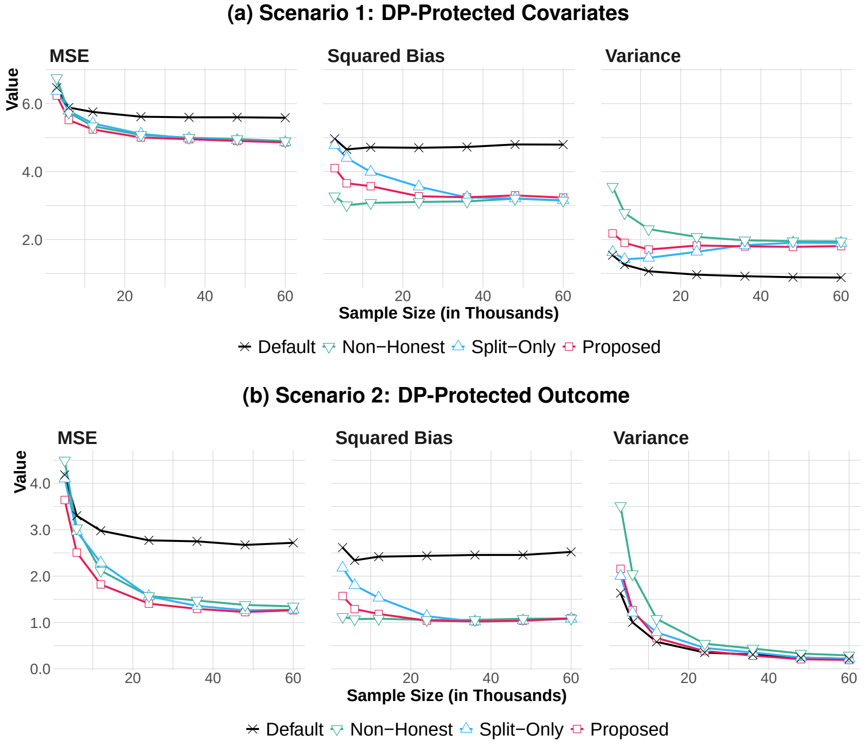
(b) Scenario 2: DP-Protected Outcome

Privacy(ϵ)	PROPOSED	SPLIT-ONLY	NON-HONEST
No	11.6% (98%)	6.6% (86%)	7.8% (83%)
50.0	14.6% (96%)	7.1% (84%)	11.3% (92%)
25.0	12.2% (98%)	4.8% (71%)	8.2% (82%)
16.7	16.4% (99%)	9.0% (73%)	9.3% (78%)
12.5	15.1% (96%)	4.5% (66%)	9.8% (74%)
10.0	14.0% (85%)	5.3% (70%)	7.6% (61%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of DR-XGBoost as the initial CATE model.

Varying Sample Sizes. Figure App-18 presents the MSE, squared bias, and variance of different methods across different sample sizes, where we apply DR-XGBoost method as the DEFAULT and the initial CATE models. Table App-22 reports the AUTOC values across different sample sizes. Table App-23 presents the value improvement of different methods across various sample sizes. Overall, the results align with the findings from the main analysis in Section 5.5.

Figure App-18: Statistical Accuracy Across Varying Sample Sizes: DR-XGBoost


(a) Scenario 1: DP-Protected Covariates

(b) Scenario 2: DP-Protected Outcome

Note: We simulate 100 replications to compute the bootstrap MSE for each individual in the holdout set. We then average the bootstrap mean over a holdout set of 10,000 individuals for each point.

Table App-22: AUTOC Values for Across Varying Sample Sizes: DR-XGBoost

(a) Scenario 1: DP-Protected Covariates

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
3k	1.51 (80%)	1.48 (66%)	1.48 (66%)	1.47
6k	1.65 (90%)	1.64 (74%)	1.62 (62%)	1.61
12k	1.72 (100%)	1.71 (92%)	1.70 (92%)	1.67
24k	1.76 (100%)	1.75 (100%)	1.75 (100%)	1.71
36k	1.77 (100%)	1.77 (100%)	1.76 (100%)	1.72
48k	1.78 (100%)	1.78 (100%)	1.77 (100%)	1.72
60k	1.79 (100%)	1.79 (100%)	1.78 (100%)	1.73

(b) Scenario 2: DP-Protected Outcome

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
3k	2.14 (92%)	2.05 (62%)	2.10 (84%)	2.03
6k	2.29 (96%)	2.25 (92%)	2.24 (80%)	2.19
12k	2.39 (100%)	2.35 (96%)	2.35 (100%)	2.28
24k	2.45 (100%)	2.43 (100%)	2.42 (100%)	2.35
36k	2.46 (100%)	2.45 (100%)	2.43 (100%)	2.36
48k	2.47 (100%)	2.46 (100%)	2.44 (100%)	2.37
60k	2.46 (100%)	2.46 (100%)	2.45 (100%)	2.38

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of DR-XGBoost as the initial CATE model.

Table App-23: Targeting Value Improvement Across Varying Sample Sizes: DR-XGBoost

(a) Scenario 1: DP-Protected Covariates

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST
3k	17.5% (94%)	-4.8% (44%)	-11.6% (30%)
6k	15.7% (93%)	4.7% (70%)	0.2% (53%)
12k	14.2% (100%)	7.7% (92%)	1.6% (56%)
24k	10.9% (100%)	7.5% (95%)	1.5% (72%)
36k	11.2% (100%)	8.9% (100%)	2.0% (83%)
48k	9.4% (100%)	7.1% (100%)	1.0% (88%)
60k	10.0% (100%)	9.1% (100%)	1.2% (84%)

(b) Scenario 2: DP-Protected Outcome

Sample Size	PROPOSED	SPLIT-ONLY	NON-HONEST
3k	13.2% (100%)	3.4% (76%)	0.8% (56%)
6k	11.5% (100%)	6.8% (84%)	2.3% (72%)
12k	8.8% (100%)	5.1% (96%)	2.4% (76%)
24k	7.1% (100%)	5.9% (100%)	1.2% (80%)
36k	5.3% (100%)	4.1% (100%)	1.7% (90%)
48k	4.9% (100%)	4.0% (100%)	1.2% (90%)
60k	4.3% (100%)	4.2% (100%)	0.5% (70%)

Note: We calculate the value improvement from 100 simulation replications, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of DR-XGBoost as the initial CATE model.

Web Appendix E Further Details about Hillstrom Case Study

In this section, we provide the summary statistics, implementation details, and robustness checks for the Hillstrom case study in Section 6.4.

Web Appendix E.1 Summary Statistics

Table App-24 presents the summary statistics for the Hillstrom data. The definitions of the pre-treatment covariates are:

1. Recency: The number of months since the customer's last purchase.
2. History: The actual dollar value the customer has spent in the past year.
3. Mens: A binary variable indicating whether the customer purchased men's merchandise in the past year (1 = Yes).
4. Womens: A binary variable indicating whether the customer purchased women's merchandise in the past year (1 = Yes).

5. Zip_Code: A classification of the customer's zip code as Urban, Suburban, or Rural
6. Newbie: A binary variable indicating whether the customer was acquired in the past twelve months (1 = Yes).
7. Channel: The channel(s) the customer purchased from in the past year.

Table App-24: Summary Statistics for Hillstrom Data

Discrete Variables

Variable	N	Unique Values	Distributions
visit (Outcome)	42,693	2	0: 37,193, 1: 5,500
email (Treatment)	42,693	2	0: 21,306, 1: 21,387
mens	42,693	2	0: 19,166, 1: 23,527
womens	42,693	2	0: 19,260, 1: 23,433
newbie	42,693	2	0: 21,235, 1: 21,458
channel	42,693	3	Phone: 18,781, Website: 18,727, Multichannel: 5,185
zip_code	42,693	3	Suburban: 19,275, Urban 17,098, Rural: 6,320

Continuous Variables

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
history	42,693	241.71	254.04	29.99	65.16	158.46	326.05	3345.93
newbie	42,693	0.5	0.5	0	0	1	1	1
recency	42,693	5.76	3.5	1	2	5	9	12

Web Appendix E.2 Covariate Balance Check

To assess covariate balance, we compare the distributions of each covariate for both treated and non-treated customers. In particular, we use the *standardized mean difference* measure, which is the mean difference between the treated and non-treated groups divided by the pooled standard deviation. Generally, it is considered small if the value is less than 0.20 (Cohen 2013). Table App-25 reports the summary for the covariate balance check. Note that the standardized mean differences are close to zero for all covariates, suggesting that the experiment is properly randomized.

We also conduct a linear regression test to assess whether any covariate predicts the treatment assignment indicator. The joint F-test indicates that the model has no predictive power (F-stat = 0.4292, p-value = 0.9202), suggesting that randomization was successfully implemented.

Table App-25: Covariate Balance Check for Hillstrom Data

Variable	Mean Diff.	Pooled St. Dev.	Standardized Mean Diff.
recency	0.021	3.505	0.006
history	1.803	255.307	0.007
mens	−0.003	0.497	−0.007
womens	0.003	0.498	0.006
newbie	0.000	0.500	0.001
channel_Multichannel	−0.002	0.327	−0.005
channel_Phone	0.000	0.496	0.000
channel_Web	0.001	0.496	0.003
zip_code_Rural	0.003	0.356	0.009
zip_code_Suburban	−0.003	0.498	−0.006
zip_code_Urban	0.000	0.490	0.000

Web Appendix E.3 Implementation of Differential Privacy

We examine two scenarios for implementing differential privacy: (1) DP-protected covariates and (2) a DP-protected outcome.

In the first scenario, we add Laplace noise to continuous covariates and apply randomized response to discrete variables. Specifically, for the continuous variable `recency`, we add noise drawn from a Laplace distribution with scale parameter $\sigma_{\texttt{recency}} \in \{2, 4, 6, 8, 10\}$, which corresponds to privacy budgets $\epsilon \in \{6.0, 3.0, 2.0, 1.5, 1.2\}$ under the Laplace mechanism. For the continuous variable `history`, we similarly apply Laplace noise with scale parameter $\sigma_{\texttt{history}} \in \{20, 40, 60, 80, 100\}$, which corresponds to privacy budgets $\epsilon \in \{59.2, 29.6, 19.7, 14.8, 11.8\}$. For all discrete covariates, we implement the randomized response mechanism, varying the probability parameter $p \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, which corresponds to privacy budgets $\epsilon \in \{3.66, 2.94, 2.51, 2.20, 1.95\}$.

In the second scenario, we protect the binary outcome variable using the same randomized response mechanism, again varying p across the same set of values to control the privacy level.

Web Appendix E.4 Model Specification

In the main analysis, we use the R-learner with regression forest models models and five-fold cross-fitting for the DEFAULT method and initial CATE models. We use a constant propensity score (0.5) for Robinson's transformation. For all regression forest models, we use 500 trees instead of the default 2,000 trees to accelerate training, while keeping all other parameters at their default settings in the `grf` package.

To construct the DR score for model calibration, we set a constant propensity score ($\widehat{e} = 0.5$), considering that the experiment is completely randomized. As for the conditional mean outcome models, we evaluate three candidate models: linear regression, logistic regression, and regression forest. The data is split into two sets: 70% allocated as the training set and the remaining 30% serving as the holdout set. The mean-squared error (MSE) metric is reported in Table App-26, calculated as $\frac{1}{N_{\text{holdout}}} \sum_{l \in \text{holdout set}} [Y_i - \widehat{m}_{W_l}(\widetilde{\mathbf{X}}_l)]^2$. Linear regression is chosen as our final model since it yields the smallest MSE.

Table App-26: MSE of Conditional Mean Outcome Models

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	Linear Regression	Logistic Regression	Regression Forest	Linear Regression	Logistic	Regression Forest
No	0.1088	0.1089	0.1099	0.1088	0.1089	0.1099
Very Low	0.1094	0.1094	0.1098	0.1118	0.1119	0.1128
Low	0.1118	0.1119	0.1123	0.1107	0.1109	0.1122
Medium	0.1092	0.1092	0.1093	0.1184	0.1184	0.1205
High	0.1155	0.1155	0.1159	0.1186	0.1186	0.1204
Very High	0.1108	0.1109	0.1114	0.1202	0.1202	0.1220

For the calibration models in the model calibration procedures, we use linear regression to ensure simplicity and computational efficiency. We set the number of subgroups to 10 and the maximum number of iterations to 20 for the subgroup cross-learning algorithm.

Web Appendix E.5 Small Sample Performance

We demonstrate that in small-sample settings, the PROPOSED method outperforms the SPLIT-ONLY method. Instead of using 70% for model construction and 30% for holdout evaluation, we now only use 10% of the data to construct models and 90% to evaluate the performance.

Table App-27 presents the RMSE across varying privacy levels. The results show that the PROPOSED method significantly outperforms the SPLIT-ONLY method, with both improving accuracy compared to the DEFAULT approach. Additionally, the NON-HONEST model calibration method can even degrade performance.

Table App-27: RMSE of GATE (Multiplied by 100) Across Varying Privacy Levels: Small Samples

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	4.00 (100%)	4.13 (100%)	7.20 (4%)	6.88	4.00 (100%)	4.13 (100%)	7.20 (4%)	6.88
Very Low	4.02 (100%)	4.19 (100%)	6.62 (24%)	6.46	4.26 (100%)	4.38 (100%)	7.55 (16%)	7.24
Low	3.90 (100%)	4.06 (100%)	6.46 (30%)	6.33	4.66 (100%)	4.69 (100%)	8.21 (4%)	7.92
Medium	4.01 (100%)	4.09 (100%)	6.52 (16%)	6.35	5.07 (100%)	5.14 (100%)	8.75 (4%)	8.31
High	3.90 (100%)	4.07 (100%)	6.45 (20%)	6.28	5.23 (100%)	5.36 (100%)	9.06 (4%)	8.74
Very High	3.92 (100%)	4.03 (100%)	6.43 (24%)	6.26	5.21 (100%)	5.41 (100%)	9.19 (0%)	8.84

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-28 reports the AUTOC values across different privacy levels. Consistent with previous findings, the PROPOSED method outperforms the SPLIT-ONLY method, with both achieving better accuracy than the DEFAULT approach. While the NON-HONEST model calibration method also enhances treatment prioritization ability, it suffers from high predictive error.

Table App-28: AUTOC (Multiplied by 100) Across Varying Privacy Levels: Small Samples

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	1.08 (96%)	1.01 (92%)	0.87 (96%)	0.70	1.08 (96%)	1.01 (92%)	0.87 (96%)	0.70
Very Low	1.00 (94%)	1.00 (94%)	0.83 (100%)	0.62	0.83 (96%)	0.84 (96%)	0.76 (100%)	0.59
Low	0.86 (96%)	0.85 (94%)	0.72 (100%)	0.54	0.63 (88%)	0.66 (76%)	0.64 (92%)	0.45
Medium	0.83 (84%)	0.76 (80%)	0.77 (98%)	0.59	0.55 (80%)	0.51 (64%)	0.50 (96%)	0.34
High	0.78 (90%)	0.74 (74%)	0.68 (84%)	0.49	0.56 (82%)	0.52 (80%)	0.53 (82%)	0.40
Very High	0.72 (90%)	0.67 (76%)	0.64 (86%)	0.47	0.50 (86%)	0.49 (72%)	0.45 (88%)	0.25

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-29 reports the average value improvement across different privacy levels. The results indicate that both the PROPOSED method and the SPLIT-ONLY method consistently outperform the default approach, with the proposed method achieving slightly better performance than the split-only method.

Table App-29: Targeting Value Improvement Across Varying Privacy Levels: Small Samples

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	PROPOSED	SPLIT-ONLY	NON-HONEST	PROPOSED	SPLIT-ONLY	NON-HONEST
No	2.26% (94%)	2.22% (88%)	-0.03% (56%)	3.02% (100%)	2.98% (96%)	0.29% (60%)
Very Low	2.75% (98%)	2.42% (90%)	0.29% (72%)	2.96% (100%)	2.91% (100%)	0.33% (64%)
Low	2.69% (98%)	2.43% (92%)	0.21% (66%)	3.02% (100%)	3.08% (100%)	0.13% (60%)
Medium	2.71% (100%)	2.44% (92%)	0.2% (66%)	3.27% (100%)	3.11% (94%)	0.39% (68%)
High	2.95% (98%)	2.53% (94%)	0.26% (68%)	2.89% (96%)	2.60% (92%)	0.29% (68%)
Very High	2.61% (94%)	2.43% (92%)	0.20% (64%)	3.00% (96%)	2.56% (96%)	0.36% (76%)

Note: We calculate the value improvement from 50 splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-XGBoost models as the initial CATE model.

Web Appendix E.6 Robustness Checks of Other DEFAULT Models

Web Appendix E.6.1 Forest-based Models

We begin by considering two alternative forest-based CATE models as the DEFAULT method and the inital CATE models for the model calibration algorithms: Causal Forest and DR-learner with regression forests. Results for the T-learner (Künzel et al. 2019) are also available upon request and show consistent patterns with the findings reported here.

1. **(Causal Forest)** We use the causal forest function implemented in the `grf` package with 500 trees and other default parameters. Note that we choose 500 trees instead the default 2,000 trees to accelerate the model training process.

2. **(DR-learner)** We construct both the DR score and the CATE model using regression forests. To speed up training, we reduce the number of trees from the default 2,000 to 500, while keeping all other parameters at their default settings in the `grf` package.

Table App-30 reports the RMSE for the two CATE models. The results indicate that (i) the PROPOSED method significantly outperforms all other methods, (ii) the SPLIT-ONLY method also improves upon the DEFAULT method, while (iii) the NON-HONEST model calibration method fail to enhance predictive performance.

Table App-31 reports the AUTOC values (multiplied by 100) for different methods across various privacy levels. Consistent with the main findings in Section 6 of the paper, the PROPOSED method outperforms all other methods, and all the model calibration methods outperform the DEFAULT approach.

Table App-30: RMSE of GATE (Multiplied by 100) Across Varying Privacy Levels

(a) Causal Forest

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	2.56 (100%)	2.54 (96%)	3.54 (44%)	3.41	2.56 (100%)	2.54 (96%)	3.54 (44%)	3.41
Very Low	2.30 (96%)	2.42 (100%)	3.27 (20%)	3.13	2.68 (100%)	2.62 (96%)	3.9 (28%)	3.73
Low	2.35 (98%)	2.36 (94%)	3.32 (12%)	3.16	2.41 (96%)	2.46 (96%)	3.67 (32%)	3.56
Medium	2.37 (98%)	2.32 (100%)	3.39 (28%)	3.23	2.56 (100%)	2.62 (96%)	3.95 (28%)	3.89
High	2.18 (100%)	2.33 (92%)	3.09 (28%)	2.97	2.49 (100%)	2.65 (100%)	4.29 (44%)	4.29
Very High	2.27 (100%)	2.26 (100%)	3.37 (12%)	3.13	2.78 (100%)	2.73 (100%)	4.39 (36%)	4.33

(b) DR-learner with Regression Forests

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	3.96 (100%)	4.05 (100%)	7.23 (38%)	7.23	3.96 (100%)	4.05 (100%)	7.23 (38%)	7.23
Very Low	3.89 (100%)	3.93 (100%)	7.16 (36%)	7.13	4.05 (100%)	4.17 (100%)	7.63 (54%)	7.61
Low	3.59 (100%)	3.66 (100%)	5.88 (40%)	5.85	4.46 (100%)	4.49 (100%)	8.17 (29%)	8.11
Medium	3.41 (100%)	3.42 (100%)	5.51 (24%)	5.47	4.82 (100%)	4.81 (100%)	8.72 (42%)	8.72
High	3.51 (100%)	3.56 (100%)	5.61 (64%)	5.63	4.73 (100%)	4.75 (100%)	8.73 (50%)	8.70
Very High	3.57 (100%)	3.59 (100%)	5.65 (24%)	5.58	5.24 (100%)	5.33 (100%)	9.52 (42%)	9.50

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table App-31: AUTOC Values (Multiplied by 100) Across Varying Privacy Levels

(a) Causal Forest

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	1.71 (60%)	1.68 (68%)	1.57 (48%)	1.59	1.71 (60%)	1.68 (68%)	1.57 (48%)	1.59
Very Low	1.55 (96%)	1.52 (84%)	1.44 (96%)	1.26	1.41 (76%)	1.43 (68%)	1.29 (60%)	1.26
Low	1.30 (100%)	1.32 (88%)	1.22 (96%)	1.03	1.56 (68%)	1.55 (72%)	1.45 (64%)	1.43
Medium	1.30 (96%)	1.21 (80%)	1.17 (96%)	0.97	1.27 (88%)	1.27 (84%)	1.18 (92%)	1.07
High	1.36 (92%)	1.29 (92%)	1.24 (98%)	0.99	1.26 (96%)	1.25 (96%)	1.02 (96%)	0.89
Very High	1.11 (100%)	1.07 (88%)	1.09 (94%)	0.89	1.06 (88%)	1.02 (78%)	1.00 (82%)	0.85

(b) DR-learner with Regression Forests

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	1.26 (72%)	1.27 (62%)	1.12 (62%)	1.11	1.26 (72%)	1.27 (62%)	1.12 (62%)	1.11
Very Low	1.23 (64%)	1.26 (72%)	1.14 (60%)	1.13	1.24 (76%)	1.25 (80%)	1.05 (58%)	1.03
Low	1.02 (88%)	1.03 (88%)	0.75 (76%)	0.71	1.06 (72%)	1.06 (62%)	0.96 (66%)	0.95
Medium	1.11 (82%)	1.12 (72%)	0.96 (84%)	0.90	0.86 (78%)	0.88 (84%)	0.75 (88%)	0.72
High	0.97 (88%)	0.98 (90%)	0.72 (84%)	0.63	0.96 (68%)	0.96 (62%)	0.91 (76%)	0.85
Very High	0.85 (78%)	0.86 (76%)	0.70 (66%)	0.61	0.66 (70%)	0.66 (70%)	0.57 (68%)	0.51

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Table App-32 reports the average value improvement across different privacy levels. The results indicate that both the Proposed method and the Split-Only method consistently outperform the default approach, with the proposed method achieving slightly better performance than the split-only method.

Table App-32: Targeting Value Improvement Across Varying Privacy Levels

(a) Causal Forest

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	Proposed	Split-only	Non-Honest	Proposed	Split-only	Non-Honest
No	1.79% (96%)	1.67% (92%)	0.14% (68%)	1.79% (96%)	1.67% (92%)	0.14% (68%)
Very Low	1.4% (100%)	1.35% (96%)	0.11% (60%)	1.77% (92%)	1.65% (88%)	0.24% (60%)
Low	0.87% (84%)	0.90% (84%)	-0.13% (44%)	1.76% (84%)	1.59% (88%)	-0.04% (56%)
Medium	1.31% (96%)	1.54% (96%)	-0.11% (44%)	1.99% (92%)	2.19% (100%)	0.08% (48%)
High	1.01% (92%)	0.98% (92%)	-0.15% (28%)	2.63% (92%)	2.53% (96%)	-0.08% (36%)
Very High	1.40% (96%)	1.30% (92%)	-0.15% (52%)	2.85% (100%)	2.75% (100%)	0.38% (64%)

(b) DR-learner with Regression Forests

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	Proposed	Split-only	Non-Honest	Proposed	Split-only	Non-Honest
No	0.97% (96%)	0.97% (96%)	-0.01% (42%)	0.97% (96%)	0.97% (96%)	-0.01% (42%)
Very Low	2.54% (94%)	2.29% (92%)	0.19% (56%)	1.30% (94%)	1.32% (96%)	-0.02% (42%)
Low	2.44% (96%)	2.39% (100%)	0.01% (54%)	1.44% (100%)	1.41% (96%)	-0.03% (46%)
Medium	2.57% (92%)	2.34% (92%)	0.34% (76%)	1.57% (96%)	1.55% (94%)	-0.02% (50%)
High	3.13% (100%)	2.95% (94%)	0.16% (56%)	1.51% (92%)	1.50% (90%)	0.02% (50%)
Very High	3.13% (100%)	3.07% (96%)	0.19% (64%)	2.10% (96%)	2.07% (96%)	-0.11% (38%)

Note: We calculate the value improvement from 50 splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

Web Appendix E.6.2 XGBoost-based Models

Next, we consider two additional boosting-based CATE models: the R-learner and the DR-learner. Results for the T-learner are available upon request and are consistent with the findings reported here. To fine-tune the hyperparameters in XGBoost models, we perform a single train-holdout split under the non-privacy setting. For conditional mean outcome models, we select parameters that minimize mean squared prediction error, while for the CATE model in R-learner and DR-learner, we choose parameters that maximize the AUTOC value. The search ranges for key hyperparameters include: learning rate $\eta \in \{0.20, 0.40, 0.60, 0.80, 1.00\}$, maximum depth in each tree $\{2, 4, 6, 8, 10\}$, and the maximum number of iterations $\{10, 20, 30, 40, 50\}$. The optimal hyperparameters are as follows:

1. **(R-learner)** Conditional mean outcome model: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 40; CATE model: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 10.

2. **(DR-learner)**: Conditional mean outcome model for the treatment group: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 20; Conditional mean outcome model for the control group: $\eta = 0.20$, maximum tree depth: 2, the maximum number of

iterations: 10; CATE model: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 10.

Table App-33 reports the RMSE for the two CATE models. Table App-34 reports the AUTOC values (multiplied by 100) for different methods across various privacy levels. The findings are consistent with the main results in Section 6.

Table App-33: RMSE of GATE (Multiplied by 100) Across Varying Privacy Levels

(a) R-XGBoost Model

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	2.16 (100%)	2.18 (100%)	2.08 (100%)	4.82	2.16 (100%)	2.18 (100%)	2.08 (100%)	4.82
Very Low	2.16 (100%)	2.18 (100%)	2.08 (100%)	5.02	2.16 (100%)	2.26 (100%)	2.34 (100%)	4.85
Low	2.16 (100%)	2.26 (100%)	2.34 (100%)	4.85	2.18 (98%)	2.39 (96%)	2.26 (100%)	4.88
Medium	2.18 (96%)	2.39 (96%)	2.26 (100%)	4.78	2.05 (100%)	2.25 (100%)	2.45 (100%)	4.85
High	2.05 (100%)	2.25 (100%)	2.45 (100%)	4.65	2.17 (100%)	2.33 (100%)	2.38 (100%)	4.88
Very High	2.17 (100%)	2.33 (100%)	2.38 (100%)	4.78	2.29 (100%)	2.32 (100%)	2.51 (100%)	4.92

(b) DR-learners with XGBoost Models

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	2.10 (100%)	2.14 (100%)	2.03 (100%)	5.09	2.10 (100%)	2.14 (100%)	2.03 (100%)	5.09
Very Low	2.34 (100%)	2.16 (100%)	2.26 (100%)	5.27	2.19 (100%)	2.16 (100%)	2.28 (100%)	5.09
Low	1.89 (100%)	2.01 (100%)	1.97 (100%)	4.98	2.11 (100%)	2.21 (100%)	2.23 (100%)	4.66
Medium	2.16 (100%)	2.07 (100%)	2.15 (100%)	5.25	2.09 (100%)	2.13 (100%)	2.37 (100%)	4.83
High	1.98 (100%)	2.08 (100%)	2.09 (100%)	5.30	2.12 (98%)	2.42 (100%)	2.54 (100%)	4.56
Very High	2.06 (100%)	2.04 (100%)	2.12 (100%)	5.24	2.09 (100%)	2.28 (96%)	2.59 (96%)	4.86

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-34: AUTOC Values (Multiplied by 100) Across Varying Privacy Levels

(a) R-learners with XGBoost Models

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	1.78 (72%)	1.79 (84%)	1.78 (80%)	1.55	1.78 (72%)	1.79 (84%)	1.78 (80%)	1.55
Very Low	1.81 (96%)	1.83 (96%)	1.78 (96%)	1.45	1.67 (74%)	1.68 (76%)	1.61 (90%)	1.58
Low	1.71 (84%)	1.74 (88%)	1.77 (96%)	1.43	1.77 (72%)	1.75 (60%)	1.78 (80%)	1.61
Medium	1.57 (72%)	1.61 (80%)	1.57 (76%)	1.42	1.68 (68%)	1.67 (76%)	1.64 (68%)	1.55
High	1.55 (68%)	1.60 (76%)	1.62 (76%)	1.41	1.61 (64%)	1.65 (60%)	1.69 (68%)	1.56
Very High	1.54 (67%)	1.60 (75%)	1.58 (83%)	1.39	1.62 (76%)	1.67 (68%)	1.66 (76%)	1.49

(b) DR-learners with XGBoost Models

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	1.80(80%)	1.78 (84%)	1.78 (92%)	1.49	1.80(80%)	1.78 (84%)	1.78 (92%)	1.49
Very Low	1.72 (86%)	1.74 (92%)	1.66 (82%)	1.49	1.71 (72%)	1.69 (74%)	1.68 (76%)	1.58
Low	1.69 (76%)	1.67 (64%)	1.68 (72%)	1.46	1.82 (84%)	1.72 (68%)	1.71 (76%)	1.51
Medium	1.68 (88%)	1.68 (82%)	1.64 (96%)	1.47	1.69 (78%)	1.61 (72%)	1.57 (72%)	1.46
High	1.53 (84%)	1.55 (88%)	1.49 (88%)	1.35	1.67 (76%)	1.64 (78%)	1.63 (72%)	1.53
Very High	1.54 (86%)	1.52 (76%)	1.45 (84%)	1.30	1.69 (74%)	1.64 (62%)	1.60 (72%)	1.50

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-47 presents the average value improvement across different methods. Overall, the calibrated models do not produce significant gains over the DEFAULT model. This is because all models, including the DEFAULT and the various calibrated versions, predict positive treatment effects for over 99.9% of customers. As a result, when targeting is based solely on whether the predicted CATE is positive, all models recommend targeting nearly the entire customer base. Nevertheless, the substantial improvements in RMSE and AUTOC highlight the value of honest model calibrations. This benefit becomes particularly important in practical applications where targeting is constrained—such as when firms can only reach a subset of customers due to budget limitations or strategic priorities.

Table App-35: Targeting Value Improvement Across Varying Privacy Levels

(a) R-learner with XGBoost Models

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	PROPOSED	SPLIT-ONLY	NON-HONEST	PROPOSED	SPLIT-ONLY	NON-HONEST
No	0.01% (54%)	0.01% (54%)	0.00% (52%)	0.01% (54%)	0.01% (54%)	0.00% (52%)
Very Low	0.18% (52%)	-0.1% (56%)	0.05% (56%)	-0.02% (48%)	-0.01% (46%)	0.00% (56%)
Low	-0.06% (50%)	-0.06% (44%)	-0.08% (48%)	-0.02% (50%)	-0.01% (48%)	-0.01% (48%)
Medium	-0.03% (44%)	0.05% (56%)	0.10% (50%)	0.02% (60%)	-0.01% (56%)	0.01% (50%)
High	-0.14% (46%)	-0.35% (43%)	-0.01% (52%)	0.01% (58%)	-0.01% (64%)	-0.01% (52%)
Very High	0.11% (56%)	-0.10% (44%)	-0.01% (60%)	-0.01% (48%)	-0.03% (56%)	-0.01% (60%)

(b) DR-learner with XGBoost Models

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	PROPOSED	SPLIT-ONLY	NON-HONEST	PROPOSED	SPLIT-ONLY	NON-HONEST
No	0.03% (60%)	0.02% (54%)	0.01% (58%)	0.03% (60%)	0.02% (54%)	0.01% (58%)
Very Low	0.02% (60%)	0.01% (64%)	-0.01% (62%)	-0.02% (46%)	-0.02% (46%)	-0.01% (52%)
Low	0.00% (58%)	0.00% (54%)	0.01% (52%)	0.01% (58%)	0.01% (60%)	0.01% (54%)
Medium	0.01% (52%)	-0.02% (44%)	0.00% (54%)	0/02% (64%)	0.00% (50%)	0.02% (52%)
High	-0.01% (48%)	-0.02% (42%)	0.01% (58%)	-0.01% (64%)	-0.02% (56%)	0.00% (48%)
Very High	0.00% (68%)	0.01% (52%)	-0.02% (64%)	0.03% (58%)	0.02% (56%)	0.00% (56%)

Note: We calculate the value improvement from 50 splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Web Appendix F Further Details about Starbucks Case Study

In this section, we provide the summary statistics, implementation details, and robustness checks for the Starbucks case study in Section 6.4.

Web Appendix F.1 Summary Statistics

Table App-36 provides summary statistics for the Starbucks data. Note that there are five categorical and two continuous pre-treatment covariates.

Table App-36: Summary Statistics for Starbucks Data

Discrete Variables

Variable	N	Unique Values	Distribution
Purchase (Outcome)	126,184	2	0: 124,664, 1: 1,520
Promotion (Treatment)	126,184	2	0: 63,072, 1: 63,112
V1	126,184	4	0: 15,846, 1: 47,410, 2: 47,134, 3: 15,794
V4	126,184	2	1: 40,379, 2: 85,805
V5	126,184	4	1: 23,179, 2: 46,597, 3: 48,643, 4: 7,765
V6	126,184	4	1: 31,435, 2: 31,420, 3: 3,1651, 4: 31,678
V7	126,184	2	1: 37,545, 2: 88,639

Continuous Variables

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
V2	126,184	29.98	5.00	7.10	26.596	29.98	33.354	55.108
V3	126,184	0.00	1.00	−1.69	−0.91	−0.04	0.83	1.69

Web Appendix F.2 Covariate Balance Check

Table App-37 reports the summary statistics for the covariate balance check. The result suggests that the experiment is properly randomized as the standardized mean differences of all the covariates are close to zero.

Table App-37: Covariate Balance Check for Starbucks Data

Variable	Mean Diff.	Pooled St. Dev.	Standardized Mean Diff.
V1 = 0	−0.002	0.331	−0.005
V1 = 1	−0.003	0.484	−0.006
V1 = 2	0.002	0.484	0.004
V1 = 3	0.003	0.331	0.009
V2	−0.011	5.001	−0.002
V3	0.012	1.000	0.012
V4 = 1	−0.001	0.466	−0.002
V4 = 2	0.001	0.466	0.002
V5 = 1	0.003	0.387	0.007
V5 = 2	−0.001	0.483	−0.001
V5 = 3	−0.001	0.487	−0.001
V5 = 4	−0.002	0.240	−0.006
V6 = 1	0.000	0.433	0.000
V6 = 2	0.000	0.432	0.000
V6 = 3	0.001	0.433	0.001
V6 = 4	−0.001	0.434	−0.002
V7 = 1	−0.001	0.457	−0.002
V7 = 2	0.001	0.457	0.002

We also conduct a linear regression test to assess whether any covariate predicts the treatment assignment indicator. The joint F-test indicates that the model has no predictive power (F-stat = 0.8966, p-value = 0.5560), suggesting that randomization was successfully implemented.

Web Appendix F.3 Implementation of Differential Privacy

In the scenario when the outcome variable is protected by DP, we apply the randomized response mechanism, setting p to values in the set $\{0.05, 0.10, 0.15, 0.20, 0.25\}$, which corresponds to $\epsilon \in \{3.66, 2.94, 2.51, 2.20, 1.95\}$. In the scenario when covaraites are protected by DP, we deploy the Laplace mechanism for the `V2` variable, setting $\sigma_{\texttt{V2}}$ to values in the set $\{5, 10, 15, 20, 25\}$ (which corresponds to $\epsilon \in \{8.40, 4.20, 2.80, 2.10, 1.68\}$), and for the `V3` variable, setting $\sigma_{\texttt{V3}}$ to values in the set $\{1, 2, 3, 4, 5\}$ (which corresponds to $\epsilon \in \{3.50, 1.75, 1.17, 0.88, 0.70\}$). For all discrete variables, we implement the randomized response mechanism, with p in the set $\{0.08, 0.16, 0.24, 0.32, 0.40\}$ (which corresponds to $\epsilon \in \{3.18, 2.44, 1.94, 1.65, 1.39\}$).

Web Appendix F.4 Model Specification

In the main analysis, we use the R-learner with regression forest models models and five-fold cross-fitting for the DEFAULT method and initial CATE models. We use a constant propensity score (0.5) for Robinson's transformation. For all regression forest models, we use 500 trees instead of the default 2,000 trees to accelerate training, while keeping all other parameters at their default settings in the `grf` package.

To construct the DR score for model calibration, we set a constant propensity score ($\widehat{e} = 0.5$), considering that the experiment is completely randomized. As for the conditional mean outcome models, we evaluate three candidate models: linear regression, logistic regression, and regression forest. The data is split into two sets: 70% allocated as the training set and the remaining 30% serving as the holdout set. The mean-squared error (MSE) metric is reported in Table App-38, calculated as $\frac{1}{N_{\text{holdout}}} \sum_{l \in \text{holdout set}} [Y_i - \widehat{m}_{W_l}(\widetilde{\mathbf{X}}_l)]^2$. Linear regression is chosen as our final model since it yields the smallest MSE.

Table App-38: MSE of Conditional Mean Outcome Models

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	Linear Regression	Logistic Regression	Regression Forest	Linear Regression	Logistic	Regression Forest
No	0.0119	0.0119	0.0119	0.0119	0.0119	0.0119
Very Low	0.0119	0.0119	0.0119	0.0360	0.0360	0.0360
Low	0.0114	0.0114	0.0114	0.0559	0.0559	0.0560
Medium	0.0123	0.0123	0.0123	0.0766	0.0766	0.0768
High	0.0118	0.0118	0.0118	0.0981	0.0981	0.0982
Very High	0.0117	0.0117	0.0117	0.1160	0.1160	0.1164

For the calibration models in the model calibration procedures, we use linear regression to ensure simplicity and computational efficiency. We set the number of subgroups to 5 and the maximum number of iterations to 5 for the subgroup cross-learning algorithm.

Web Appendix F.5 Small Sample Performance

Here, we demonstrate that in small-sample settings, the PROPOSED method outperforms the SPLIT-ONLY method. Instead of using 70% for model construction and 30% for holdout evaluation, we only use 10% of the data to construct models and 90% to evaluate the performance.

Table App-39 presents the RMSE across varying privacy levels. The results show that the PROPOSED method significantly outperforms the SPLIT-ONLY method, with both improving accuracy compared to the DEFAULT approach. Additionally, the NON-HONEST model calibration method can even degrade performance, underscoring the importance of our PROPOSED method in small-sample settings.

Table App-39: RMSE of GATE (Multiplied by 100) Across Varying Privacy Levels

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	1.00 (100%)	1.03 (100%)	1.72 (20%)	1.70	1.00 (100%)	1.03 (100%)	1.72 (20%)	1.70
Very Low	0.95 (100%)	0.98 (100%)	1.67 (8%)	1.60	1.86 (100%)	1.88 (100%)	3.15 (16%)	3.12
Low	0.98 (100%)	0.99 (100%)	1.67 (16%)	1.62	2.45 (100%)	2.51 (100%)	4.12 (28%)	4.06
Medium	1.01 (100%)	1.05 (100%)	1.71 (16%)	1.64	2.94 (100%)	2.98 (100%)	4.86 (12%)	4.78
High	1.03 (100%)	1.04 (100%)	1.72 (4%)	1.65	3.29 (100%)	3.45 (100%)	5.42 (4%)	5.35
Very High	1.05 (100%)	1.06 (100%)	1.70 (4%)	1.63	3.60 (100%)	3.69 (100%)	6.00 (0%)	5.91

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-40 reports the AUTOC values across different privacy levels. Consistent with previous findings, the PROPOSED method outperforms the SPLIT-ONLY method, with both achieving better accuracy than the DEFAULT approach. While the NON-HONEST model calibration method also enhances treatment prioritization ability, it suffers from higher predictive error. Table App-41 reports the average value improvement across different privacy levels. The results indicate that both the PROPOSED method and the SPLIT-ONLY method consistently outperform the default approach, with the proposed method achieving slightly better performance than the split-only method.

Table App-40: AUTOC (Multiplied by 100) Across Varying Privacy Levels: Small Sample

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	0.24 (92%)	0.26 (96%)	0.24 (100%)	0.18	0.24 (92%)	0.26 (96%)	0.24 (100%)	0.18
Very Low	0.27 (92%)	0.23 (76%)	0.25 (88%)	0.18	0.14 (92%)	0.16 (88%)	0.12 (92%)	0.08
Low	0.19 (84%)	0.15 (84%)	0.15 (76%)	0.14	0.12 (72%)	0.09 (56%)	0.10 (92%)	0.07
Medium	0.15 (88%)	0.13 (68%)	0.12 (76%)	0.11	0.10 (72%)	0.09 (68%)	0.08 (84%)	0.04
High	0.12 (80%)	0.10 (68%)	0.11 (78%)	0.07	0.07 (74%)	0.03 (60%)	0.05 (84%)	0.05
Very High	0.08 (72%)	0.08 (72%)	0.07 (86%)	0.05	0.08 (84%)	0.07 (72%)	0.07 (82%)	0.04

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-41: Targeting Value Improvement Across Varying Privacy Levels: Small Sample

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	PROPOSED	SPLIT-ONLY	NON-HONEST	PROPOSED	SPLIT-ONLY	NON-HONEST
No	5.24% (100%)	5.02% (92%)	1.75% (92%)	5.24% (100%)	5.02% (92%)	1.75% (92%)
Very Low	4.94% (100%)	4.81% (100%)	2.06% (96%)	4.92% (96%)	4.69% (88%)	1.75% (88%)
Low	4.67% (100%)	4.56% (96%)	1.22% (72%)	4.30% (84%)	3.93% (76%)	1.01% (64%)
Medium	4.80% (100%)	4.42% (100%)	1.13% (72%)	3.98% (76%)	3.05% (80%)	0.96% (72%)
High	6.27% (100%)	5.40% (100%)	0.62% (60%)	3.21% (78%)	2.79% (68%)	0.66% (68%)
Very High	4.89% (100%)	3.65% (88%)	-0.11% (52%)	3.08% (80%)	2.33% (74%)	1.01% (74%)

Note: We calculate the value improvement from 50 splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses). The results presented here are based on the use of R-learner with regression forests as the initial CATE model.

Web Appendix F.6 Robustness Checks of Other CATE Models

Web Appendix F.6.1 Forest-based Models

We consider three alternative forest-based CATE models as the DEFAULT method and the inital CATE models: Causal Forest and DR-learner with regression forests. Results for the T-learner are also available upon request and show consistent patterns with the findings reported here.

1. **(Causal Forest)** We use the causal forest function implemented in the `grf` package with 500 trees and other default parameters. Note that we choose 500 trees instead the default 2,000 trees to accelerate the model training process.

2. **(DR-learner)** We construct both the DR score and the CATE model using regression forests. To speed up training, we reduce the number of trees from the default 2,000 to 500, while keeping all other parameters at their default settings in the `grf` package.

Table App-42 reports the RMSE for the two CATE models. The results indicate that (i) the PROPOSED method significantly outperforms all other methods, (ii) the SPLIT-ONLY method also improves upon the DEFAULT method, while (iii) the NON-HONEST model calibration method fail to reduce RMSE.

Table App-42: RMSE of GATE (Multiplied by 100) Across Varying Privacy Levels

(a) Causal Forest

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	0.46 (96%)	0.48 (88%)	0.76 (36%)	0.75	0.46 (96%)	0.48 (88%)	0.76 (36%)	0.75
Very Low	0.44 (100%)	0.44 (100%)	0.78 (24%)	0.74	0.72 (100%)	0.75 (100%)	1.37 (28%)	1.35
Low	0.42 (100%)	0.44 (100%)	0.76 (28%)	0.72	0.99 (100%)	1.00 (100%)	1.79 (52%)	1.79
Medium	0.44 (100%)	0.44 (100%)	0.78 (36%)	0.75	1.23 (100%)	1.26 (100%)	2.15 (48%)	2.14
High	0.45 (100%)	0.46 (100%)	0.78 (20%)	0.75	1.42 (100%)	1.44 (100%)	2.42 (52%)	2.41
Very High	0.44 (100%)	0.45 (100%)	0.78 (16%)	0.73	1.55 (100%)	1.57 (100%)	2.63 (48%)	2.63

(b) DR-learner with Regression Forests

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	0.38 (100%)	0.38 (100%)	0.7 (68%)	0.69	0.38 (100%)	0.38 (100%)	0.7 (68%)	0.69
Very Low	0.82 (100%)	0.83 (100%)	1.57 (32%)	1.56	0.70 (100%)	0.71 (100%)	1.31 (70%)	1.31
Low	0.87 (100%)	0.88 (100%)	1.6 (20%)	1.59	0.93 (100%)	0.93 (100%)	1.65 (84%)	1.65
Medium	0.95 (100%)	0.94 (100%)	1.67 (32%)	1.66	1.07 (100%)	1.07 (100%)	1.9 (74%)	1.89
High	0.89 (100%)	0.91 (100%)	1.6 (52%)	1.60	1.22 (100%)	1.22 (100%)	2.13 (72%)	2.12
Very High	0.92 (100%)	0.92 (100%)	1.57 (40%)	1.57	1.36 (100%)	1.37 (100%)	2.36 (76%)	2.36

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-43 reports the AUTOC values (multiplied by 100) for different methods across various privacy levels. Consistent with the main findings in Section 6 of the paper, the PROPOSED method outperforms all other methods, and all the model calibration methods outperform the DEFAULT approach.

Table App-43: AUTOC Values (Multiplied by 100) Across Varying Privacy Levels

(a) Causal Forest

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	0.53 (72%)	0.53 (72%)	0.51 (72%)	0.50	0.53 (72%)	0.53 (72%)	0.51 (72%)	0.50
Very Low	0.51 (100%)	0.50 (100%)	0.43 (96%)	0.41	0.33 (92%)	0.33 (96%)	0.30 (96%)	0.26
Low	0.43 (100%)	0.42 (96%)	0.38 (88%)	0.35	0.20 (88%)	0.21 (86%)	0.18 (85%)	0.14
Medium	0.34 (96%)	0.33 (100%)	0.31 (92%)	0.28	0.17 (88%)	0.18 (92%)	0.16 (93%)	0.11
High	0.28 (96%)	0.28 (92%)	0.25 (96%)	0.21	0.14 (82%)	0.14 (76%)	0.12 (86%)	0.08
Very High	0.24 (90%)	0.23 (92%)	0.22 (92%)	0.18	0.14 (88%)	0.14 (80%)	0.12 (82%)	0.09

(b) DR-learner with Regression Forests

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	0.44 (84%)	0.44 (88%)	0.39 (68%)	0.38	0.44 (84%)	0.44 (88%)	0.39 (68%)	0.38
Very Low	0.39 (80%)	0.4 (80%)	0.34 (88%)	0.33	0.22 (78%)	0.23 (72%)	0.19 (96%)	0.16
Low	0.3 (76%)	0.3 (68%)	0.26 (96%)	0.24	0.12 (76%)	0.14 (84%)	0.10 (92%)	0.07
Medium	0.19 (76%)	0.21 (84%)	0.16 (92%)	0.14	0.15 (84%)	0.17 (80%)	0.13 (86%)	0.10
High	0.2 (80%)	0.2 (76%)	0.19 (100%)	0.16	0.11 (82%)	0.11 (76%)	0.08 (76%)	0.06
Very High	0.15 (60%)	0.16 (64%)	0.15 (100%)	0.12	0.06 (74%)	0.04 (72%)	0.04 (86%)	0.01

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-44 reports the average value improvement across different privacy levels. The results indicate that both the PROPOSED method and the SPLIT-ONLY method consistently outperform the DEFAULT approach.

Table App-44: Value Improvement Across Varying Privacy Levels

(a) Causal Forest

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	PROPOSED	SPLIT-ONLY	NON-HONEST	PROPOSED	SPLIT-ONLY	NON-HONEST
No	1.84% (92%)	1.62% (92%)	0.30% (64%)	1.84% (92%)	1.62% (92%)	0.30% (64%)
Very Low	3.18% (92%)	3.39% (92%)	0.5% (72%)	5.55% (100%)	5.67% (100%)	1.22% (84%)
Low	3.12% (92%)	3.49% (88%)	0.02% (44%)	7.65% (100%)	8.43% (100%)	2.02% (92%)
Medium	3.9% (92%)	3.36% (92%)	-0.55% (40%)	6.52% (100%)	7.17% (100%)	1.74% (84%)
High	3.73% (96%)	3.75% (100%)	-0.77% (36%)	6.67% (96%)	7.31% (92%)	1.74% (86%)
Very High	4.21% (96%)	3.88% (96%)	0.07% (56%)	6.78% (100%)	6.50% (92%)	2.16% (82%)

(b) DR-learner with Regression Forests

Privacy	Scenario 1: DP-Protected Covariates			Scenario 2: DP-Protected Outcome		
	PROPOSED	SPLIT-ONLY	NON-HONEST	PROPOSED	SPLIT-ONLY	NON-HONEST
No	0.43% (84%)	0.43% (84%)	0.02% (68%)	0.43% (84%)	0.43% (84%)	0.02% (68%)
Very Low	5.58% (96%)	5.36% (96%)	0.51% (68%)	2.20% (96%)	2.20% (96%)	0.07% (60%)
Low	5.33% (92%)	5.51% (96%)	0.72% (76%)	3.95% (96%)	3.98% (96%)	0.33% (62%)
Medium	5.77% (94%)	6.02% (92%)	0.16% (56%)	4.71% (98%)	4.72% (98%)	0.33% (60%)
High	5.75% (92%)	6.05% (96%)	0.63% (68%)	5.58% (96%)	5.52% (92%)	0.29% (58%)
Very High	5.32% (92%)	5.23% (92%)	0.46% (60%)	7.05% (100%)	6.92% (100%)	0.40% (60%)

Note: We calculate the value improvement from 50 splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Web Appendix F.6.2 XGBoost-based Models

Next, we consider two additional boosting-tree models: R-learner and DR-learner with XGBoost models. Results for the T-learner are available upon request and are consistent with the findings reported here. To fine-tune the hyperparameters in XGBoost models, we perform a single train-holdout split under the non-privacy setting. For conditional mean outcome models, we select parameters that minimize mean squared prediction error, while for the CATE model in the R-learner, we choose parameters that maximize the AUTOC value. The search ranges for key hyperparameters include: learning rate $\eta \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$, maximum depth in each tree $\{1, 2, 5, 10, 20\}$, and the maximum number of iterations $\{10, 20, 30, 40, 50\}$. The optimal hyperparameters are as follows:

- **(R-learner)** Conditional mean outcome model: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 50; CATE model: $\eta = 0.15$, maximum tree depth: 4, the maximum number of iterations: 10.

- **(DR-learner)** Conditional mean outcome model for the treatment group: $\eta = 0.15$, maximum tree depth: 2, the maximum number of iterations: 50; Conditional mean outcome model for the control group: $\eta = 0.20$, maximum tree depth: 2, the maximum number of iterations: 20; CATE model: $\eta = 0.15$, maximum tree depth: 4, the maximum number of iterations: 10.

Table App-45 reports the RMSE for the two CATE models. For R-learner, all the model calibration methods significantly outperform the DEFAULT method. For T-learner, both the PROPOSED and SPLIT-ONLY methods outperform the DEFAULT method, while the NON-HONEST model calibration method fail to reduce RMSE in the setting with DP-protected outcomes.

Table App-45: RMSE of GATE (Multiplied by 100) Across Varying Privacy Levels

(a) R-learner wtih XGBoost Models

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	0.46 (100%)	0.48 (100%)	0.48 (100%)	9.62	0.46 (100%)	0.48 (100%)	0.48 (100%)	9.62
Very Low	0.50 (100%)	0.47 (100%)	0.45 (100%)	9.65	0.53 (100%)	0.59 (100%)	0.52 (100%)	9.63
Low	0.48 (100%)	0.43 (100%)	0.43 (100%)	9.65	0.58 (100%)	0.71 (100%)	0.58 (100%)	9.69
Medium	0.48 (100%)	0.44 (100%)	0.45 (100%)	9.66	0.66 (100%)	0.8 (100%)	0.70 (100%)	9.61
High	0.47 (100%)	0.41 (100%)	0.41 (100%)	9.61	0.73 (100%)	0.9 (100%)	0.77 (100%)	9.56
Very High	0.49 (100%)	0.42 (100%)	0.41 (100%)	9.66	0.84 (100%)	1.01 (100%)	0.92 (100%)	9.52

(c) DR-learner with XGBoost Models

Privacy	Scenario 1: DP-Protected Covariates				Scenario 2: DP-Protected Outcome			
	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT	PROPOSED	SPLIT-ONLY	NON-HONEST	DEFAULT
No	0.56 (100%)	0.50 (100%)	0.46 (100%)	9.70	0.56 (100%)	0.50 (100%)	0.46 (100%)	9.70
Very Low	0.47 (100%)	0.44 (100%)	0.4 (100%)	9.69	0.82 (100%)	0.59 (100%)	0.59 (100%)	9.70
Low	0.51 (100%)	0.43 (100%)	0.43 (100%)	9.68	0.85 (100%)	0.62 (100%)	0.59 (100%)	9.59
Medium	0.46 (100%)	0.4 (100%)	0.4 (100%)	9.65	1.03 (100%)	0.82 (100%)	0.79 (100%)	9.55
High	0.52 (100%)	0.45 (100%)	0.41 (100%)	9.67	1.01 (100%)	0.81 (100%)	0.73 (100%)	9.71
Very High	0.53 (100%)	0.42 (100%)	0.41 (100%)	9.63	1.22 (100%)	1.02 (100%)	1.03 (100%)	9.64

Note: We calculate the RMSE from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the DEFAULT approach (given in parentheses).

Table App-46 reports the AUTOC values (multiplied by 100) for different methods across various privacy levels. For the R-learner, all model calibration methods significantly outperform the DEFAULT method.

Table App-47 presents the average value improvement across different methods. For both the R-learner and DR-learner, the calibrated models do not produce significant gains over the DEFAULT model. This is because all models, including the DEFAULT and the various calibrated versions, predict positive treatment effects for over 99.9% of customers. As a result, when targeting is based solely on whether the predicted CATE is positive, all models recommend targeting nearly the entire customer base. Nevertheless, the substantial improvements in RMSE and AUTOC highlight the value of honest model calibrations. This benefit becomes particu-

Table App-46: AUTOC Values (Multiplied by 100) Across Varying Privacy Levels

(a) R-learners with XGBoost Models

Privacy	**Scenario 1: DP-Protected Covariates**				**Scenario 2: DP-Protected Outcome**			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.62 (64%)	0.60 (44%)	0.62 (56%)	0.61	0.62 (64%)	0.60 (44%)	0.62 (56%)	0.61
Very Low	0.60 (65%)	0.56 (60%)	0.55 (80%)	0.55	0.57 (56%)	0.52 (38%)	0.57(56%)	0.56
Low	0.48 (72%)	0.49 (76%)	0.50 (84%)	0.47	0.47 (80%)	0.41 (36%)	0.39 (28%)	0.44
Medium	0.42 (80%)	0.42 (72%)	0.42 (80%)	0.42	0.39 (84%)	0.38 (64%)	0.38 (68%)	0.33
High	0.39 (74%)	0.39 (76%)	0.39 (72%)	0.36	0.35 (82%)	0.30 (64%)	0.30 (56%)	0.26
Very High	0.31 (76%)	0.31 (68%)	0.30 (70%)	0.29	0.33 (76%)	0.29 (68%)	0.32 (74%)	0.25

(b) DR-learner with XGBoost Models

Privacy	**Scenario 1: DP-Protected Covariates**				**Scenario 2: DP-Protected Outcome**			
	Proposed	Split-only	Non-Honest	Default	Proposed	Split-only	Non-Honest	Default
No	0.68 (80%)	0.68 (84%)	0.65 (82%)	0.60	0.68 (80%)	0.68 (84%)	0.65 (82%)	0.60
Very Low	0.58 (78%)	0.57 (56%)	0.58 (72%)	0.56	0.61 (78%)	0.59 (74%)	0.61 (82%)	0.57
Low	0.51 (88%)	0.50 (76%)	0.49 (78%)	0.46	0.53 (74%)	0.52 (72%)	0.53 (74%)	0.49
Medium	0.43 (74%)	0.43 (72%)	0.44 (64%)	0.40	0.49 (72%)	0.44 (78%)	0.42 (78%)	0.35
High	0.40 (72%)	0.39 (64%)	0.40 (72%)	0.36	0.42 (88%)	0.43 (88%)	0.46 (84%)	0.38
Very High	0.32 (70%)	0.32 (66%)	0.33 (74%)	0.30	0.31 (92%)	0.29 (88%)	0.29 (90%)	0.17

Note: We calculate the AUTOC values from 50 random splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

larly important in practical applications where targeting is constrained—such as when firms can only reach a subset of customers due to budget limitations or strategic priorities.

Table App-47: Targeting Value Improvement Across Varying Privacy Levels

(a) R-learner with XGBoost Models

Privacy	**Scenario 1: DP-Protected Covariates**			**Scenario 2: DP-Protected Outcome**		
	Proposed	Split-only	Non-Honest	Proposed	Split-only	Non-Honest
No	0.16% (64%)	0.16% (62%)	0.02% (52%)	0.16% (64%)	0.16% (62%)	0.02% (52%)
Very Low	0.01% (52%)	0.01% (52%)	0.00% (54%)	0.03% (52%)	0.03% (64%)	0.02% (48%)
Low	-0.03% (46%)	-0.03% (40%)	0.00% (52%)	0.04% (56%)	-0.04% (52%)	0.00% (52%)
Medium	-0.01% (48%)	-0.01% (56%)	-0.01% (60%)	0.02% (54%)	0.01% (56%)	0.02% (48%)
High	0.05% (60%)	-0.05% (48%)	0.01% (58%)	0.05% (52%)	-0.04% (46%)	-0.01% (56%)
Very High	-0.02% (50%)	-0.06% (46%)	-0.01% (44%)	0.12% (52%)	0.14% (56%)	-0.04% (46%)

(c) DR-learner with XGBoost Models

Privacy	**Scenario 1: DP-Protected Covariates**			**Scenario 2: DP-Protected Outcome**		
	Proposed	Split-only	Non-Honest	Proposed	Split-only	Non-Honest
No	0.03% (60%)	0.02% (54%)	0.01% (58%)	0.03% (60%)	0.02% (54%)	0.01% (58%)
Very Low	0.02% (60%)	0.01% (64%)	-0.01% (62%)	-0.02% (46%)	-0.02% (46%)	-0.01% (52%)
Low	0.00% (58%)	0.00% (54%)	0.01% (52%)	0.01% (58%)	0.01% (60%)	0.01% (54%)
Medium	0.01% (52%)	-0.02% (44%)	0.00% (54%)	0/02% (64%)	0.00% (50%)	0.02% (52%)
High	-0.01% (48%)	-0.02% (42%)	0.01% (58%)	-0.01% (64%)	-0.02% (56%)	0.00% (48%)
Very High	0.00% (68%)	0.01% (52%)	-0.02% (64%)	0.03% (58%)	0.02% (56%)	0.00% (56%)

Note: We calculate the value improvement from 50 splits, along with the percentage of replications in which the value of the focal method is greater than the value of the Default approach (given in parentheses).

References

Abel AB (2018) Classical measurement error with several regressors. *Working Paper* .

Agarwal A, Singh R (2021) Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780* .

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Battistin E, Chesher A (2014) Treatment effect estimation with covariate measurement error. *Journal of Econometrics* 178(2):707–715.

Bonhomme S, Robin JM (2010) Generalized non-parametric deconvolution with an application to earnings dynamics. *The Review of Economic Studies* 77(2):491–533.

Bound J, Brown CC, Duncan G, Rodgers WL (1989) Measurement error in cross-sectional and longitudinal labor market surveys: Results from two validation studies.

Chen X, Hong H, Tamer E (2005) Measurement error models with auxiliary data. *The Review of Economic Studies* 72(2):343–366.

Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I (2018) Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india.

Chesher A (1991) The effect of measurement error. *Biometrika* 78(3):451–462.

Cohen J (2013) *Statistical power analysis for the behavioral sciences* (Academic press).

Erickson T, Whited TM (2002) Two-step gmm estimation of the errors-in-variables model using high-order moments. *Econometric Theory* 18(3):776–799.

Fan J, Truong YK (1993) Nonparametric regression with errors in variables. *The Annals of Statistics* 1900–1925.

Friedman JH (2002) Stochastic gradient boosting. *Computational statistics & data analysis* 38(4):367–378.

Hausman J, Newey W, Ichimura H, Powell J (1991a) Measurement errors in polynomial regression models. *Journal of Econometrics* 50(3):273–295.

Hausman JA, Newey WK, Ichimura H, Powell JL (1991b) Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics* 50(3):273–295.

Hu Y, Ridder G (2012) Estimation of nonlinear models with mismeasured regressors using marginal information. *Journal of Applied Econometrics* 27(3):347–385.

Hu Y, Schennach SM (2008) Instrumental variable treatment of nonclassical measurement error models. *Econometrica* 76(1):195–216.

Kennedy EH (2023) Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17(2):3008–3049.

Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Li T (2002) Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics* 110(1):1–26.

Newey WK (2001) Flexible simulated moment estimation of nonlinear errors-in-variables models. *Review of Economics and statistics* 83(4):616–627.

Nie X, Wager S (2021) Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108(2):299–319.

Pal M (1980) Consistent moment estimators of regression coefficients in the presence of errors in variables. *Journal of Econometrics* 14(3):349–364.

Schennach SM (2004) Estimation of nonlinear models with measurement error. *Econometrica* 72(1):33–75.

Schennach SM (2007) Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica* 75(1):201–239.

Schennach SM, Hu Y (2013) Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association* 108(501):177–186.

Sepanski JH, Carroll RJ (1993) Semiparametric quasilikelihood and variance function estimation in measurement error models. *Journal of Econometrics* 58(1-2):223–256.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge university press).

Whitehouse J, Jung C, Syrgkanis V, Wilder B, Wu ZS (2024) Orthogonal causal calibration. *arXiv preprint arXiv:2406.01933* .

Wolter KM, Fuller WA (1982) Estimation of nonlinear errors-in-variables models. *The Annals of Statistics* 539–548.